Structured Probabilistic Reasoning

(Incomplete draft)

Bart Jacobs
Institute for Computing and Information Sciences, Radboud University Nijmegen,
P.O. Box 9010, 6500 GL Nijmegen, The Netherlands.
bart@cs.ru.nl http://www.cs.ru.nl/~bart

Version of: September 11, 2025

Contents

Preface vi

1 Collections 1
1.1 Notation 3
1.2 Coefficients 5
1.3 Cartesian products 9
1.4 Lists 12
1.5 Subsets 22
1.6 Multisets 31
1.7 Multisets in summations 41
1.8 Coefficients of multisets 48
1.9 Multiset partitions and the triangular prism 57
1.10 Combinatorics 63
1.11 Channels 71
1.12 The role of category theory 75

2 Discrete probability distributions 82
2.1 Probability distributions 83
2.2 Frequentist learning and stick breaking 97
2.3 Parallel products 106
2.4 Probabilistic channels 118
2.5 String diagrams and Bayesian networks 130
2.6 Draw distributions 141
2.7 Convolution 153
2.8 A glimpse of probabilistic logic programming 165
2.9 A glimpse of probabilistic programming using WebPPL 178
2.10 Divergence between distributions 185
2.11 Exchangeability for positions and elements 197

3 Drawing from an urn 207
3.1 Zipping multisets 208
3.2 Single draws 216
3.3 The multinomial channel 227
3.4 Divergence and multinomials 238
3.5 The hypergeometric channel 243
3.6 The Pólya channel 249
3.7 The parallel multinomial law 258

3.8 The parallel multinomial law: basic properties 262
3.9 Parallel multinomials as law of monads 274
3.10 Discrete Poisson point processes 281

4 Observables and validity 289
4.1 Validity 290
4.2 The mean of a distribution 298
4.3 Random variables 308
4.4 The structure of observables 311
4.5 Transformation of observables 324
4.6 Validity and drawing 333
4.7 Distributions parameterised by predicates and factors 341
4.8 Validity-based distances 346

5 Variance and covariance 361
5.1 Variance and shared-state covariance 362
5.2 Draw distributions and their (co)variances 375
5.3 Joint-state covariance and correlation 384
5.4 Independence for random variables 393
5.5 Moments 396
5.6 The laws of large draws and of large numbers 405

6 Updating distributions 410
6.1 Update basics 411
6.2 Examples of forward and backward inference 425
6.3 Analysis of forward and backward inference 440
6.4 Inference in Bayesian networks 447
6.5 Updating in probabilistic logic programming 453
6.6 Updating in probabilistic programming 460
6.7 Hidden Markov models 464
6.8 Updating draw distributions 473
6.9 Discretisation, and coin bias learning 480
6.10 Frequentist and Bayesian probability 488
6.11 Updating and (co)variance 494
6.12 Updating with e-powers of observables 500

7 Daggers and disintegrations 510
7.1 Bayesian inversion: the dagger of a channel 511
7.2 Disintegration of joint distributions 523
7.3 Disintegration, in general form 531
7.4 Disintegration for learning with missing data 540
7.5 Disintegration and inversion in machine learning 544
7.6 Daggers of draw channels and Laplace's rule of succession 554
7.7 Sufficient statistics 562
7.8 Factorisation of joint states 574
7.9 Pearl's and Jeffrey's update rules 583
7.10 Disintegration of Bayesian Networks 599
7.11 Categorical aspects of Bayesian inversion 609

8 Learning of Distributions and Channels 616
8.1 When to use conjunctions 617
8.2 A general sum-increase method 623
8.3 Validity and updating with multisets of factors 632
8.4 Properties of updating with multisets of factors 638
8.5 Ideal expectation maximisation via double daggers 650
8.6 Expectation maximisation without parameters 654
8.7 Expectation maximisation with parameters 664
8.8 Learning Markov chains 673
8.9 Learning hidden Markov models 681

9 Causal diagram surgery and counterfactual reasoning 692
9.1 Simpson's paradox 695
9.2 Two causality challenges of Pearl 703
9.3 Two illustrations of counterfactual reasoning 710
9.4 Counterfactual reasoning and randomness pushback 718

10 Probabilistic Automata 724
10.1 Markov models with output 725
10.2 Markov models with rewards 736
10.3 Markov decision processes 741
10.4 Probabilistic automata as coalgebras 746
10.5 Invariants and nexttime operators 754
10.6 Behavioural distances 763
References 771

Preface

No victor believes in chance.

Friedrich Nietzsche,

Die fröhliche Wissenschaft, §258, 1882.

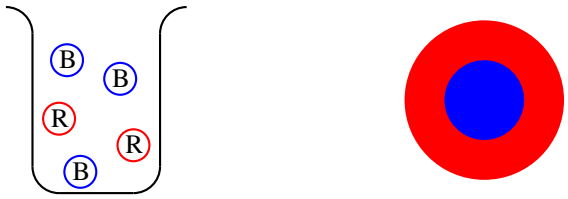
Originally in German:

Kein Sieger glaubt an den Zufall.

Probability is for losers — a defiant rephrase of the above aphorism of the German philosopher Friedrich Nietzsche. According to him, winners do not reason with probabilities, but with certainties, via Boolean logic, one could say. However, this goes against the current trend, in which reasoning with probabilities has become the norm, in large scale data analytics and in artificial intelligence (AI); it is Boolean reasoning that is now losing influence, see also the famous *End of Theory* article Anderson (2008) from 2008. This book is about the mathematical structures underlying the reasoning, not of Nietzsche's winners, but of today's apparent winners.

The phrase 'structure in probability' in the title of this book may sound like a *contradictio in terminis*: it seems that probability is about randomness, like in the tossing of coins, in which one may not expect to find much structure. Still, as we know since the seventeenth century, via the pioneering work of Christiaan Huygens, Pierre de Fermat, and Blaise Pascal, there is quite some mathematical structure in the area of probability. The *raison d'être* of this book is that there is more structure — especially algebraic and categorical — than is commonly emphasised.

One can distinguish two important sources of probabilities, namely *counting* and *measuring*. This is illustrated in the two pictures below. On the left we see an urn filled with five coloured balls: two red (R) and three blue (B). In a random draw of a single ball from the (well-mixed, stirred) urn the probability of drawing a red ball is $\frac{2}{5}$. This probability arises from counting: two of the five balls in total are red.



(0.1)

probability-from-counting: draw a ball **probability-from-measuring: throw a dart**

On the right the two coloured circles have radius 1 and 2. The outer red area is thus three times bigger than the inner blue area. If we now randomly throw a dart at the board — and know for sure that we will hit it — then the probability that it hits red is $\frac{3}{4}$. This 'continuous' situation on the right can be approximated by the 'discrete' approach on the left: when the two red and blue areas are covered with finitely many, increasinly small areas, which are treated as balls in an urn, then the probability of drawing red approximates $\frac{3}{4}$.

In this book we shall describe probability-from-counting systematically via *multisets*, via an operation that we call frequentist learning. Informally, a multiset is like a subset, except that elements may occur multiple times. Urns filled with coloured balls are physical realisations of multisets. In multiset form, the above urn will be written as $2|R\rangle + 3|B\rangle$, expressing that it contains 2 red (R) items and 3 blue (B) ones. The ket notation $|-\rangle$ is just syntactic sugar that is used to separate the multiplicities (2 and 3) from the elements (R and B). This probability-from-counting is often referred to as *discrete* probability, in contrast to *continuous* probability, which is based on measuring the size of subsets, that is on probability-from-measuring. Continuous probability theory is mathematically more challenging than discrete probability theory. In computer science, discrete probabilities suffice in many cases. Mathematicians typically go directly to the continuous case and largely ignore multisets — a historic mistake. When urns are represented not as multisets, but as subsets, they can have at most one ball per colour. That does not work. Counting elements of subsets — a form of measurement — is not a substitute for counting elements of multisets.

In the sequel we put much more emphasis on discrete probability theory than on the continuous case. The interplay between them is relevant. One can reduce the continuous case to the discrete case by discretisation: chopping up the whole area into finitely many small areas. It involves a certain loss of precision. One can move from discrete to continuous via a (consistent) limit process, as expressed by the central limit theorem, or by de Finetti's theorem, as we shall see later.

The scientific roots of this book's author lie outside probability theory, in type theory and logic (including some quantum logic), in semantics and specification of programming languages, in computer security and privacy, in state-based computation (coalgebra), and in category theory. This scientific distance to probability theory has advantages and disadvantages. Its obvious disadvantage is that there is no deeply engrained familiarity with the field and with its development. But at the same time this distance may be an advantage, since it provides a fresh perspective, without sacred truths and without adherence to common practices and notations. For instance, the terminology and notation in this book are influenced by quantum theory, *e.g.* in using ket notation $|-\rangle$ for multisets and discrete probability distributions, in using daggers as reversals, in analogy with conjugate transposes (for Hilbert spaces), or in using the words 'state', 'observable'

and 'test' — as synonyms for 'distribution', for $\mathbb{R}$-valued function on a sample space, and for compatible (summable) predicates

It should be said: for someone trained in formal methods, the area of probability theory can be rather sloppy: everything is called 'P', types are hardly ever used, equational reasoning seldom happens, crucial ingredients (like distributions in expected values or in random variables) are left implicit, basic notions (like conjugate prior) are introduced only via examples, calculation recipes and algorithms are regularly just given, without explanation, goal or justification, *etc.* This hurts, especially because there is so much beautiful mathematical structure around. For instance, the notion of a channel (see below) formalises the idea of a conditional probability and carries a rich mathematical structure that can be used in compositional reasoning, with both sequential and parallel composition and with reversal: the Bayesian inversion ('dagger') of a channel does not only come with appealing mathematical (categorical) properties — *e.g.* smooth interaction with sequential and parallel composition — but is also extremely useful in inference and learning. Via this dagger we can connect forward and backward inference (see Corollary 7.1.7: backward inference is forward inference with the dagger, and vice-versa) and capture the difference between Pearl's and Jeffrey's update rules (see Theorem 6.1.6: Pearl increases validity, whereas Jeffrey decreases divergence).

We even dare to think that this 'sloppiness' is ultimately a hindrance to further development of the field, especially in computer science, where computer-assisted reasoning and explaining requires a formal syntax and a clear semantics. For instance, it is hard to even express the above-mentioned Corollary 7.1.7 and Theorem 6.1.6 in standard probabilistic notation. Wittgenstein said that the limits of our language determine the limits of our thinking. This applies certainly in probability theory, where the traditional language introduced by Kolmogorov does not always allow us to express the relevant matters. This is clear in the habit to keep probability distributions implicit, for instance in the notation $\mathbb{E}[X]$ for the expected value of a random variable X, or in the notion of random variable itself. The latter is a pair of a distribution together with a function, from the underlying sample space of the distribution to the reals. This distribution is usually not mentioned. For instance, it is often completely missing in the definition of random variable in *e.g.* (Casella and Berger, 2002, Def. 1.4.1). This absence makes it not only difficult to understand what is going on, but also easily leads to confusion. For instance, when we have two random variables at hand, do they involve the same distribution or not? Also, some notions, like convolution, are traditionaly defined in terms of random variables, but convolution really is an operation on distributions.

One can speculate that states / distributions are kept implicit in traditional probability theory because in many examples they are used as a fixed implicit assumption in the background. Indeed, in mathematical notation one tends to omit — for efficiency — the least relevant (implicit) parameters. But the essence of probabilistic computation is state transformation, where it has become highly relevant to know explicitly in which state (distribution) one is working at which stage. The notation developed in this book helps in such situations — and in many other situations as well, we hope. There is some understandable resistance against new notation and reluctance to adopt it, especially in a mathematical context. In contrast, computer scientists are more pragmatic and switch relatively easily to a new (programming) language if it better allows them to express what they aim to address.

Apart from having beautiful structure, probability theory also has *magic*. It can be found, for instance, in the following two points.

1 Probability distributions can be updated, making it possible to absorb information (evidence) into them and learn. Multiple updates, based on data, can be used for training, so that a distribution absorbs more and more information and can subsequently be used for prediction or classification.
2 The components of a joint distribution, over a product space, can 'listen' to each other, so that updating in one (product) component has crossover effects in other components. These ripple effects look like what happens in quantum physics, where measuring one part of an entangled quantum systems changes other parts.

The combination of these two points is very powerful and forms the basis for probabilistic reasoning. For instance, if we know that two phenomena are related, and we have new information about one of them, then we also learn something new about the other phenomenon, after updating. We shall see that such crossover ripple effects can be described in two equivalent ways, starting from a joint distribution with evidence in one component.

- We can use the 'weaken-update-marginalise' approach, where we first weaken the evidence from one component to the whole product space, so that it fits the joint distribution and can be used for updating; subsequently, we marginalise the updated state to the component that we wish to learn more about. That's where the ripple effect through the joint distribution becomes visible.
- We can also use the 'extract-infer' technique, where we first extract a conditional probability (channel) from the joint distribution and then do (forward or backward) inference with the evidence, along the channel. This is what happens if we reason in Bayesian networks, represented as a diagram of channels, when information at one point in the network is transported up and down the connections, in order to draw conclusions at another point in the network.

The equivalence of these two approaches will be demonstrated and exploited at various places in this book, see *e.g.* Remark 7.2.9.

Here is a characteristic illustration of our structure-based approach. A well known property of the Poisson distribution *pois* is commonly expressed as: if $X_1 \sim pois[\lambda_1]$ and $X_2 \sim pois[\lambda_2]$ then $X_1 + X_2 \sim pois[\lambda_1 + \lambda_2]$. This formulation uses random variables X_1, X_2, which are Poisson-distributed. We shall formulate this fact as an (algebraic) structure preservation property of the Poisson distribution, without using any random variables. The property is: $pois[\lambda_1 + \lambda_2] = pois[\lambda_1] + pois[\lambda_2]$. It says that the Poisson channel *pois* is a homomorphism of monoids, from the commutative monoid non-negative reals $\mathbb{R}_{\geq 0}$ with addition, to the commutative monoid of distributions $\mathcal{D}(\mathbb{N})$ with convolution, see Proposition 2.7.9 for details. This result uses a convolution commutative monoid structure on distributions whose underlying space is itself a commutative monoid — in this case $\mathbb{N}$ with addition. This convolution monoid structure plays a role in many other situations, for instance, by iterated convoluttion sums $n \cdot \omega = \omega + \cdots + \omega$ of an arbitrary discrete distribution ω on the real numbers one obtains a Bell shape, when n goes to infinity, corresponding to a (continuous) normal distribution. Also, convolution plays a role in the fundamental distributive law that turns a multiset of distributions into a distribution of multisets.

The following aspects characterise the approach of this book.

1 *Channels* are used as a cornerstone in probabilistic reasoning. The concept of (communication) channel is widely used elsewhere, under various names, such as conditional probability,

stochastic matrix, probabilistic classifier, Markov kernel, statistical model, conditional probability table (in Bayesian networks), probabilistic function / computation, signal (in Bayesian persuasion theory), and finally as Kleisli map (in category theory). Channels can be composed sequentially and in parallel, and can be reversed. They can be used to transform both states and predicates, in forward and backward directions. Channels exist for all relevant collection types (lists, subsets, multisets, distributions), for instance for non-deterministic, and for probabilistic computation. However, after the first chapter about collection types, channels will be used exclusively for distributions, in probabilistic form.

2 *Multisets* play a central role to capture various forms of data, like coloured balls in an urn, draws from such an urn, tables, data inputs for successive learning steps, *etc.* The interplay between multisets and distributions, notably in learning, is a recurring theme. Also, many distributions over multisets will occur, for instance the standard distributions (multinomial, hypergeometric, Pólya) for drawing a handful of balls from an urn. Indeed, not only the urn is a multiset, but also a draw of balls from the urn.

3 *Distributions* (states) are used as first class citizens and are always written explicitly. Distributions are treated as separate from, and dual to, *predicates*. These predicates are used as evidence and form the ingredients of an (implicit) probabilistic logic, with, for instance conjunction and negation operations and updating. Distributions are really different from predicates, with their own operations, without, for instance conjunction and negation. In this book, predicates standardly occur in fuzzy (soft, non-sharp) form, taking values in the unit interval $[0, 1]$. They form fuzzy events. Along a channel one can transfer states forward, and predicates backward. A central notion is the validity of a predicate p in a distribution ω, written as $\omega \models p$. It is standardly called expected value, written as $\mathbb{E}[p]$, leaving ω implicit, or alternatively as $\mathop{\mathbb{E}}\limits_{x \sim \omega} p(x)$. Conditioning involves updating a 'prior' distribution with evidence — in the form of a predicate — resulting in a new 'posterior' distribution.

4 Probabilistic reasoning in this book is done in an *exact* manner, using the relevant mathematical formulas. Approximations via sampling of distributions is very relevant in practice, because state space explosions associated with joint distributions quickly make calculations intractable. Nevertheless, the focus here is on the mathematical structure in probability and not on what is practically computable.

It is not purely mathematical formality and aesthetics that drive the developments in this book. Probability theory nowadays forms the basis of large parts of big data analytics and of artificial intelligence. These areas are of increasing societal relevance and provide the basis of the modern view of the world — more based on correlation than on causation — and also provide the basis for much of modern decision making, that may affect the lives of billions of people in profound ways. There are increasing demands for justification and explanation of such probabilistic reasoning methods and decisions, for instance in the legal setting provided by Europe's General Data Protection Regulation (GDPR). Its recital 71 is about automated decision-making and talks about a right to obtain an explanation:

> In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

It is not acceptable that your mortgage application is turned down because you drive a blue car

— even in presence of a clear correlation between driving blue cars and being late on one's mortgage payments.

These and other developments have led to a new area called Explainable Artificial Intelligence (XAI), which strives to provide decisions with explanations that can be understood easily by humans, without bias or discrimination. Although this book will not contribute to XAI as such, it aims to provide a mathematically solid basis for such explanations.

In this context it is appropriate to quote Judea Pearl Pearl (1989) from 1989 about a divide that is still wide today.

> To those trained in traditional logics, symbolic reasoning is the standard, and nonmonotonicity a novelty. To students of probability, on the other hand, it is symbolic reasoning that is novel, not nonmonotonicity. Dealing with new facts that cause probabilities to change abruptly from very high values to very low values is a commonplace phenomenon in almost every probabilistic exercise and, naturally, has attracted special attention among probabilists. The new challenge for probabilists is to find ways of abstracting out the numerical character of high and low probabilities, and cast them in linguistic terms that reflect the natural process of accepting and retracting beliefs.

This book does not pretend to fill this gap. One of the big embarrassments of the field is that there is no widely accepted symbolic logic for probability — including updating — together with proof rules and a denotational semantics. Such a logic for symbolic reasoning about probability will be non-trivial, because it will have to be non-monotonic[1] — a property that many logicians shy away from. This book does aim to contribute towards bridging the divide mentioned by Pearl, by providing a mathematical basis for such a symbolic probabilistic logic, consisting of channels, distributions, predicates, transformations, multisets, conditioning, disintegration, *etc.*

From the perspective of this book, the structured categorical approach to probability theory began with the work of Bill Lawvere (already in the 1960s) and his student Michèle Giry — later continued by another student of Lawvere's, namely Anders Kock Kock (2012). Lawever and Giry recognised that taking probability distributions has the structure of a monad, which was published in the early 1980s in Giry (1982). Roughly at the same time Dexter Kozen started the systematic investigation of probabilistic programming languages and logics, published in Kozen (1981, 1985). The monad introduced back then is now called the Giry monad $\mathcal{G}$, whose restriction to finite discrete probability distributions is written as $\mathcal{D}$. Most of this book concentrates on this discrete form. The language and notation that is used, however, covers both discrete and continuous probability — and quantum probability too (inspired by the general categorical notion of effectus, see Cho et al. (2015); Jacobs (2015)).

Since the early 1980s the area of categorical probability theory remained relatively calm. It is only in the new millenium that there is renewed attention, sparked in particular by several developments.

- The grown interest in probabilistic programming languages that incorporate updating (conditioning) and/or higher order features, see *e.g.* Dahlqvist et al. (2016); Dahlqvist and Kozen (2020); Danos and Ehrhard (2011); Ehrhard et al. (2014); Raedt and Kimmig (2015); Olmedo et al. (2018); Staton et al. (2016); Heunen et al. (2017); Bacci et al. (2018); Staton (2021).
- The compositional approach to Bayesian networks Coecke and Spekkens (2012); Fong (2012) and to Bayesian reasoning Culbertson and Sturtz (2014); Jacobs and Zanasi (2016, 2021).

[1] Informally, a logic is non-monotonic if adding assumptions may make a conclusion less true. For instance, I may think that scientists are civilised people, until, at some conference dinner, a heated scientific debate ends in a fist fight.

- The use of categorical and diagrammatic methods in quantum foundations, including quantum probability, see Coecke and Kissinger (2016); Piedeleu and Zanasi (2025) for overviews.
- The efforts to develop 'synthetic' probability theory via a categorical axiomatisation, see *e.g.* Fritz (2020); Fritz and Rischel (2020); Ścibior et al. (2018); Cho and Jacobs (2017); Jacobs (2021c).

This book builds on these developments.

The intended audience consists of students and professionals — in mathematics, computer science, artificial intelligence, natural sciences (see also Dill and Bromberg (2010)), and related fields — with a basic background in probability and in algebra and logic — and with an interest in formal, logically oriented approaches. This book's goal is not to provide intuitive explanations of probability, like Tijms (2007), but to provide clear and precise formalisations of the relevant structures. Mathematical abstraction (esp. categorical air guitar playing) is not a goal in itself. Instead, the book tries to uncover relevant abstractions in concrete problems, illustrated by many examples. It includes several basic algorithms, with a focus on the algorithms' correctness, not their efficiency. Each section ends with a series of exercises, so that the book can also be used for teaching and/or self-study. It aims at an undergraduate level. Some exercises invite the reader to perform some calculations, in order to understand better how definitions and results work. Other exercises contain additional facts and theory that the reader may wish to elaborate and prove. No familiarity with category theory is assumed. The basic, necessary notions are explained along the way. People who wish to learn more about category theory can use the references in the text, consult modern introductory texts like Awodey (2006); Leinster (2014), or use online resources such as ncatlab.org or Wikipedia.

Contents overview

We give a high-level overview of the various chapters of this book. Each chapter starts with a more detailed description of its contents.

The first chapter of the book covers introductory material that is meant to set the scene. It provides a crash course in discrete mathematics, building on the basic collection types of subsets, lists, and multisets, together with the basic (natural) transformations between them: (Multisets) $\to$ (Lists) $\to$ (Subsets). The chapter discusses the (free) monoid structure on all these collection types and introduces 'unit' and 'flatten' maps as their common, underlying structure. It also introduces the basic concept of a channel, for these three collection types, and shows how channels can be used for state transformation and how they can be composed, both sequentially and in parallel. At the end, the chapter provides definitions of the relevant notions from category theory. This first chapter contains various combinatorial results related to the basic collection types. They lay the foundation for many discrete probability distributions later on.

In the second chapter (discrete) probability distributions first emerge, as a special collection type, with their own associated form of (probabilistic) channel. The subtleties of parallel products of distributions (states), with entwinedness / correlation between components and the non-naturality of copying, are discussed at this early stage. This culminates in an illustration of Bayesian networks in terms of (probabilistic) channels. It shows how predictions are made within such Bayesian networks via state transformation and via compositional reasoning, basically by translating the network structure into (sequential and parallel) composites of channels. In this

chapter we start using string diagrams as graphical representation. They have been developed in physics and in category theory, to deal with the relevant compositional structure (given by symmetric monoidal categories). The chapter also introduces convolution sums of distributions and the Kullback-Leibler divergences between them.

Blindly drawing coloured balls from an urn is a basic model in discrete probability. Such draws are introduced briefly in Chapter 2 but analysed systematically in Chapter 3 in three basic forms: draw-delete (called 'hypergeometric'), draw-replace ('multinomial') and draw-duplicate ('Pólya'). Formulated in terms of channels, these distributions satisfy various compositionality properties. Such properties are typical for this book's approach and are (largely) absent in traditional treatments of this topic. Urns, and draws from urns, are both described as multisets. The interplay between multisets and distributions is an underlying theme in this chapter. There is a fundamental distributive law between multisets and distributions that expresses basic structural properties.

The fourth chapter is more logically oriented, via observables of the form $X \to \mathbb{R}$. Such maps include what we call factors, predicates and events as special forms, with more restricted codomains, like $[0, 1] \subseteq \mathbb{R}$). They are defined on sample spaces X of outcomes, providing numerical information. The chapter concentrates on validity (expected value) of obervables in states and on transformation of observables. Validity shows up in many definitions and properties. For instance, it is used to define the mean of a distribution or the (total variation) distance between distributions and also between predicates. In terms of this distance we formulate the 'laws of large urns' stating that the distance between the distributions of hypergeometric draws and multinomials draws goes to zero as the urn becomes very large. The same holds for Pólya and multinomial draws. The intuitive reason is that for draws of a fixed size, the removal or addition of drawn balls to the urn does not really matter if the urn is large. Where the second chapter introduces state transformation along a probabilistic channel in a forward direction, this fourth chapter adds observable (predicate) transformation in a backward direction. The two operations are of fundamental importance in program semantics, and also in quantum computation — where they are distinguished as Schrödinger's (forward) and Heisenberg's (backward) approach.

In this context, a random variable is a pair of a distribution and an observable, on the same underlying sample space. The statistical notions of variance and covariance are described in terms of validity for such random variables in Chapter 5. Two forms of covariance are distinghuished, with a 'shared' or a 'joint' state, which satisfy subtly different properties. In this setting the 'law of large draws' appears, describing that the distance goes to zero between large multinomial draws, when normalised, and the urn (as distribution) from which they are drawn. From this metric law the usual formulations of the laws of large numbers are derived.

A very special technique in the area of probability theory is conditioning, also known as belief revision, learning, or simply as updating. It involves the incorporation of evidence into a distribution (state), so that the distribution better fits the evidence. In traditional probability such conditioning is only indirectly available, via a rule $P(B \mid A)$ for computing conditional probabilities, namely of event B given event A. In Chapter 6 we formulate conditioning as an explicit operation, mapping a distribution ω and a predicate p to a new updated distribution $\omega|_p$, called the Bayesian update. In this way we can express that Bayesian updating is commutative: $\omega|_p|_q = \omega|_q|_p$. This leads to an important conclusion: it does not make sense to consider (multiple pieces of) evidence as a list, with a particular order, since the order of Bayesian updating is irrelevant. Thus, evidence is best considered as a multiset.

Another crucial property of Bayesian updating is the inequality $(\omega|_p \models p) \geq (\omega \models p)$. It expresses that the validity of the evidence predicate p is higher in the updated distribution $\omega|_p$ than in the original distribution ω. Informally we shall say: learning (updating) makes you wiser. This inequality cannot be expressed when distributions are left implicit — and indeed, it does not occur in the traditional literature. The same holds for the commutativity equation $\omega|_p|_q = \omega|_q|_p$. Thus, a proper treatment of probabilistic updating requires an update of language. We shall see that updating of distributions satisfies Bayes' rule, for fuzzy evidence. The combination with forward and backward transformation along a channel leads to the powerful techniques of forward inference (causal reasoning) and backward inference (evidential reasoning). These inference techniques are illustrated in many examples, including Bayesian networks and hidden Markov models.

A channel from X to Y is a probabilistic computation, turning elements of X into distributions on Y. Interestingly, such a channel can be reversed — under suitable circumstances — giving a channel / computation from Y to X. This corresponds to turning a conditional probability $P(y \mid x)$ into $P(x \mid y)$, essentially via Bayes' rule. Such reversal is also called Bayesian inversion and will be described here using string diagrams and 'daggers' of channels. In fact, there is a more general mechanism called disintegration that allows us to 'bend wires around'. It is the topic of Chapter 7. These reversed, dagger channels turn out to be important for basic tools in machine learning (see *e.g.* Bishop and Bishop (2024)), such as naive Bayesian classification, decision trees and expectation maximisation. Moreover, they form the basis of a second update rule. The standard rule, based on backward inference, will be called Pearl's rule, and the alternative new rule is called Jeffrey's rule. These two rules are quite different, but their differences are poorly understood. What we do offer is a mathematical characterisation: Pearl's rule increases validity and Jeffrey's rule decreases divergence. More informally, one learns via Pearl's rule by improving what's going well and via Jeffrey's rule by reducing what's going wrong. Jeffrey's rule is thus an error correction mechanism. This fits the basic idea in predictive coding theory Hohwy (2013); Clark (2016) that the human mind is seen as a Bayesian prediction engine that operates by reducing prediction errors via updating. Along the same lines, from (Goodman and Tenenbaum, 2016, §3): "Much of cognition can be understood in terms of conditional inference", see also Griffiths et al. (2024). Indeed, updating of probability distributions is at the heart of computational neuroscience Pouget et al. (2013).

One of the themes running through this book is how 'crossover' influence can be captured via channels — extracted from joint distributions via disintegration — in particular via forward and backward inference. This phenomenon is what makes (reasoning in) Bayesian networks possible. Disintegration is of interest in itself, but also provides an intuitive formalisation of the Bayesian inversion of a channel. At this stage we like to quote (Kallenberg, 2021, Ch. 8).

> Modern probability theory can be said to begin with the notions of conditioning and disintegration.

Disintegration is incorporated into the language of string diagrams via 'shaded' normalisation boxes. It turns out that there are suitable rules for manipulating diagrams with such boxes so that disintegration can be applied in a compositional manner to entire diagrams — especially to Bayesian networks. This is also part of Chapter 7.

Updating is described in Chapter 6 in terms of a single piece of evidence that is incorporated into a distribution. Things become more interesting when there are multiple pieces of evidence. This is the topic of Chapter 8. Suppose there is some disease with a certain prevalence and a

test for the disease, which is not perfect, as usual — expressed in terms of its sensitivity and specificity. Suppose three tests have been done, say with two positive and one negative outcome. How do we then compute the posterior disease probability? The information that is missing so far is whether the three tests were performed on the same individual, or on different people. This matters for how one computes the posterior. In this setting the different update mechanisms associated with Pearl and Jeffrey show up again. The evidence that is used is described as a multiset of factors (or predicates), like in the two positive-test and one negative-test illustration. With such a multiset we shall associate different forms of validity, according to Pearl and according to Jeffrey, with corresponding forms of updating — and corresponding increases of validity. These two forms of validity and updating occur in the literature, as will be illustrated, but are not clearly identified and distinguished. Making explicit which form of updating is used under which circumstances is important for the resulting guarantees — in the form of suitable validity increases — and thus for reasoning about such updates, for instance for explainable and trustworthy AI — including probabilistic guarantees Dalrymple et al. (2024); Tull et al. (2024).

This same chapter shows how updating is used not only for distributions but also for channels. This happens in particular in a basic technique in machine learning, called expectation maximisation (EM). It is shown how the updating of a channel (in EM) can be described systematically in terms of 'double daggers', that is, in terms of double reversals of the original, prior channel. The updating of the distribution (in EM) goes via Jeffrey's rule. The same updating of both a channel and a distribution happens in learning for (hidden) Markov models.

People — especially scientists — like to understand the world in terms of causes and effects. This is difficult. Often we only have statistical data in which one can find correlations. Causality and correlation are quite different things. Moreover, these correlations may be 'polluted' by confounding factors. Judea Pearl has developed a technique for removing such confounding influences (or 'backdoors') so that a channel remains that can be considered to provide the direct causal influence. Chapter 9 gives an impression of this approach, in terms of string diagrams, using what is called diagrammatic surgery — to remove the unwanted parts. General principles are described, but the main goal of this chapter is to convey the basic idea by elaborating several examples. These same surgery techniques can be used for what is called counterfactual reasoning. Typically, counterfactual arguments takes the 'had' form describing something that did not happen, as in: had the author been more disciplined, the book would have been ready in time. Counterfactual reasoning can also be done diagrammatically, by suitably connecting two string diagrams: one representing the factual situation and one for the counterfactual world.

The last chapter using *discrete* probability theory is Chapter 10. It deals with probabilistic automata — of which (hidden) Markov models are examples. It also has the aim to give an impression of an area and it has no claim to completeness. The chapter describes probabilistic automata as instances of the notion of coalgebra. This enables a systematic approach, using (again) the language of category theory. For instance, the (logical) notions of invariant and bisimulation are introduced in such a manner. An invariant is a predicate that, once it holds in a particular position, continues to hold in all successor positions. Invariants are used for safety properties. Bisimulations are relations that are similarly closed under transitions. The chapter concentrates on bisimulations in the form behaviour (pseudo)metrics, which have as key property that observationally indistinguishable states have distance zero.

The final chapter makes the transition from discrete to continuous probability theory. [To be finished.]

This book includes many examples, often copied from familiar sources, with the deliberate aim of illustrating how the channel-based approach actually works. Since many of these examples are taken from the literature, the interested reader may wish to compare the channel-based description used here with the original description. Typically, our aim is to make the underlying mathematical structure in these examples explicit and also the approach that is used to calculate an outcome. Several examples are also illustrated within the probabilistic programming languages ProbLog and WebPPL. Both these language are based on sampling techniques and not on exact computations. Also, they do not use the language of channels that is central in this book. Nevertheless, it is shown that one can recognise channels — and updating and other probabilistic constructions — in these programming languages, so that connections become clear. These examples from probabilistic programming are included in order to bring theory and practice closer together.

Status of the current incomplete version

An incomplete version of this book is made available online, in order to generate feedback and to justify a pause in the writing process. Feedback is most welcome, both positive and negative, especially when it suggests concrete improvements of the text. This may lead to occasional updates of this text. The date on the title page indicates the current version.

Some additional points.

- The (non-trivial) calculations in this book have been carried out with (a follow-up version of) the EfProb library Cho and Jacobs (2017) for channel-based probability. Several calculations in this book can be done by hand, typically when the outcomes are described as fractions, like $\frac{117}{2012}$. Such calculations are meant to be reconstructable by a motivated reader who really wishes to learn the 'mechanics' of the field. Doing such calculations is a great way to really understand the topic — and the approach of this book[2]. Outcomes written in decimal notation 0.1234, as approximations, or as plots, serve to give an impression of the results of a computation.
- For the rest of this book, beyond Chapter 10, only one additional chapter exists in unfinished form, on continuous probability. It will be incorporated in due course.

Bart Jacobs, Nijmegen, May 20, 2025.

[2] Doing the actual calculations can be a bit boring and time consuming, but there are useful online tools for calculating fractions, such as https://www.mathpapa.com/fraction-calculator.html. Calculations in EfProb can also be done fractional form.

1

Collections

The notion of a set X containing elements $x \in X$ is used without further explanation. When we have several elements $x_1, \ldots, x_n \in X$ from the same set X, we can somehow put them together to form a 'collection' of elements. Examples of such collections are subsets, lists, multisets, and discrete probability distributions. The latter distributions form the main topic of this book. But it is useful to first review collections in general, in order to make similarities explicit. For instance, lists, subsets and multisets all form monoids, by suitable unions of collections. Unions of distributions are more subtle and take the form of convex combinations. Also, subsets, multisets and distributions can be combined naturally via parallel products $\otimes$, though lists cannot.

This first chapter reviews several basic constructions and properties of lists, subsets and multisets (also known as bags). Multisets are the least familiar of these collection types. They do not get the attention that they deserve, even though they play a key role in probability theory, for instance as urns filled with coloured balls, as draws from such urns, or as collections of data in learning. For now it suffices that multisets are 'subsets' in which elements may occur multiple times, or 'lists' in which the order of the elements does not matter. We shall use 'ket' notation $|-\rangle$ for multisets, so that $3|R\rangle + 2|G\rangle + 1|B\rangle$ is a multiset that represents an urn with three red balls, two green ones, and one blue ball.

The main differences between lists, multisets and subsets are summarised in the table below.

	lists	**multisets**	**subsets**
order of elements matters	+	-	-
multiplicity of elements matters	+	+	-

For instance, the lists $[a, a, b]$, $[a, b, a]$ and $[a, b]$ are all different, since they involve differences in order and multiplicity (occurrence frequency) of their elements a, b. The multisets $2|a\rangle + 1|b\rangle$ and $1|b\rangle + 2|a\rangle$, with the element a occurring twice and the element b occurring once, are considered to be the same. However, $1|a\rangle + 1|b\rangle$ is a different multiset, since its multiplicities are different. The subsets $\{a, b\}$, $\{b, a\}$, and $\{a\} \cup \{a, b\}$ are all the same, since in a subset we only care about whether an element occurs or not, and not about its order or multiplicity.

We shall formalise the relationship between lists, multisets and subsets as a diagram of the form:

$$\begin{tikzcd} \text{Lists} \arrow[r, Rightarrow, "\text{accumulation}", "\text{forget order}"'] & \text{Multisets} \arrow[r, Rightarrow, "\text{support}", "\text{forget multiplicity}"'] & \text{Subsets} \end{tikzcd} \tag{1.1}$$

The double arrows are used to denote 'natural' transformations, turning a list into a multiset,

and a multiset into a subset. We shall see that the collection types of list, multiset and subset all form 'monads' and that accumulation and support transformations are 'maps of monads' and also 'maps of monoids'. These statements summarise many mathematically relevant facts.

These different properties of different collections are relevant in a probabilistic setting.

1 If one successively learns from multiple data items and the learning method is sensitive to the order of the items, then the data should be organised as a list. If the order in which one learns is irrelevant, the appropriate collection for the data items is a multiset. In general, the multiplicity of occurrences of data items matters, but if not, one can use a *set* of data items.
2 It is a crucial property of a (discrete) probability distribution that its probabilities add up to 1. This property usually relies on combinatorial formulas for basic datatypes. For instance, there are formulas for the number of subsets (or multisets) of size m of a set with $n \geq m$ elements, or for the number of set (or multiset) partitions of a certain kind. These formulas typically involve binomial $\binom{n}{m}$ and multichoose $\left(\!\binom{n}{m}\!\right)$ coefficients, or Stirling numbers of the first or second kind.

The various properties of lists, subsets and multisets that we review in this chapter will be related to basic questions from combinatorics in Section 1.10.

This chapter not only reviews the basic collection types, but also the transformations between them, such as turning a list into a multiset (via 'accumulation') or into a set partition (via 'matching'). Ultimately, the situation is summarised in a triangular prism Diagram (1.52), with many associated combinatorial properties that form the basis for later probability distributions.

The collection types that we review are important in themselves, in many ways, as just illustrated. There is a another, less standard way, in which they will be used, namely as outputs of what we call channels. Such channels are functions of the form $\mathit{input} \to T(\mathit{output})$, where T is a 'collection' operator. It may combine output elements into lists, subsets, multisets, or distributions. Such channels capture a form of computation, directly linked to the kind of collection (of outputs). For instance, channels where T is powerset are used as non-deterministic computations, in which each input element produces a subset of possible output elements. In the probabilistic case these channels produce distributions — for a suitable instantiation of the operator T. Channels will be used as typed computations, which can be used to build more complicated computations via sequential and parallel composition.

As mentioned, this chapter does not contain any probability theory yet. It first exposes the reader to some general considerations that set the scene. This requires some level of patience. Later on it will be useful to see the similarities between probability distributions (in Chapter 2) and other collections, so that constructions, techniques, notation, terminology and intuition that we use for distributions can be put in a wider perspective and thus may become more natural. Instead of fully submerging oneself in this first chapter, one can skim or skip it, move quickly to Chapter 2 on probability distributions, and return later on a call-by-need basis.

The final section of this chapter is more methodological in nature. It explains where the abstractions that we use come from, namely from an area of mathematics called category theory. This area focuses on the structural aspects of the field and makes similarities between various constructions and areas explicit. This last section gives a quick overview of the most relevant parts of this theory and also illustrates how category theory will be used in the remainder of this book, namely in order to present the essentials, often in diagrammatic form. We use category theory pragmatically, as a tool, and not as a goal in itself. No prior knowledge is assumed about category theory. What is needed will be introduced along the way.

1.1 Notation

Much of the (mathematical) notation in this book will be explained at the point where it is introduced. Moreover, at the end, at page **??**, there is an overview of notation. Some of the most basic matters are collected in this section.

We generally use capital letters for sets and small letters for elements. Frequently we use the same letter, as in $x \in X$ and $y \in Y$, so that the name of the element gives an indication of the set that it inhabits.

We write $\mathbb{N} = \{0, 1, 2, \ldots\}$ for the set of natural numbers and $\mathbb{N}_{>0} = \{1, 2, \ldots\}$ for the (sub)set of positive natural numbers. Similarly, $\mathbb{R}$ is the set of real numbers, with subsets $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} \mid x \geq 0\}$ and $\mathbb{R}_{>0} = \{x \in \mathbb{R} \mid x > 0\}$ for non-negative and positive numbers. The unit interval is the subset $[0, 1] = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\} \subseteq \mathbb{R}_{\geq 0}$. We may use a round bracket when the boundary is not included, as in: $(0, 1] = \{x \in \mathbb{R} \mid 0 < x \leq 1\}$. Simililary there are sets $[0, 1)$ and $(0, 1)$.

For a finite set X, say $X = \{x_1, \ldots, x_n\}$ we write $|X| = n$ for its number of elements. We call this number $|X|$ the size of the set X. Sometimes we also write $\mathit{size}(X) = |X|$. We recall that $|X \times Y| = |X| \cdot |Y|$ and $|X + Y| = |X| + |Y|$, where $\times$ is Cartesian product and $+$ is disjoint union (also known as coproduct). Further, $|\mathcal{P}(X)| = 2^{|X|}$, where $\mathcal{P}$ is used for powerset: $\mathcal{P}(X)$ is the set of subsets $U \subseteq X$ of X.

For a number $n \in \mathbb{N}$ we write in bold face $\boldsymbol{n} = \{0, 1, \ldots, n-1\}$ for a canonical set with n elements. Thus $\mathbf{0} = \emptyset$, $\mathbf{1} = \{0\}$ and $\mathbf{2} = \{0, 1\}$. This boldface notation $\boldsymbol{n}$ is short and convenient. For instance, a function of the form $X \to \mathbf{2}$ can be identified with a predicate on the set X, or with a subset of X. Clearly, there is precisely one function $X \to \mathbf{1}$, namely the function that sends every element $x \in X$ to the sole element 0 in $\mathbf{1} = \{0\}$. In the other direction, functions $\mathbf{1} \to X$ can be identified with elements of X.

Sometimes we like to be explicit about whether we start counting from 0 or from 1. In that case we may write:

$$[n) \coloneqq \{0, 1, \ldots, n-1\} = \boldsymbol{n} \qquad \text{and} \qquad (n] \coloneqq \{1, 2, \ldots, n\}.$$

We use the symbol $\coloneqq$ for definitions and typically write it when we introduce new notation.

A function f from a set X to a set Y is written as $f \colon X \to Y$ or as $X \xrightarrow{f} Y$. In that case we call X the domain and Y the codomain of the function f. We include the domain and codomain in the definition of a function and treat them as input and output types. An alternative name for function is 'map' or 'mapping'. We may define a function via the symbol $\mapsto$, that describes which elements are mapped to which. Thus $x \mapsto x + 1$ on $\mathbb{N}$ describes the increment function $\mathit{inc} \colon \mathbb{N} \to \mathbb{N}$ with $\mathit{inc}(x) = x + 1$. For each set X there is an identity function $\mathit{id}_X \colon X \to X$, defined as $\mathit{id}_X(x) = x$. When confusion is unlikely, we omit X from id_X and simply write $\mathit{id} \colon X \to X$.

Given two functions $f \colon X \to Y$ and $g \colon Y \to Z$ we use $g \circ f \colon X \to Z$ for the composite function 'g after f'. It is defined as $(g \circ f)(x) = g(f(x))$. Then $\mathit{id}_Y \circ f = f = f \circ \mathit{id}_X$. We like to be explicit about these elementary matters because later on we introduce channels as 'probabilistic functions'. We will write them with a special arrow $\twoheadrightarrow$ and with their own composition $\circledcirc$.

We often use diagrammatic notation for functions. Consider for instance the two diagrams

below.

$$\begin{tikzcd} & A \arrow[dr, "h"] & \\ X \arrow[ur, "g"] \arrow[rr, "f"'] & & Y \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "f"] \arrow[d, "h"'] & B \arrow[d, "k"] \\ X \arrow[r, "g"'] & Y \end{tikzcd}$$

The triangle on the left expresses the equality of functions $f = h \circ g$. This means $f(x) = h(g(x))$ for each $x \in X$. Similarly, the rectangle on the right expresses the identity $k \circ f = g \circ h$. Commutation of diagrams means that all possible paths in the diagram are equal. Via such diagrams the situation at hand can be clarified visually. Later on we shall use such commuting diagrams also for probabilistic functions (channels), via arrows $\multimap$ with a circle on their shaft. In addition, we will use another graphical formalism for channels, namely string diagrams, with boxes and wires between them.

When two sets X, Y are isomorphic we write $X \cong Y$. This means that there are functions $f\colon X \to Y$ and $g\colon Y \to X$ with $g \circ f = id_X$ and $f \circ g = id_Y$. These maps f, g are then called isomorphisms. One oftens writes $g = f^{-1}$ and where f^{-1} is called the inverse function of f. Similarly, one could write $f = g^{-1}$. We sometimes write $f\colon X \xrightarrow{\cong} Y$ to indicate that f is an isomorphism, with $f^{-1}\colon Y \xrightarrow{\cong} X$ as associated inverse. An isomorphism is also called a bijection or a permutation. For a finite set X, with $n = |X|$ elements, there are $n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$ many bijections $X \xrightarrow{\cong} X$.

We may write a surjective function as $X \twoheadrightarrow Y$ an an injective function as $X \rightarrowtail Y$. A function that is surjective (injective) is also called a surjection (injection). A function of the form $f\colon X \to X$ with the same domain and codomain may be called an endofunction. Such an endofunction is called an idempotent if $f \circ f = f$. An split idempotent is a function $f = s \circ r$, where $r \circ s = id$.

For an arbitrary function $f\colon X \to Y$ and a subset $V \subseteq Y$ we write $f^{-1}(V) := \{x \in X \mid f(x) \in V\}$ for the inverse image. For the special case when V is a singleton $\{y\}$, we simply write $f^{-1}(y)$ for $f^{-1}(\{y\})$. The function f is surjective if and only if all inverse images $f^{-1}(y)$, for $y \in Y$, are non-empty.

For a function $\varphi\colon X \to \mathbb{R}$ there may be a maximum value $\max \varphi \in \mathbb{R}$ and an 'argument maximum' subset $\operatorname{argmax} \varphi \subseteq X$, defined as:

- $\max \varphi$ is the least number $s \in \mathbb{R}_{\geq 0}$ with $\varphi(x) \leq s$, for all $x \in X$; this means that $\max \varphi$ is the supremum of the subset $\{\varphi(x) \mid x \in X\} \subseteq \mathbb{R}$;
- when this maximum / supremum exists, we define the 'argmax' as the subset of arguments where this maximum is actually reached:

$$\operatorname{argmax} \varphi := \{x \in X \mid \varphi(x) = \max \varphi\}.$$

  Sometimes we write $\operatorname{argmax}_{x \in X} \varphi(x)$ for $\operatorname{argmax} \varphi$ to emphasise the variable involved, and its range.

For a finite set X, each function $\varphi\colon X \to \mathbb{R}$ has a maximum; it may be reached for several arguments, so the argmax subset need not be a singleton. For instance, when $X = \{a, b, c, d\}$ and $\varphi(a) = \varphi(c) = 4$, $\varphi(b) = \frac{3}{2}$ and $\varphi(d) = 3$, one has $\max \varphi = 4$ and $\operatorname{argmax} \varphi = \{a, c\}$. When the set X is infinite, a maximum may not exist, like for the inclusion $\mathbb{N} \to \mathbb{R}$. It may happen that the maximum exists, but still the argmax subset is empty, for instance for the function $\mathbb{N} \to \mathbb{R}$ given by $n \mapsto 1 - \frac{1}{n+1}$. Its maximum is 1, but it is never reached.

In a similar manner, one can have a minimum value $\min \varphi$, and an argument minimum subset $\operatorname{argmin} \varphi$.

1.2 Coefficients

This section recalls the basics of various coefficients, commonly written as $\binom{n}{m}$, $\left(\!\binom{n}{m}\!\right)$, $\left[{n \atop m}\right]$ and $\left\{{n \atop m}\right\}$. These numbers play a fundamental role in combinatorics, for counting items in different scenarios, and thereby for associated probabilities. These scenarios will be described later on. Here we only give an overview of the definitions of these coefficients, together with some relevant mathematical properties, for future reference.

For a number $n \in \mathbb{N}$ the *factorial* $n! \in \mathbb{N}_{>0}$ is defined as:

$$n! \coloneqq 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n, \qquad \text{which means:} \qquad \begin{cases} 0! = 1 \\ (n+1)! = (n+1) \cdot n!. \end{cases}$$

If we have n different items, then $n!$ is the number of ways that we can arrange these items in lists of length n.

These factorials are used in the following two constructions.

Definition 1.2.1 Let numbers $n \geq m \geq 0$ be given.

1 The *falling factorial* $(n)_m$ is defined as:

$$(n)_m \coloneqq \frac{n!}{(n-m)!} = n \cdot (n-1) \cdot \ldots \cdot (n-m+1) = \prod_{0 \leq i < m} n - i.$$

The latter formulation shows that we can use the falling factorial notation $(n)_m$ for arbitrary (real) numbers n.

2 The *binomial coefficient* $\binom{n}{m}$ is given by the formula:

$$\binom{n}{m} \coloneqq \frac{n!}{m! \cdot (n-m)!} = \frac{(n)_m}{m!}.$$

3 For numbers $m_1, \ldots, m_K \in \mathbb{N}$ with $n = \sum_i m_i$ there is the *multinomial coefficent* given by:

$$\binom{n}{m_1 \ldots m_K} \coloneqq \frac{n!}{m_1! \cdot \ldots \cdot m_K!}.$$

The falling factorial $(n)_m$ is the number of *lists* containing m elements, each occuring once, out of n given items. The binomial coefficient $\binom{n}{m}$ is the number of *subsets* with m elements, taken from a set of n items. The multinomial coefficient is used for multiple subsets. Later on, in Definition 1.7.1 (5), we shall redefine it for multisets.

Binomial coefficients satisfy a recurrence relation that is called Pascal's rule: for $0 < m \leq n$,

$$\binom{n+1}{m} = \binom{n}{m} + \binom{n}{m-1}. \tag{1.2}$$

Definition 1.2.2 Let $n \geq 1$ and $m \geq 0$.

1 The *rising factorial* $(n)^m$ is defined as:

$$(n)^m \coloneqq \frac{(n+m-1)!}{(n-1)!} = n \cdot (n+1) \cdot \ldots \cdot (n+m-1) = \prod_{0 \leq i < m} n + i.$$

The latter formulation shows that the rising factorial notation can also be used when n is a real number.

2 The *multichoose* coefficient is defined as:

$$\left(\!\binom{n}{m}\!\right) := \frac{(n)^m}{m!} = \frac{(n+m-1)!}{m!\cdot(n-1)!} = \binom{n+m-1}{m}.$$

The rising factorial $(n)^m$ is also written as $n^{\overline{m}}$ or as $n^{(m)}$. We prefer the notation $(n)^m$ for the rising factorial because of its similarity to the falling factorial $(n)_m$. The falling and rising factorials are used in a similar way to define coefficients as $\binom{n}{m} = \frac{(n)_m}{m!}$ and $\left(\!\binom{n}{m}\!\right) = \frac{(n)^m}{m!}$.

As mentioned, $\binom{n}{m}$ is the number of *subsets* of size m contained in a set with n elements. It is less well-known that $\left(\!\binom{n}{m}\!\right)$ is the number of *multisets* of size m over an n-element set. Details appear in Proposition 1.8.7.

The following basic result will be useful later.

Lemma 1.2.3 *Fix a number $m \in \mathbb{N}$. Then:*

$$\lim_{n\to\infty} \frac{\binom{n}{m}}{n^m} = \frac{1}{m!} = \lim_{n\to\infty} \frac{\left(\!\binom{n}{m}\!\right)}{n^m}.$$

Proof. We may assume $n \geq m$. Then:

$$\begin{aligned}\lim_{n\to\infty} \frac{\binom{n}{m}}{n^m} &= \frac{1}{m!}\cdot \lim_{n\to\infty} \frac{n}{n}\cdot\frac{n-1}{n}\cdot\ldots\cdot\frac{n-m+1}{n} \\ &= \frac{1}{m!}\cdot\left(\lim_{n\to\infty}\frac{n}{n}\right)\cdot\left(\lim_{n\to\infty}\frac{n-1}{n}\right)\cdot\ldots\cdot\left(\lim_{n\to\infty}\frac{n-m+1}{n}\right) = \frac{1}{m!}.\end{aligned}$$

Similarly, for the multichoose coefficient:

$$\lim_{n\to\infty} \frac{\left(\!\binom{n}{m}\!\right)}{n^m} = \lim_{n\to\infty} \frac{(n+m-1)!}{m!\cdot(n-1)!\cdot n^m} = \frac{1}{m!}\cdot\lim_{n\to\infty}\frac{n}{n}\cdot\frac{n+1}{n}\cdot\ldots\cdot\frac{n+m-1}{n} = \frac{1}{m!}. \qquad \square$$

1.2.1 Stirling numbers

So-called Stirling numbers are used for counting various objects in combinatorics, see *e.g.* (Graham et al., 1994, §6.1). They exists 'of the first kind' and 'of the second kind', written respectively as $\left[{n \atop m}\right]$ and $\left\{{n \atop m}\right\}$. We briefly recall the essentials of both varieties.

Stirling numbers *of the first kind* are determined by the following equations, for $n \geq 0$ and $m > 0$.

$$\begin{bmatrix}0\\0\end{bmatrix} = 1 \qquad \begin{bmatrix}0\\m\end{bmatrix} = \begin{bmatrix}m\\0\end{bmatrix} = 0 \qquad \begin{bmatrix}n+1\\m\end{bmatrix} = n\cdot\begin{bmatrix}n\\m\end{bmatrix} + \begin{bmatrix}n\\m-1\end{bmatrix} \tag{1.3}$$

The next result collects some basic facts about Stirling numbers of the first kind. The proofs are left as an exercise below.

Lemma 1.2.4 *Let $n \geq 0$.*

1 $\left[{n \atop m}\right] = 0$ *when $m > n$;*
2 $\left[{n \atop n}\right] = 1$*;*
3 $\left[{n+1 \atop 1}\right] = n!$*;*

4 $\left[{n \atop n-1}\right] = \binom{n}{2}$, *when* $n \geq 2$*;*
5 When $n \geq 1$*,*

$$\sum_{1 \leq m \leq n} \begin{bmatrix} n \\ m \end{bmatrix} = n! = \sum_{0 \leq m \leq n} \begin{bmatrix} n \\ m \end{bmatrix}.$$

6 For $r \in \mathbb{R}$*,*

$$\sum_{0 \leq m \leq n} \begin{bmatrix} n \\ m \end{bmatrix} \cdot r^m = (r)^n = r \cdot (r+1) \cdot \ldots \cdot (r+n-1)$$

For $n > 0$ *we may as well restrict the latter summation to* $1 \leq m \leq n$*, since* $\left[{n \atop 0}\right] = 0$ *when* $n > 0$*.* ❑

We turn to Stirling numbers *of the second kind*, written as $\left\{{n \atop m}\right\}$, for $n, m \geq 0$. We use them primarily to count the number of ways to cover an n-element set with m non-empty subsets, see Proposition 1.5.7. At this stage we introduce these Stirling numbers via the following recurrence relations, for $n \geq 0$ and $m > 0$.

$$\begin{Bmatrix} 0 \\ 0 \end{Bmatrix} = 1 \qquad \begin{Bmatrix} m \\ 0 \end{Bmatrix} = \begin{Bmatrix} 0 \\ m \end{Bmatrix} = 0 \qquad \begin{Bmatrix} n+1 \\ m \end{Bmatrix} = m \cdot \begin{Bmatrix} n \\ m \end{Bmatrix} + \begin{Bmatrix} n \\ m-1 \end{Bmatrix}. \tag{1.4}$$

For convenience and future use we collect some basic facts about Stirling numbers of the second kind.

Lemma 1.2.5

1 $\left\{{n \atop m}\right\} = 0$ *when* $m > n$*;*
2 $\left\{{n \atop 1}\right\} = 1$ *when* $n \geq 1$*;*
3 $\left\{{n \atop n}\right\} = 1$ *for all* $n \geq 0$*;*
4 For $n \in \mathbb{N}$ *and* $r \in \mathbb{R}$*,*

$$\sum_{0 \leq m \leq n} \begin{Bmatrix} n \\ m \end{Bmatrix} \cdot (r)_m = r^n \quad \text{where} \quad (r)_m = r \cdot (r-1) \cdot \ldots \cdot (r-m+1). \qquad ❑$$

In Lemma 1.2.4 (5) we saw the summation formula $\sum_{1 \leq m \leq n} \left[{n \atop m}\right] = n!$. The analogue $\sum_{1 \leq m \leq n} \left\{{n \atop m}\right\}$ for Stirling numbers of the second kind produces, for $n = 1, 2, 3, \ldots$, the Bell numbers 1, 2, 5, 15, 52, 203, 877, 4140, ..., see Conway and Guy (1996). They will re-appear later, in Subsection 1.5.4, as the sizes of finite set partitions.

Summations

For convenience we collect some basic results about finite summations below. Infinite summations will appear later on, in Theorem 1.7.4.

Proposition 1.2.6 *We use the following standard summation results, for* $n \in \mathbb{N}$*.*

1 $\displaystyle\sum_{0 \leq i \leq n} i = \frac{n \cdot (n+1)}{2} = \binom{n+1}{2}$.

2 $\displaystyle\sum_{0 \leq i \leq n} i^2 = \frac{n \cdot (n+1) \cdot (2n+1)}{6} = \frac{1}{4} \cdot \left(\!\binom{2n}{6}\!\right)$*; the second equality is (Benjamin and Quinn, 2003, Identity 215).*

3 $\displaystyle\sum_{0\le i\le n} i^3 = \frac{n^2\cdot(n+1)^2}{4}.$

These results can be generalised to K-ary sums, for $K \ge 1$.

4 $\displaystyle\sum_{0\le i_1,\ldots,i_K\le n} i_1+\cdots+i_K = \frac{K\cdot n\cdot(n+1)^K}{2}.$

5 $\displaystyle\sum_{0\le i_1,\ldots,i_K\le n} \big(i_1+\cdots+i_K\big)^2 = \frac{K\cdot n\cdot(n+1)^K\cdot((3K+1)\cdot n+2)}{12}.$

Proof. The first three items are standard, and easy to prove by induction on n. The last two items are consequences, obtained again by induction, now on $K \in \mathbb{N}$. Notice that all of the above sums start at 0; this is relevant for the multiple sums, in (4) and (5). ❑

Exercises

1.2.1 Prove Pascal's rule (1.2).

1.2.2 Show that for $n, m \ne 0$,

$$\binom{n}{m}\cdot m = \binom{n-1}{m-1}\cdot n \qquad \text{and} \qquad \left(\!\!\binom{n}{m}\!\!\right)\cdot m = \left(\!\!\binom{n+1}{m-1}\!\!\right)\cdot n.$$

1.2.3 1 Use this rule of Pascal (1.2) to prove:

$$\sum_{0\le i\le n}\binom{m+i}{m} = \binom{n+m+1}{m+1} = \left(\!\!\binom{n+1}{m+1}\!\!\right).$$

2 Now use the previous equation and (1.2) to prove:

$$\sum_{0\le i\le n}(n+1-i)\cdot\binom{m+i}{m} = \binom{n+m+2}{m+2} = \left(\!\!\binom{n+1}{m+2}\!\!\right).$$

1.2.4 Prove for $n \ge 0$ and $m \ge 1$,

$$\sum_{0\le i\le n}\binom{n}{i} = 2^n \qquad \text{and} \qquad \sum_{0\le i\le m}\left(\!\!\binom{m}{i}\!\!\right) = \binom{2m}{m}.$$

Hint: The second equation follows directly from Exercise 1.2.3 (1).

1.2.5 Prove for $n \ge 0$ that:

$$\sum_{0\le i\le n}\binom{n}{i}\cdot i = n\cdot 2^{n-1} \qquad \text{and} \qquad \sum_{0\le i\le n}\binom{n}{i}\cdot i^2 = n\cdot(n+1)\cdot 2^{n-2} = \binom{n+1}{2}\cdot 2^{n-1}.$$

Induction is not needed. The analogues of these results for multichoose coefficient appear in Exercise 1.8.14.

1.2.6 Let $n_1,\ldots,n_K \in \mathbb{N}$ be given, for $K > 2$, with $n = \sum_i n_i$. Show that the multinomial coefficient can be written as a product of binomial coefficients, via the equation:

$$\binom{n}{n_1,\ldots,n_K} = \binom{n}{n_1}\cdot\binom{n-n_1}{n_2,\ldots,n_K}$$

1.2.7 Prove the statements about Stirling numbers of the first kind that are listed in Lemma 1.2.4.

1.2.8 Prove that for $n \geq m \geq 0$,

$$\begin{bmatrix} n \\ m \end{bmatrix} = \sum_{U \subseteq \{1,\ldots,n\},\, |U| = n-m} \prod_{k \in U} k - 1.$$

1.2.9 Prove that for $1 \leq n \leq N$,

$$\frac{\left[{N \atop n+1} \right]}{(N-1)!} = \sum_{n \leq i < N} \frac{1}{i!} \cdot \begin{bmatrix} i \\ n \end{bmatrix}.$$

1.2.10 Prove also Lemma 1.2.5, about Stirling numbers of the second kind. For item (4) the equation $m \cdot (x)_m + (x)_{m+1} = x \cdot (x)_m$ is useful.

1.2.11 Prove the following equations for Stirling numbers of the second kind, for $n \geq 2$.

1 $\left\{ {n \atop n-1} \right\} = \binom{n}{2}$.

2 $\left\{ {n \atop 2} \right\} = 2^{n-1} - 1$.

1.2.12 Prove Proposition 1.2.6.

1.3 Cartesian products

This section briefly reviews some (standard) terminology and notation related to Cartesian products $\times$ of sets. These products $\times$ are defined not only for sets but also for functions between them.

Let X_1 and X_2 be two arbitrary sets. We can form their Cartesian product $X_1 \times X_2$, as the new set containing all pairs of elements from X_1 and X_2, as in:

$$X_1 \times X_2 := \{(x_1, x_2) \mid x_1 \in X_1 \text{ and } x_2 \in X_2\}.$$

We thus write (x_1, x_2) for the 'pair' or 'tuple' of elements $x_1 \in X_1$ and $x_2 \in X_2$. We have just defined a *binary* product set, constructed from two given sets X_1, X_2. We can also do this in n-ary form, for n sets $X_1, \ldots, X_n$. We then get an n-ary Cartesian product:

$$X_1 \times \cdots \times X_n := \{(x_1, \ldots, x_n) \mid x_1 \in X_1, \ldots, x_n \in X_n\}.$$

The tuple $(x_1, \ldots, x_n)$ is sometimes called an n-tuple. For convenience, it may be abbreviated as a vector $\vec{x}$. The product $X_1 \times \cdots \times X_n$ is sometimes written differently using the symbol $\prod$, as:

$$\prod_{1 \leq i \leq n} X_i \qquad \text{or more informally as:} \qquad \prod X_i.$$

In the latter case it is left implicit what the range is of the index element i.

We allow $n = 0$. The resulting 'empty' product is then written as a singleton set $\mathbf{1}$. For $n = 1$ the product $X_1 \times \cdots \times X_n$ is (isomorphic to) the set X_1.

If one of the sets X_i in a product $X_1 \times \cdots \times X_n$ is empty, then the whole product is empty. Also, if all of the sets X_i are finite, then so is the product $X_1 \times \cdots \times X_n$. In fact, the number of elements of $X_1 \times \cdots \times X_n$ is then obtained by multiplying all the numbers of elements of the sets X_i. Thus:

$$\left| X_1 \times \cdots \times X_n \right| = \left| X_1 \right| \cdot \ldots \cdot \left| X_n \right|,$$

where $| - |$ is used for the size of a finite set, that is, for its number of elements.

1.3.1 Projections and tuples

If we have sets $X_1, \ldots, X_n$ as above, then for each number i with $1 \leq i \leq n$ there is a projection function π_i out of the product to the set X_i, as in:

$$X_1 \times \cdots \times X_n \xrightarrow{\pi_i} X_i \qquad \text{given by} \qquad \pi_i(x_1, \ldots, x_n) := x_i.$$

This gives us functions *out of* a product. We also wish to be able to define functions *into* a product, via tuples of functions: if we have a set Y and n functions $f_1 \colon Y \to X_1, \ldots, f_n \colon Y \to X_n$, then we can form a new function $Y \to X_1 \times \cdots \times X_n$, namely:

$$Y \xrightarrow{\langle f_1, \ldots, f_n \rangle} X_1 \times \cdots \times X_n \qquad \text{via} \qquad \langle f_1, \ldots, f_n \rangle(y) := (f_1(y), \ldots, f_n(y)).$$

There is an obvious result about projecting after tupling of functions:

$$\pi_i \circ \langle f_1, \ldots, f_n \rangle = f_i. \tag{1.5}$$

This is an equality of functions. It can be proven easily by applying both sides to an arbitrary element $y \in Y$.

There are some more 'obvious' equations about tupling of functions:

$$\langle f_1, \ldots, f_n \rangle \circ g = \langle f_1 \circ g, \ldots, f_n \circ g \rangle \qquad \langle \pi_1, \ldots, \pi_n \rangle = id, \tag{1.6}$$

where $g \colon Z \to Y$ is an arbitrary function. In the last equation, id is the identity function on the product $X_1 \times \cdots \times X_n$.

In a Cartesian product we place sets 'in parallel'. We can also place functions between them in parallel. Suppose we have n functions $f_i \colon X_i \to Y_i$. Then we can form the parallel composition:

$$X_1 \times \cdots \times X_n \xrightarrow{f_1 \times \cdots \times f_n} Y_1 \times \cdots \times Y_n$$

via:

$$f_1 \times \cdots \times f_n = \langle f_1 \circ \pi_1, \ldots, f_n \circ \pi_n \rangle$$

so that:

$$(f_1 \times \cdots \times f_n)(x_1, \ldots, x_n) = (f_1(x_1), \ldots, f_n(x_n)).$$

The latter formulation clearly shows how the functions f_i are applied in parallel to the elements x_i.

We overload the product symbol $\times$, since we use it both for sets and for functions. This may be a bit confusing at first, but it is in fact quite convenient.

1.3.2 Powers and exponents

Let X be an arbitrary set. A *power* of X is an n-fold product of X's, for some n. We write the n-th power of X as X^n, in:

$$X^n := \underbrace{X \times \cdots \times X}_{n \text{ times}} = \{(x_1, \ldots, x_n) \mid x_i \in X \text{ for each } i\}.$$

As special cases we have $X^1 = X$ and $X^0 = \mathbf{1}$, a singleton. Since powers are special cases of Cartesian products, they come with projection functions $\pi_i \colon X^n \to X$ and tuple functions

$\langle f_1, \ldots, f_n \rangle \colon Y \to X^n$ for n functions $f_i \colon Y \to X$. Finally, for a function $f \colon X \to Y$ we write $f^n \colon X^n \to Y^n$ for the obvious n-fold parallelisation of f.

There is a *copy* function $\Delta[K] \colon X \to X^K = X \times \cdots \times X$ (K times), given as:

$$\Delta[K](x) := (x, \ldots, x) \quad (K \text{ times } x).$$

We often omit the subscript K, when it is clear from the context, especially when $K = 2$. These Δ functions are alternatively called *copiers* or *diagonals*.

More generally, for two sets X, Y we shall occasionally write:

$$X^Y := \{\, \text{functions } f \colon Y \to X \,\}.$$

This new set X^Y is sometimes called the *function space* or the *exponent* of X and Y. Notice that this exponent notation is consistent with the above notation X^n for powers, since functions $\boldsymbol{n} \to X$ can be identified with n-tuples of elements in X.

These exponents X^Y are related to products in an elementary and useful way, namely via a bijective correspondence:

$$\frac{Z \times Y \xrightarrow{f} X}{\overline{\overline{\phantom{Z \xrightarrow{g} X^Y}}}} \qquad \text{where } Z \xrightarrow{g} X^Y \tag{1.7}$$

This means that for a function $f \colon Z \times Y \to X$ there is a corresponding function $\overline{f} \colon Z \to X^Y$, and vice-versa, for $g \colon Z \to X^Y$ there is a function $\overline{g} \colon Z \times Y \to X$, in such a way that $\overline{\overline{f}} = f$ and $\overline{\overline{g}} = g$. It is not hard to see that we can take $\overline{f}(z) \in X^Y$ to be the function $\overline{f}(z)(y) = f(z, y)$, for $z \in Z$ and $y \in Y$. Similarly, we use $\overline{g}(z, y) = g(z)(y)$. In type theory one says that $\overline{f}$ is obtained from f by Currying, after the logician Haskell Curry.

The correspondence (1.7) is characteristic for so-called Cartesian closed categories.

Exercises

1.3.1 Check what a tuple function $\langle \pi_2, \pi_3, \pi_6 \rangle$ does on a product set $X_1 \times \cdots \times X_8$. What is the codomain of this function?

1.3.2 Check that, in general, the tuple function $\langle f_1, \ldots, f_n \rangle$ is the *unique* function $h \colon Y \to X_1 \times \cdots \times X_n$ with $\pi_i \circ h = f_i$ for each i.

1.3.3 Prove, using Equations (1.5) and (1.6) for tuples and projections, that:

$$(g_1 \times \cdots \times g_n) \circ (f_1 \times \cdots \times f_n) = (g_1 \circ f_1) \times \cdots \times (g_n \circ f_n).$$

1.3.4 Check that the copy function $\Delta \colon X \to X^K$ is 'natural', in the following sense. For each function $f \colon X \to Y$ the following diagram commutes.

$$\begin{tikzcd} X \arrow[r, "\Delta"] \arrow[d, "f"'] & X^K \arrow[d, "f^K"] \\ Y \arrow[r, "\Delta"] & Y^K \end{tikzcd}$$

Check that $\pi_i \circ \Delta = id$, for each $1 \leq i \leq K$.

1.3.5 Define functions in both directions, using tuples and projections, that yield isomorphisms:

$$X \times Y \cong Y \times X \qquad 1 \times X \cong X \qquad X \times (Y \times Z) \cong (X \times Y) \times Z.$$

Try to use Equations (1.5) and (1.6) to prove these isomorphisms, without reasoning with elements.

1.3.6 Similarly, show that exponents satisfy:

$$X^1 \cong X \qquad 1^Y \cong 1 \qquad (X \times Y)^Z \cong X^Z \times Y^Z \qquad X^{Y \times Z} \cong (X^Y)^Z.$$

1.3.7 For $K \in \mathbb{N}$ and sets X, Y define:

$$X^K \times Y^K \xrightarrow{zip[K]} (X \times Y)^K$$

by:

$$zip[K]((x_1, \ldots, x_K), (y_1, \ldots, y_K)) := ((x_1, y_1), \ldots, (x_K, y_K)).$$

1 Show that *zip* is an isomorphism, with inverse function $unzip[K] := \langle (\pi_1)^K, (\pi_2)^K \rangle$.

2 Show that the following diagram commutes.

$$\begin{tikzcd}
 & X^K \times Y^K \arrow[dr, bend left, "\pi_1"] \arrow[d, "zip", "\cong"'] & \\
X \times Y \arrow[ur, bend left, "\Delta \times \Delta"] \arrow[r, bend right, "\Delta"'] & (X \times Y)^K \arrow[r, bend right, "(\pi_1)^K"'] & X^K
\end{tikzcd}$$

1.4 Lists

The datatype of (finite) lists of elements from a given set is well-known in computer science, especially in functional programming. This section collects some basic constructions and properties, especially about the close relationship between lists and monoids. Separately, some of the combinatorial aspects associated with lists are discussed. Lists are also called sequences.

For an arbitrary set X we write $\mathcal{L}(X)$ for the set of all finite lists $[x_1, \ldots, x_n]$ of elements $x_i \in X$, for arbitrary $n \in \mathbb{N}$. Notice that we use square brackets $[-]$ for lists, to distinguish them from tuples, which are typically written with round brackets $(-)$. The difference between tuples and lists is cosmetic.

Thus, the set of lists over X can be defined as a union of all powers X^n of X, as in:

$$\mathcal{L}(X) := \bigcup_{n \in \mathbb{N}} X^n = \coprod_{n \in \mathbb{N}} X^n. \tag{1.8}$$

The fancy sum $\coprod$ is used to emphasise that this is a disjoint union. When the elements of X are seen as letters of an alphabet, then $\mathcal{L}(X)$ is the set of words — the language — over this alphabet. The set $\mathcal{L}(X)$ is alternatively written as $X^\star$, and called the Kleene star of X.

We zoom in on some trivial cases. One has $\mathcal{L}(\mathbf{0}) \cong \mathbf{1}$, since one can only form the empty word $[]$ over the empty alphabet $\mathbf{0} = \emptyset$. If the alphabet contains only one letter, a word consists of a finite number of occurrences of this single letter. Thus: $\mathcal{L}(\mathbf{1}) \cong \mathbb{N}$.

We shall write the length of a list $\ell \in \mathcal{L}(X)$ as $\|\ell\| \in \mathbb{N}$ or sometimes also as $len(\ell)$. It is the number n of elements in a list $\ell = [x_1, \ldots, x_n]$. We shall write this length as function $len \colon \mathcal{L}(X) \to \mathbb{N}$, or also as $\| - \| \colon \mathcal{L}(X) \to \mathbb{N}$. See Exercise 1.4.4 for more details.

Lists over a set X, that is, elements of the set $\mathcal{L}(X)$, collect elements of X in a particular

manner: elements may occur multiple times, and the order of occurrence matters: the lists $[a, b]$ and $[b, a]$ are different. In contrast, in subsets elements may occur at most once and the order of elements does not matter; in multisets elements may occur multiple times, but the order of elements is irrelevant.

Let $f\colon X \to Y$ be an arbitrary function. It can be used to map lists over X to lists over Y by applying the function f elementwise. This is what functional programmers call map-list. It is convenient to overload the notation and apply $\mathcal{L}$ not only to sets but also to functions. Thus we write $\mathcal{L}(f)\colon \mathcal{L}(X) \to \mathcal{L}(Y)$ for the map-list function, defined as:

$$\mathcal{L}(f)\Big([x_1, \ldots, x_n]\Big) \coloneqq [f(x_1), \ldots, f(x_n)].$$

In this way identity maps and compositions are preserved:

$$\mathcal{L}(id) = id \qquad \mathcal{L}(g \circ f) = \mathcal{L}(g) \circ \mathcal{L}(f).$$

We shall say: the operation $\mathcal{L}$ is *functorial*, or simply $\mathcal{L}$ is a *functor*.

Functoriality can be used to define the *marginal* of a list on a product set, via $\mathcal{L}(\pi_i)$, where π_i is a projection map. For instance, let $\ell \in \mathcal{L}(X \times Y)$ be a list of tuples of the form $\ell = [(x_1, y_1), \ldots, (x_n, y_n)]$. The first marginal $\mathcal{L}(\pi_1)(\ell) \in \mathcal{L}(X)$ is then computed as:

$$\mathcal{L}(\pi_1)(\ell) = \mathcal{L}(\pi_1)\Big([(x_1, y_1), \ldots, (x_n, y_n)]\Big) = [\,\pi_1(x_1, y_1), \ldots, \pi_1(x_n, y_n)\,] = [x_1, \ldots, x_n].$$

1.4.1 Monoids

A monoid is a very basic mathematical structure. For convenience we define it explicitly.

Definition 1.4.1 A *monoid* consists of a set M with a binary operation $M \times M \to M$, written for instance as infix $+$, together with an identity element, say $0 \in M$. The binary operation $+$ is required to be associative and must have 0 as identity on both sides. That is, for all $a, b, c \in M$,

$$a + (b + c) = (a + b) + c \qquad \text{and} \qquad 0 + a = a = a + 0.$$

The monoid is called *commutative* when $a + b = b + a$, for all $a, b \in M$. It is called *idempotent* when $a + a = a$ for all elements $a \in M$.

Let $(M, 0, +)$ and $(N, 1, \cdot)$ be two monoids. A function $f\colon M \to N$ is called a *homomorphism of monoids* if f preserves the unit and binary operation, in the sense that:

$$f(0) = 1 \qquad \text{and} \qquad f(a + b) = f(a) \cdot f(b), \quad \text{for all } a, b \in M.$$

Diagrammatically we can express the second equation as:

$$\begin{tikzcd} M \times M \arrow[r, "f \times f"] \arrow[d, "+"'] & N \times N \arrow[d, "\cdot"] \\ M \arrow[r, "f"] & N \end{tikzcd} \tag{1.9}$$

For brevity we also say that such an f is a *map of monoids*, or simply a *monoid map*.

The natural numbers $\mathbb{N}$ with addition form a commutative monoid $(\mathbb{N}, 0, +)$. But also with multiplication they form a commutative monoid $(\mathbb{N}, 1, \cdot)$. The function $f(n) = 2^n$ is a homomorphism of monoids $f\colon (\mathbb{N}, 0, +) \to (\mathbb{N}, 1, \cdot)$.

Various forms of collection types form monoids, with 'union' as binary operation. We start with lists, in the next result. The proof is left as (an easy) exercise to the reader.

Lemma 1.4.2 *1 For each set X, the set $\mathcal{L}(X)$ of lists over X is a monoid, with the empty list $[] \in \mathcal{L}(X)$ as identity element, and with concatenation $+\!\!+ \colon \mathcal{L}(X) \times \mathcal{L}(X) \to \mathcal{L}(X)$ as binary operation:*

$$[x_1, \ldots, x_n] +\!\!+ [y_1, \ldots, y_m] \coloneqq [x_1, \ldots, x_n, y_1, \ldots, y_m].$$

This monoid $(\mathcal{L}(X), [], +\!\!+)$ is neither commutative nor idempotent.

2 For each function $f \colon X \to Y$ the associated map $\mathcal{L}(f) \colon \mathcal{L}(X) \to \mathcal{L}(Y)$ is a homomorphism of monoids. ❑

Thus, lists are monoids via concatenation. But there is more to say: lists are *free* monoids. Informally, this means that $\mathcal{L}(X)$ is the minimal monoid that contains X. This is made precise by the proposition below. We shall occasionally make use of this freeness property and so we like to make it explicit. We shall encounter similar properties for other collection types.

Each element $x \in X$ yields a singleton list, written as $\mathit{unit}(x) \coloneqq [x] \in \mathcal{L}(X)$. The resulting function $\mathit{unit} \colon X \to \mathcal{L}(X)$ plays a special role, see also the next subsection.

Proposition 1.4.3 *Let X be an arbitrary set and let $(M, 0, +)$ be an arbitrary monoid, with a function $f \colon X \to M$. Then there is a unique homomorphism of monoids $\overline{f} \colon (\mathcal{L}(X), [], +\!\!+) \to (M, 0, +)$ with $\overline{f} \circ \mathit{unit} = f$.*

The homomorphism $\overline{f}$ is called the free extension *of f. Its freeness can be expressed via a diagram, as below, where the vertical arrow is dashed, to indicate uniqueness.*

$$\begin{tikzcd} X \arrow[r, "\mathit{unit}"] \arrow[dr, bend right, "f"'] & \mathcal{L}(X) \arrow[d, dashed, "\overline{f},\ \mathit{homomorphism}"] \\ & M \end{tikzcd} \tag{1.10}$$

This diagram expresses $\mathcal{L}(X)$ as the minimal monoid containing X, via the unit map.

Proof. Since $\overline{f}$ preserves the identity element and satisfies $\overline{f} \circ \mathit{unit} = f$ it is determined on empty and singleton lists as:

$$\overline{f}([]) = 0 \qquad \text{and} \qquad \overline{f}([x]) = f(x).$$

Further, on an list $[x_1, \ldots, x_n]$ of length $n \geq 2$ we necessarily have:

$$\overline{f}\big([x_1, \ldots, x_n]\big) = \overline{f}([x_1] +\!\!+ \cdots +\!\!+ [x_n]) = \overline{f}([x_1]) + \cdots + \overline{f}([x_n]) = f(x_1) + \cdots + f(x_n).$$

Thus, there is only one way to define $\overline{f}$. By construction, this $\overline{f} \colon \mathcal{L}(X) \to M$ is a homomorphism of monoids. ❑

The exercises at the end of this section contain illustrations of this freeness result. For future use we introduce the notion of monoid action and the associated homomorphisms.

Definition 1.4.4 Let $(M, 0, +)$ be a monoid.

1 An *action* of the monoid M on a set X is a function $\alpha \colon M \times X \to X$ satisfying:

$$\alpha(0, x) = x \qquad \text{and} \qquad \alpha(a + b, x) = \alpha(a, \alpha(b, x)),$$

for all $a, b \in M$ and $x \in X$.

(Sometimes monoid actions occur of the form $X \times M \to X$, with the monoid as second input; the difference is immaterial.)

2 A *homomorphism* or *map* of monoid actions from $(M \times X \xrightarrow{\alpha} X)$ to $(M \times Y \xrightarrow{\beta} Y)$ is a function $f\colon X \to Y$ satisfing:

$$f(\alpha(a, x)) = \beta(a, f(x)), \qquad \text{for all } a \in M, x \in X.$$

This last equation corresponds to commutation of the following diagram.

$$\begin{tikzcd} M \times X \arrow[r, "id \times f"] \arrow[d, "\alpha"'] & M \times Y \arrow[d, "\beta"] \\ X \arrow[r, "f"] & Y \end{tikzcd}$$

Monoid actions are quite common in mathematics. For instance, scalar multiplication $s \cdot v$ in a vector space forms an action, using the multiplicative monoid structure on scalars: $1 \cdot v = v$ and $s \cdot (s' \cdot v) = (s \cdot s') \cdot v$. Also, as we shall see, probabilistic updating can be described via monoid actions, for shifting and rescaling distributions, but also for updating them. The action map $\alpha\colon M \times X \to X$ can be understood intuitively as pushing the elements in X forward with a quantity from M. It then makes sense that the zero-push is the identity, and that a sum-push is the composition of two individual pushes.

A *group action* is simply a monoid action $\alpha\colon M \times X \to X$ as above where M happens to be a group. Pushes can then be undone, via $\alpha(-a, \alpha(a, x)) = \alpha(-a + a, x) = \alpha(0, x) = x$.

There is also a notion of 'monoid without unit'. It is often called a *semigroup*, and thus consists of a set with (only) an associative operation. A semigroup may be commutative, in an obvious sense, see Exercise 1.4.3. One can also have semigroup actions.

1.4.2 Unit and flatten for lists

We proceed to describe more elementary structure for lists, in terms of special 'unit' and 'flatten' functions. In subsequent sections we shall see that this same structure exists for other collection types, like powerset, multiset and distribution. This unit and flatten structure will turn out to be essential for sequential composition. At the end of this chapter we will see that it is characteristic for what is called a 'monad' in category theory.

We have already seen the singleton-list function $\mathit{unit}\colon X \to \mathcal{L}(X)$, given by $\mathit{unit}(x) := [x]$. There is also a 'flatten' function which turns a list of lists into a list by removing inner brackets. This function is written as $\mathit{flat}\colon \mathcal{L}(\mathcal{L}(X)) \to \mathcal{L}(X)$. It is defined as:

$$\mathit{flat}([[x_{11}, \ldots, x_{1n_1}], \ldots, [x_{k1}, \ldots, x_{kn_k}]]) := [x_{11}, \ldots, x_{1n_1}, \ldots, x_{k1}, \ldots, x_{kn_k}].$$

Alternatively, one can describe flattening via concatenation, see Exercise 1.4.5.

The next result contains some basic properties about unit and flatten. These properties will first be formulated in terms of equations, and then, alternatively as commuting diagrams. The latter style is preferred in this book.

Lemma 1.4.5 *1 For each function $f\colon X \to Y$ one has:*

$$\mathit{unit} \circ f = \mathcal{L}(f) \circ \mathit{unit} \qquad \text{and} \qquad \mathit{flat} \circ \mathcal{L}(\mathcal{L}(f)) = \mathcal{L}(f) \circ \mathit{flat}.$$

Equivalently, the following two diagrams commute.

$$\begin{tikzcd} X \arrow[r, "unit"] \arrow[d, "f"'] & \mathcal{L}(X) \arrow[d, "\mathcal{L}(f)"] \\ Y \arrow[r, "unit"'] & \mathcal{L}(Y) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{L}(\mathcal{L}(X)) \arrow[r, "flat"] \arrow[d, "\mathcal{L}(\mathcal{L}(f))"'] & \mathcal{L}(X) \arrow[d, "\mathcal{L}(f)"] \\ \mathcal{L}(\mathcal{L}(Y)) \arrow[r, "flat"'] & \mathcal{L}(Y) \end{tikzcd}$$

2 *One further has:*

$$flat \circ unit \;=\; id \;=\; flat \circ \mathcal{L}(unit) \qquad\qquad flat \circ flat \;=\; flat \circ \mathcal{L}(flat).$$

These two equations can equivalently be expressed via commutation of:

$$\begin{tikzcd} \mathcal{L}(X) \arrow[r, "unit"] \arrow[dr, equal] & \mathcal{L}(\mathcal{L}(X)) \arrow[d, "flat"] & \mathcal{L}(X) \arrow[l, "\mathcal{L}(unit)"'] \arrow[dl, equal] \\ & \mathcal{L}(X) & \end{tikzcd} \qquad \begin{tikzcd} \mathcal{L}(\mathcal{L}(\mathcal{L}(X))) \arrow[r, "flat"] \arrow[d, "\mathcal{L}(flat)"'] & \mathcal{L}(\mathcal{L}(X)) \arrow[d, "flat"] \\ \mathcal{L}(\mathcal{L}(X)) \arrow[r, "flat"'] & \mathcal{L}(X) \end{tikzcd}$$

Proof. We shall do the first cases of each item, leaving the second cases to the interested reader. First, for $f\colon X \to Y$ and $x \in X$ one has:

$$\begin{aligned} (\mathcal{L}(f) \circ unit)(x) \;=\; \mathcal{L}(f)(unit(x)) \;&=\; \mathcal{L}(f)([x]) \\ &=\; [f(x)] \;=\; unit(f(x)) \;=\; (unit \circ f)(x). \end{aligned}$$

Next, for the second item we take an arbitrary list $[x_1, \ldots, x_n] \in \mathcal{L}(X)$. Then:

$$\begin{aligned} (flat \circ unit)([x_1, \ldots, x_n]) \;&=\; flat([[x_1, \ldots, x_n]]) \\ &=\; [x_1, \ldots, x_n] \\ (flat \circ \mathcal{L}(unit))([x_1, \ldots, x_n]) \;&=\; flat([unit(x_1), \ldots, unit(x_n)]) \\ &=\; flat([[x_1], \ldots, [x_n]]) \\ &=\; [x_1, \ldots, x_n]. \end{aligned}$$

❑

The equations in item 1 of this lemma are so-called naturality properties. They express that *unit* and *flat* work uniformly, independent of the set X involved. The equations in item 2 show that $\mathcal{L}$ is a monad, see Section 1.12 for more information.

The next result connects monoids with the unit and flatten maps.

Proposition 1.4.6 *Let X be an arbitrary set.*

1 *To give a monoid structure $(u, +)$ on X is the same as giving an* $\mathcal{L}$-algebra*, that is, a map $\alpha\colon \mathcal{L}(X) \to X$ satisfying $\alpha \circ unit = id$ and $\alpha \circ flat = \alpha \circ \mathcal{L}(\alpha)$, as in:*

$$\begin{tikzcd} X \arrow[r, "unit"] \arrow[dr, "id"'] & \mathcal{L}(X) \arrow[d, "\alpha"] \\ & X \end{tikzcd} \qquad \begin{tikzcd} \mathcal{L}(\mathcal{L}(X)) \arrow[r, "\mathcal{L}(\alpha)"] \arrow[d, "flat"'] & \mathcal{L}(X) \arrow[d, "\alpha"] \\ \mathcal{L}(X) \arrow[r, "\alpha"] & X \end{tikzcd} \tag{1.11}$$

2 *Let $(M_1, u_1, +_1)$ and $(M_2, u_2, +_2)$ be two monoids, with corresponding $\mathcal{L}$-algebras $\alpha_1\colon \mathcal{L}(M_1) \to M_1$ and $\alpha_2\colon \mathcal{L}(M_2) \to M_2$. A function $f\colon M_1 \to M_2$ is then a homomorphism of monoids if and only if the diagram*

$$\begin{tikzcd} \mathcal{L}(M_1) \arrow[r, "\mathcal{L}(f)"] \arrow[d, "\alpha_1"'] & \mathcal{L}(M_2) \arrow[d, "\alpha_2"] \\ M_1 \arrow[r, "f"] & M_2. \end{tikzcd} \tag{1.12}$$

commutes.

This result says that instead of giving a binary operation + with an identity element u we can give a single operation α that works on all sequences of elements. This is not so surprising, since we can apply the sum multiple times. The more interesting part is that the monoid equations can be captured uniformly by the diagrams / equations (1.11). We shall see that the same diagrams also work for other types of monoids (and collection types).

Proof. 1 If $(X, u, +)$ is a monoid, we can define $\alpha\colon \mathcal{L}(X) \to X$ in one go as $\alpha([x_1, \ldots, x_n]) := x_1 + \cdots + x_n$. The latter sum equals the identity element u when $n = 0$. Notice that the bracketing of the elements in the expression $x_1 + \cdots + x_n$ does not matter, since the binary operation of a monoid is associative. The order does matter, since we do not assume that the monoid is commutative. It is easy to check the equations (1.11). From a more abstract perspective we define $\alpha\colon \mathcal{L}(X) \to X$ via freeness (1.10).

In the other direction, assume an $\mathcal{L}$-algebra $\alpha\colon \mathcal{L}(X) \to X$. We then define an identity element $u := \alpha([]) \in X$ and the sum of $x, y \in X$ as $x + y := \alpha([x, y]) \in X$. We have to check that u is identity for + and that + is associative. This requires some fiddling with the equations (1.11):

$$\begin{aligned} x + u \;=\; \alpha([x, \alpha([])]) &\overset{(1.11)}{=} \alpha([\alpha(\mathit{unit}(x)), \alpha([])]) \\ &= \alpha\big(\mathcal{L}(\alpha)([\,[x], []\,])\big) \\ &\overset{(1.11)}{=} \alpha\big(\mathit{flat}([\,[x], []\,])\big) \;=\; \alpha([x]) \overset{(1.11)}{=} x. \end{aligned}$$

Similarly one shows $u + y = y$. Next, associativity of + is obtained in a similar manner:

$$\begin{aligned} x + (y + z) \;=\; \alpha([x, \alpha([y, z])]) &\overset{(1.11)}{=} \alpha([\alpha(\mathit{unit}(x)), \alpha([y, z])]) \\ &= \alpha\big(\mathcal{L}(\alpha)([\,[x], [y, z]\,])\big) \\ &\overset{(1.11)}{=} \alpha\big(\mathit{flat}([\,[x], [y, z]\,])\big) \\ &= \alpha([\,x, y, z\,]) \\ &= \alpha\big(\mathit{flat}([\,[x, y], [z]\,])\big) \\ &\overset{(1.11)}{=} \alpha\big(\mathcal{L}(\alpha)([\,[x, y], [z]\,])\big) \\ &= \alpha([\alpha([x, y]), \alpha(\mathit{unit}(z))]) \\ &\overset{(1.11)}{=} \alpha([\alpha([x, y]), z]) \;=\; (x + y) + z. \end{aligned}$$

2 Now let $f\colon M_1 \to M_2$ be a homomorphism of monoids. Diagram (1.12) then commutes:

$$\begin{aligned} (\alpha_2 \circ \mathcal{L}(f))([x_1, \ldots, x_n]) &= \alpha_2([f(x_1), \ldots, f(x_n)]) \\ &= f(x_1) + \cdots + f(x_n) \\ &= f(x_1 + \cdots + x_n) \quad \text{since } f \text{ is a homomorphism} \\ &= (f \circ \alpha_1)([x_1, \ldots, x_n]). \end{aligned}$$

In the other direction, if (1.12) commutes for a function $f\colon M_1 \to M_2$, then f is a homomorphism of monoids, since:

$$f(u_1) \;=\; f(\alpha_1([])) \overset{(1.12)}{=} \alpha_2(\mathcal{L}(f)([])) \;=\; \alpha_2([]) \;=\; u_2.$$

Similarly one checks that sums are preserved:

$$f(x +_1 y) \;=\; f(\alpha_1([x, y])) \overset{(1.12)}{=} \alpha_2(\mathcal{L}(f)([x, y])) \;=\; \alpha_2([f(x), f(y)]) \;=\; f(x) +_2 f(y). \qquad \square$$

We see that the algebraic structure $(M, u, +)$ on the set M is expressed as an algebra $\alpha\colon \mathcal{L}(M) \to M$, namely as a certain map to M. This will be a recurring theme in the coming sections.

1.4.3 List combinatorics

Combinatorics is a subarea of mathematics focused on advanced forms of counting. It is relevant for probability theory, since frequencies of occurrences play an important role. We give a first taste of this, using lists.

We use the sum and product of a list of natural numbers in the following way.

$$\begin{aligned} sum\big([n_1, \ldots, n_k]\big) &:= n_1 + \cdots + n_k \;=\; \textstyle\sum_i n_i \\ prod\big([n_1, \ldots, n_k]\big) &:= n_1 \cdot \ldots \cdot n_k \;=\; \textstyle\prod_i n_i. \end{aligned}$$

These functions are free extensions, see Exercise 1.4.7.

We restrict ourselves to lists over the subset $\mathbb{N}_{>0} = \{n \in \mathbb{N} \mid n > 0\}$ of positive natural numbers. Clearly, we obtain restrictions $sum, prod\colon \mathcal{L}(\mathbb{N}_{>0}) \to \mathbb{N}$ of the above sum and product functions.

Now fix $N \in \mathbb{N}_{>0}$. We are interested in lists $\ell \in \mathcal{L}(\mathbb{N}_{>0})$ with $sum(\ell) = N$. These lists are in the inverse image (*i.e.* preimage):

$$sum^{-1}(N) := \{\ell \in \mathcal{L}(\mathbb{N}_{>0}) \;\big|\; sum(\ell) = N\}.$$

For instance, for $N = 4$ this inverse image contains the eight lists:

$$[1,1,1,1] \quad [1,1,2] \quad [1,2,1] \quad [2,1,1] \quad [2,2] \quad [1,3] \quad [3,1] \quad [4]. \tag{1.13}$$

We can interpret the situation as follows. Suppose we have arbitrarily many coins of each value / amount $n \in \mathbb{N}_{>0}$. Then we can ask, for an amount N: how many (ordered) ways are there to lay out the amount N in coins? For $N = 4$ the different layouts are given above, in (1.13).

Here is a first, easy counting result.

Lemma 1.4.7 *For $N \in \mathbb{N}_{>0}$, the subset $sum^{-1}(N) \subseteq \mathcal{L}(\mathbb{N}_{>0})$ has 2^{N-1} elements, that is, $\big|sum^{-1}(N)\big| = 2^{N-1}$.*

Proof. We use induction on N, starting with $N = 1$. Obviously, only the list $[1]$ sums to 1, and for $N = 1$, indeed, $2^{N-1} = 2^0 = 1$.

For the induction step we use the familiar fact that for $K \in \mathbb{N}$,

$$\sum_{0 \le k \le K} \frac{1}{2^k} = \frac{2^{K+1} - 1}{2^K}. \tag{$*$}$$

This can be shown easily by induction on K. Then, for $N > 1$,

$$\begin{aligned}
\left|sum^{-1}(N)\right| &= \left|\{[N]\} \cup \{\ell +\!\!+ [n] \,\middle|\, 1 \leq n \leq N-1 \text{ and } \ell \in sum^{-1}(N-n)\}\right| \\
&= 1 + \sum_{1 \leq n \leq N-1} \left|sum^{-1}(N-n)\right| \\
&\overset{\text{(IH)}}{=} 1 + \sum_{1 \leq n \leq N-1} 2^{N-n-1} \\
&= 1 + \left(\sum_{0 \leq n \leq N-1} 2^{N-n-1}\right) - 2^{N-1} \\
&= 1 + 2^{N-1} \cdot \left(\sum_{0 \leq n \leq N-1} \frac{1}{2^n}\right) - 2^{N-1} \\
&\overset{(*)}{=} 1 + 2^{N-1} \cdot \left(\frac{2^N - 1}{2^{N-1}}\right) - 2^{N-1} \;=\; 1 + 2^N - 1 - 2^{N-1} \;=\; 2^{N-1}.
\end{aligned}$$

❑

Next we describe an elementary fact about coin lists. It has a probabilistic flavour, since it involves what we later call a convex sum of probabilities $r_i \in [0, 1]$ with $\sum_i r_i = 1$. The proof of the first item of this result is postponed, until after Theorem 1.9.9. A proof of the second item using distributions on coin lists may be found in Jacobs (2022c).

Theorem 1.4.8 *Fix $N \in \mathbb{N}_{>0}$.*

1 For each list size $1 \leq n \leq N$,

$$\sum_{\ell \in sum^{-1}(N),\, \|\ell\| = n} \frac{1}{prod(\ell)} \;=\; \frac{n!}{N!} \cdot \begin{bmatrix} N \\ n \end{bmatrix}.$$

2 For arbitrary sizes:

$$\sum_{\ell \in sum^{-1}(N)} \frac{1}{\|\ell\|! \cdot prod(\ell)} \;=\; 1. \tag{1.14}$$

Proof. The proof of the first item is postponed until the very end of Section 1.9, after Theorem 1.9.9. For the second item, we slice up the sum, for each possible list size n and use the previous point:

$$\begin{aligned}
\sum_{\ell \in sum^{-1}(N)} \frac{1}{\|\ell\|! \cdot prod(\ell)} &= \sum_{1 \leq n \leq N} \sum_{\ell \in sum^{-1}(N),\, \|\ell\| = n} \frac{1}{n! \cdot prod(\ell)} \\
&= \sum_{1 \leq n \leq N} \frac{1}{n!} \cdot \frac{n!}{N!} \cdot \begin{bmatrix} N \\ n \end{bmatrix} && \text{by point (1)} \\
&= 1 && \text{by Lemma 1.2.4 (5).}
\end{aligned}$$

❑

We give an illustration of Equation (1.14), for $N = 4$, using the corresponding lists ℓ in (1.13).

The associated sum (1.14) is illustrated below.

$$\underbrace{\begin{array}{cccccccc}
[1,1,1,1] & [1,1,2] & [1,2,1] & [2,1,1] & [2,2] & [1,3] & [3,1] & [4] \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\frac{1}{4!\cdot 1} & \frac{1}{3!\cdot 2} & \frac{1}{3!\cdot 2} & \frac{1}{3!\cdot 2} & \frac{1}{2!\cdot 4} & \frac{1}{2!\cdot 3} & \frac{1}{2!\cdot 3} & \frac{1}{1!\cdot 4} \\
\| & \| & \| & \| & \| & \| & \| & \| \\
\frac{1}{24} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{8} & \frac{1}{6} & \frac{1}{6} & \frac{1}{4}
\end{array}}_{\text{with sum: } 1}$$

Obviously, elements in a list are ordered. Thus, in (1.13) we distinguish between coin layouts $[1,1,2]$, $[1,2,1]$ and $[2,1,1]$. However, when we are commonly discussing which coins add up to 4 we do not take the order into account, for instance in saying that we use two coins of value 1 and one coin of 2, without caring about the order. In doing so, we are not using lists as collection type, but multisets — in which the order of elements does not matter. These multisets form an important alternative collection type; they are discussed from Section 1.6 onwards.

We briefly look at how lists can be turned into subsets, by ignoring multiple occurrences. The *support* of a list is the set of elements occurring in the list. Thus:

$$supp[K]([x_1,\ldots,x_K]) := \{x_1,\ldots,x_K\}. \tag{1.15}$$

For instance, $supp[5]([a,a,b,c,b]) = \{a,b,c\}$. Clearly, this support map ignores the multiplicity of occurrences of elements in its input list. We can describe support as a function $supp[K]\colon X^K \to \mathcal{P}(X)$, where $\mathcal{P}$ is used for powerset. The parameter K in $supp[K]$, used for the length of the list, will be ommitted when it is clear from the context.

We can look at the sizes of inverse images of the support map. These sizes describe the number of lists of a certain length that map to a given non-empty, finite set U of a certain size. The formula involves Striling numbers $\left\{ {K \atop n} \right\}$ of the second kind, see Section 1.2. The proof of the next result is postponed. It will appear as a result of a more basic result about Stirling numbers, see Remark 1.5.11 (2).

Lemma 1.4.9 *Fix a number $K \in \mathbb{N}_{>0}$. For a non-empty set U with $|U| \leq K$ the number of lists of length K with U as support is:*

$$\left| supp[K]^{-1}(U) \right| = |U|! \cdot \left\{ {K \atop |U|} \right\}. \tag{1.16}$$

For instance, for the set $U = \{a,b\}$ and number $K = 4$ there are $|U|!\cdot\left\{ {K \atop |U|} \right\} = 2!\cdot\left\{ {4 \atop 2} \right\} = 2\cdot 7 = 14$ lists of length 4 with support U, namely:

$$\begin{array}{ccccccc}
[a,a,a,b] & [a,a,b,a] & [a,a,b,b] & [a,b,a,a] & [a,b,a,b] & [a,b,b,a] & [a,b,b,b] \\
[b,a,a,a] & [b,a,a,b] & [b,a,b,a] & [b,a,b,b] & [b,b,a,a] & [b,b,a,b] & [b,b,b,a].
\end{array}$$

Exercises

1.4.1 Let $X = \{a,b,c\}$ and $Y = \{u,v\}$ be sets with a function $f\colon X \to Y$ given by $f(a) = u = f(c)$ and $f(b) = v$. Write $\ell_1 = [c,a,b,a]$ and $\ell_2 = [b,c,c,c]$. Compute consecutively:

- $\ell_1 +\!\!+ \ell_2$
- $\ell_2 +\!\!+ \ell_1$

- $\ell_1 +\!\!+ \ell_1$
- $\ell_1 +\!\!+ (\ell_2 +\!\!+ \ell_1)$
- $(\ell_1 +\!\!+ \ell_2) +\!\!+ \ell_1$
- $\mathcal{L}(f)(\ell_1)$
- $\mathcal{L}(f)(\ell_2)$
- $\mathcal{L}(f)(\ell_1) +\!\!+ \mathcal{L}(f)(\ell_2)$
- $\mathcal{L}(f)(\ell_1 +\!\!+ \ell_2)$.

1.4.2 We write ln for the natural logarithm function, so that $\ln(x) = y$ iff $x = e^y$. Verify that the logarithm function ln and e-power form an isomorphism of monoid:

$$\begin{tikzcd} (\mathbb{R}_{>0}, 1, \cdot) \arrow[r, bend left, "\ln"] & (\mathbb{R}, 0, +) \arrow[l, bend left, "e^{(-)}"] \end{tikzcd}$$

These two maps form in fact an isomorphism of groups. Check that this works as well for any other logarithm function $\log_a$ with basis a, instead of e.

Often the ln function is used to simplify a computation, by turning multiplications into additions. Then one uses that ln is precisely this homomorphism of monoids. An additonal useful property is that ln preserves and reflects the order: $x \leq y$ iff $\ln(x) \leq \ln(y)$.

1.4.3 Consider the 'root of squares' function $rs\colon \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ defined by:

$$rs(x, y) \;=\; \sqrt{x^2 + y^2}.$$

Show that this yields a commutative semigroup structure on $\mathbb{R}_{>0}$.

1.4.4 Define a length function $\| - \|\colon \mathcal{L}(X) \to \mathbb{N}$ on lists via the freeneess property of Proposition 1.4.3.

1 Describe $\|\ell\|$ for $\ell \in \mathcal{L}(X)$ explicitly.
2 Elaborate what it means that $\| - \|$ is a homomorphism of monoids.
3 Write ! for the unique function $X \to \mathbf{1}$ and check that $\| - \|$ is $\mathcal{L}(!)$. Notice that the previous item can then be seen as an instance of Lemma 1.4.2 (2).

1.4.5 1 Check that the list-flatten operation $\mathcal{L}(\mathcal{L}(X)) \to \mathcal{L}(X)$ can be described in terms of concatenation $+\!\!+$ as:

$$\mathit{flat}([\ell_1, \ldots, \ell_n]) \;=\; \ell_1 +\!\!+ \cdots +\!\!+ \ell_n.$$

2 Now consider the correspondence of Proposition 1.4.6 (1). Conclude that the algebra $\mathcal{L}(\mathcal{L}(X)) \to \mathcal{L}(X)$ associated with the monoid $(\mathcal{L}(X), [], +\!\!+)$ from Lemma 1.4.2 (1) is *flat*.

1.4.6 Check Equation (1.14) yourself for $N = 5$.

1.4.7 Consider the set $\mathbb{N}$ of natural numbers with its *additive* monoid structure $(0, +)$ and also with its *multiplicative* monoid structure $(1, \cdot)$. Apply freeness from Proposition 1.4.3 with these two structures to define two monoid homomorphisms:

$$\begin{tikzcd} \mathcal{L}(\mathbb{N}) \arrow[r, bend left, "sum"] \arrow[r, bend right, "prod"'] & \mathbb{N} \end{tikzcd}$$

1 Describe these maps explicitly on a sequence $[n_1, \ldots, n_k]$ of natural numbers n_i.
2 Make explicit what it means that they preserve the monoid structure.

3 Prove that for an arbitrary set X, the list-length function $\|-\|$ from Exercise 1.4.4 satisfies:

$$sum \circ \mathcal{L}(\| - \|) = \| - \| \circ flat.$$

In other words, the following diagram commutes.

$$\begin{tikzcd} \mathcal{L}(\mathcal{L}(X)) \arrow[r, "\mathcal{L}(\|-\|)"] \arrow[d, "flat"'] & \mathcal{L}(\mathbb{N}) \arrow[d, "sum"] \\ \mathcal{L}(X) \arrow[r, "\|-\|"] & \mathbb{N} \end{tikzcd}$$

A fancy way to prove that length is such an algebra homomorphism is to use the uniqueness in Proposition 1.4.3.

1.4.8 Let $M = (M, +, 0)$ be an arbitrary monoid.

1 We use the natural numbers $\mathbb{N}$ as additive monoid. Check that there is a monoid action $\mathbb{N} \times M \to M$, given by $(n, a) \mapsto n \cdot a = a + \cdots + a$, n times.

2 Check also that M acts on itself: the addition $M \times M \to M$ is an action of M (as monoid) on M (as set).

1.4.9 Let X be an arbitrary set.

1 Show that the set X^X of functions $X \to X$ forms a monoid, via composition, with the identity function as identity element.

2 Let M be a monoid. For a function $\alpha\colon X \times M \to X$ define $\overline{\alpha}\colon M \to X^X$ by $\alpha(a)(x) = \alpha(x, a)$. Show that α is a monoid action if and only if $\overline{\alpha}$ is a map of monoids.

3 Let $X \times Y \to Y$ be an arbitrary function. Turn it into a monoid action $\mathcal{L}(X) \times Y \to Y$, either via direct definition, or via freeness, using the previous point.

1.5 Subsets

The next collection type that will be studied is powerset. The symbol $\mathcal{P}$ is commonly used for the powerset operator. We will see that there are many similarities with lists $\mathcal{L}$ from the previous section. We again pay attention to monoid structures and to basic combinatorial facts.

For an arbitrary set X we write $\mathcal{P}(X) := \{U \mid U \subseteq X\}$ for the set of *all* subsets of X. In addition, we use several variations:

$$\begin{aligned} \mathcal{P}_*(X) &:= \{U \in \mathcal{P}(X) \mid U \neq \emptyset\} \\ \mathcal{P}_{fin}(X) &:= \{U \in \mathcal{P}(X) \mid U \text{ is finite}\} \\ \mathcal{P}[K](X) &:= \{U \in \mathcal{P}_{fin}(X) \mid |U| = K\} \\ \mathcal{P}_{\leq}[K](X) &:= \{U \in \mathcal{P}_{fin}(X) \mid 1 \leq |U| \leq K\}. \end{aligned} \tag{1.17}$$

If X is a finite set itself, there is no difference between $\mathcal{P}(X)$ and $\mathcal{P}_{fin}(X)$. In the sequel we shall speak mostly about $\mathcal{P}$, but basically all properties of interest hold for $\mathcal{P}_{fin}$ as well.

First of all, $\mathcal{P}$ is a functor: it works both on sets and on functions. Given a function $f\colon X \to Y$ we can define a new function $\mathcal{P}(f)\colon \mathcal{P}(X) \to \mathcal{P}(Y)$ by taking the image of f on a subset. Explicitly, for $U \subseteq X$,

$$\mathcal{P}(f)(U) = \{f(x) \mid x \in U\}.$$

The right-hand side is clearly a subset of Y, and thus an element of $\mathcal{P}(Y)$. We have two equalities:

$$\mathcal{P}(id) \;=\; id \qquad\qquad \mathcal{P}(g \circ f) \;=\; \mathcal{P}(g) \circ \mathcal{P}(f).$$

We can use functoriality for marginalisation: for a subset (relation) $R \subseteq X \times Y$ on a product set we get its first marginal $\mathcal{P}(\pi_1)(R) \in \mathcal{P}(X)$ as the subset:

$$\mathcal{P}(\pi_1)(R) \;=\; \{\pi_1(z) \mid z \in R\} \;=\; \{\pi_1(x,y) \mid (x,y) \in R\} \;=\; \{x \mid \exists y.\, (x,y) \in R\}.$$

The next topic is the monoid structure on powersets, given by finite unions. The first result is an analogue of Lemma 1.4.2; its proof is left to the reader. What is characteristic of powerset monoids is idempotency of its binary operation: $V \cup V = V$.

Lemma 1.5.1 *1 For each set X, the powerset $\mathcal{P}(X)$ is a commutative and idempotent monoid, with empty subset $\emptyset \in \mathcal{P}(X)$ as identity element and union $\cup$ of subsets of X as binary operation.*
2 Each $\mathcal{P}(f)\colon \mathcal{P}(X) \to \mathcal{P}(Y)$ is a map of monoids, for $f\colon X \to Y$.
Both properties also hold for $\mathcal{P}_{\mathrm{fin}}$ instead of $\mathcal{P}$. ❑

Next we define unit and flatten maps for subsets, much like for lists. The function $unit\colon X \to \mathcal{P}(X)$ sends an element to a singleton subset: $unit(x) := \{x\}$. The flatten function $flat\colon \mathcal{P}(\mathcal{P}(X)) \to \mathcal{P}(X)$ is given by union: for $A \subseteq \mathcal{P}(X)$,

$$flat(A) \;:=\; \bigcup A \;=\; \{x \in X \mid \exists U \in A.\, x \in U\}.$$

We mention, without proof, the following analogue of Lemma 1.4.5.

Lemma 1.5.2 *1 For each function $f\colon X \to Y$ the 'naturality' diagrams*

$$\begin{tikzcd} X \arrow[r, "unit"] \arrow[d, "f"'] & \mathcal{P}(X) \arrow[d, "\mathcal{P}(f)"] \\ Y \arrow[r, "unit"'] & \mathcal{P}(Y) \end{tikzcd} \qquad\qquad \begin{tikzcd} \mathcal{P}(\mathcal{P}(X)) \arrow[r, "flat"] \arrow[d, "\mathcal{P}(\mathcal{P}(f))"'] & \mathcal{P}(X) \arrow[d, "\mathcal{P}(f)"] \\ \mathcal{P}(\mathcal{P}(Y)) \arrow[r, "flat"'] & \mathcal{P}(Y) \end{tikzcd}$$

commute.
2 Additionaly, the 'monad' diagrams below commute.

$$\begin{tikzcd} \mathcal{P}(X) \arrow[r, "unit"] \arrow[dr, equal] & \mathcal{P}(\mathcal{P}(X)) \arrow[d, "flat"] & \mathcal{P}(X) \arrow[l, "\mathcal{P}(unit)"'] \arrow[dl, equal] \\ & \mathcal{P}(X) & \end{tikzcd} \qquad\qquad \begin{tikzcd} \mathcal{P}(\mathcal{P}(\mathcal{P}(X))) \arrow[r, "flat"] \arrow[d, "\mathcal{P}(flat)"'] & \mathcal{P}(\mathcal{P}(X)) \arrow[d, "flat"] \\ \mathcal{P}(\mathcal{P}(X)) \arrow[r, "flat"'] & \mathcal{P}(X) \end{tikzcd}$$

❑

1.5.1 From list to powerset

We have seen that lists and subsets behave in a similar manner. The connection can be made explicit via the support function from (1.15), given by $supp([x_1, \ldots, x_n]) = \{x_1, \ldots, x_n\}$. It forms a function $supp\colon \mathcal{L}(X) \to \mathcal{P}_{fin}(X)$.

There is no systematic way to go in the other direction, via a function $\mathcal{P}_{fin}(X) \to \mathcal{L}(X)$ that works for each set X. Of course, one can for each subset *choose* an order of the elements in order to turn the subset into a list. However, this process is completely arbitrary and is not natural. By

the way, we shall see later in Exercise 2.4.17 that one can go from subsets to lists *probabilistically*, by taking a uniform distribution over all lists whose support is a given subset — via the formula in Lemma 1.4.9.

The support function interacts nicely with the structures that we have seen so far. This is expressed in the result below, where we use the same notation *unit* and *flat* for different functions, namely for $\mathcal{L}$ and for $\mathcal{P}$. The context, and especially the type of an argument, will make clear which one is meant.

Lemma 1.5.3 *Consider the support map* $supp\colon \mathcal{L}(X) \to \mathcal{P}_{fin}(X)$*, for a fixed set* X*.*

1 It is a map of monoids $(\mathcal{L}(X), [], +\!\!+) \to (\mathcal{P}_{fin}(X), \emptyset, \cup)$*.*

2 It is natural, in the sense that for each function $f\colon X \to Y$ *one has:*

$$\begin{tikzcd} \mathcal{L}(X) \arrow[r, "supp"] \arrow[d, "\mathcal{L}(f)"'] & \mathcal{P}_{fin}(X) \arrow[d, "\mathcal{P}_{fin}(f)"] \\ \mathcal{L}(Y) \arrow[r, "supp"'] & \mathcal{P}_{fin}(Y) \end{tikzcd}$$

3 It commutes with the unit and flatten maps of list and powerset, as in:

$$\begin{tikzcd} X \arrow[r, equal] \arrow[d, "unit"'] & X \arrow[d, "unit"] \\ \mathcal{L}(X) \arrow[r, "supp"'] & \mathcal{P}_{fin}(X) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{L}(\mathcal{L}(X)) \arrow[r, "\mathcal{L}(supp)"] \arrow[d, "flat"'] & \mathcal{L}(\mathcal{P}_{fin}(X)) \arrow[r, "supp"] & \mathcal{P}_{fin}(\mathcal{P}_{fin}(X)) \arrow[d, "flat"] \\ \mathcal{L}(X) \arrow[rr, "supp"'] & & \mathcal{P}_{fin}(X) \end{tikzcd}$$

Proof. The first item is easy and skipped. For item 2,

$$\begin{aligned} (\mathcal{P}_{fin}(f) \circ supp)([x_1, \ldots, x_n]) &= \mathcal{P}_{fin}(f)(supp([x_1, \ldots, x_n])) \\ &= \mathcal{P}_{fin}(f)(\{x_1, \ldots, x_n\}) \\ &= \{f(x_1), \ldots, f(x_n)\} \\ &= supp([f(x_1), \ldots, f(x_n)]) \\ &= supp(\mathcal{L}(f)([x_1, \ldots, x_n])) \;=\; (supp \circ \mathcal{L}(f))([x_1, \ldots, x_n]). \end{aligned}$$

In item 3, commutation of the first diagram is easy:

$$(supp \circ unit)(x) \;=\; supp([x]) \;=\; \{x\} \;=\; unit(x).$$

The second diagram requires a bit more work. Starting from a list of lists we get:

$$\begin{aligned} &(flat \circ supp \circ \mathcal{L}(supp))([[x_{11}, \ldots, x_{1n_1}], \ldots, [x_{k1}, \ldots, x_{kn_k}]]) \\ &= (\textstyle\bigcup \circ supp)([supp([x_{11}, \ldots, x_{1n_1}]), \ldots, supp([x_{k1}, \ldots, x_{kn_k}])]) \\ &= (\textstyle\bigcup \circ supp)([\{x_{11}, \ldots, x_{1n_1}\}, \ldots, \{x_{k1}, \ldots, x_{kn_k}\}]) \\ &= \textstyle\bigcup(\{\{x_{11}, \ldots, x_{1n_1}\}, \ldots, \{x_{k1}, \ldots, x_{kn_k}\}\}) \\ &= \{x_{11}, \ldots, x_{1n_1}, \ldots, x_{k1}, \ldots, x_{kn_k}\} \\ &= supp([x_{11}, \ldots, x_{1n_1}, \ldots, x_{k1}, \ldots, x_{kn_k}]) \\ &= (supp \circ flat)([[x_{11}, \ldots, x_{1n_1}], \ldots, [x_{k1}, \ldots, x_{kn_k}]]). \end{aligned}$$

❑

1.5.2 Finite powersets and idempotent commutative monoids

We briefly dive a bit deeper into the relation between monoids and (finite) powersets from Lemma 1.5.1. At an abstract level the situation is much like for lists, as described in Subsection 1.4.1. For instance, the commutative idempotent monoids $\mathcal{P}_{fin}(X)$ are free, like lists, in Proposition 1.4.3.

Proposition 1.5.4 *Let X be a set and $(M, 0, +)$ a commutative idempotent monoid, with a function $f\colon X \to M$ between them. Then there is a unique homomorphism of monoids of the form $\overline{f}\colon (\mathcal{P}_{fin}(X), \emptyset, \cup) \to (M, 0, +)$ with $\overline{f} \circ \mathsf{unit} = f$. We represent this situation as:*

$$\begin{tikzcd} X \arrow[r, "unit"] \arrow[dr, bend right, "f"'] & \mathcal{P}_{fin}(X) \arrow[d, dashed, "\overline{f},\ homomorphism"] \\ & M \end{tikzcd} \tag{1.18}$$

Proof. Given the requirements, the only way to define $\overline{f}$ is as:

$$\overline{f}\big(\{x_1, \ldots, x_n\}\big) := f(x_1) + \cdots + f(x_n), \qquad \text{with special case} \qquad \overline{f}(\emptyset) = 0.$$

The order in the above sum $f(x_1) + \cdots + f(x_n)$ does not matter since M is commutative. The function $\overline{f}$ sends unions to sums since $+$ is idempotent. ❑

Commutative idempotent monoids can be described as algebras, in analogy with Proposition 1.4.6.

Proposition 1.5.5 *Let X be an arbitrary set.*

1 *To specify a commutative idempotent monoid structure $(u, +)$ on X is the same as giving a $\mathcal{P}_{fin}$-algebra $\alpha\colon \mathcal{P}_{fin}(X) \to X$, namely so that the diagrams*

$$\begin{tikzcd} X \arrow[r, "unit"] \arrow[dr, "id"'] & \mathcal{P}_{fin}(X) \arrow[d, "\alpha"] \\ & X \end{tikzcd} \qquad\qquad \begin{tikzcd} \mathcal{P}_{fin}(\mathcal{P}_{fin}(X)) \arrow[r, "\mathcal{P}_{fin}(\alpha)"] \arrow[d, "flat"'] & \mathcal{P}_{fin}(X) \arrow[d, "\alpha"] \\ \mathcal{P}_{fin}(X) \arrow[r, "\alpha"] & X \end{tikzcd} \tag{1.19}$$

commute.

2 *Let $(M_1, u_1, +_1)$ and $(M_2, u_2, +_2)$ be two commutative idempotent monoids, with corresponding $\mathcal{P}_{fin}$-algebras $\alpha_1\colon \mathcal{P}_{fin}(M_1) \to M_1$ and $\alpha_2\colon \mathcal{P}_{fin}(M_2) \to M_2$. A function $f\colon M_1 \to M_2$ is a map of monoids if and only if the rectangle*

$$\begin{tikzcd} \mathcal{P}_{fin}(M_1) \arrow[r, "\mathcal{P}_{fin}(f)"] \arrow[d, "\alpha_1"'] & \mathcal{P}_{fin}(M_2) \arrow[d, "\alpha_2"] \\ M_1 \arrow[r, "f"] & M_2. \end{tikzcd} \tag{1.20}$$

commutes.

Proof. This works very much like in the proof of Proposition 1.4.6. If $(X, u, +)$ is a monoid, we define $\alpha\colon \mathcal{P}_{fin}(X) \to X$ by freeness as $\alpha(\{x_1, \ldots, x_n\}) := x_1 + \cdots + x_n$. In the other direction, given $\alpha\colon \mathcal{P}_{fin}(X) \to X$ we define a sum as $x + y := \alpha(\{x, y\})$ with unit $u := \alpha(\emptyset)$. Clearly, the sum $+$ on X is commutative and idempotent. ❑

This result concentrates on the finite powerset functor $\mathcal{P}_{fin}$. One can also consider algebras $\mathcal{P}(X) \to X$ for the (general) powerset functor $\mathcal{P}$. Such algebras turn the set X into a complete lattice, see Barr and Wells (1985); Johnstone (1982); Manes (1974) for details.

1.5.3 Extraction

So far we have emphasised the similarity between lists and subsets: the only structural difference that we have seen up to now is that subsets form an *idempotent* and *commutative* monoid. But there are other important differences. Here we look at subsets of product sets, also known as relations.

The observation is that one can extract functions from a relation $R \subseteq X \times Y$ two functions of the form $extr_1(R)\colon X \to \mathcal{P}(Y)$ and $extr_2(R)\colon Y \to \mathcal{P}(X)$. They are given by:

$$extr_1(R)(x) := \{y \in Y \mid (x, y) \in R\} \qquad\qquad extr_2(R)(y) := \{x \in X \mid (x, y) \in R\}.$$

In fact, one can easily reconstruct the relation R from $extr_1(R)$, and also from $extr_2(R)$, via:

$$R = \{(x, y) \mid y \in extr_1(R)(x)\} = \{(x, y) \mid x \in extr_2(R)(y)\}.$$

This all looks rather trivial, but such function extraction is less trivial for other data types, as we shall see later on for distributions, where it will be called disintegration, see Section 7.2.

Using the exponent notation from Subsection 1.3.2 we can summarise the situation as follows. There are two isomorphisms:

$$\mathcal{P}(Y)^X \cong \mathcal{P}(X \times Y) \cong \mathcal{P}(X)^Y. \tag{1.21}$$

Functions of the form $X \to \mathcal{P}(Y)$ will later be called 'channels' from X to Y, see Section 1.11. What we have just seen will then be described in terms of 'extraction of channels'.

There are *no* isomorphisms $\mathcal{L}(Y)^X \cong \mathcal{L}(X \times Y) \cong \mathcal{L}(X)^Y$ for lists.

1.5.4 Set partitions

In this text we use two kinds of partitions, namely *set* partitions and *multiset* partitions. The latter will appear in Definition 1.9.1. Here we introduce set partitions. They are also known as disjoint covers: collections of pairwise disjoint subsets that cover the set. They correspond in a standard way to equivalence relations, see Exercise 1.5.4.

Definition 1.5.6 1 For a non-empty set X we write we write $SP(X)$ for the set of set partitions of X:

$$SP(X) := \Big\{ P \subseteq \mathcal{P}(X) \,\Big|\, \bigcup P = X \text{ and } \forall B \in P.\, B \neq \emptyset \text{ and } \forall B, B' \in P.\, B \neq B' \Rightarrow B \cap B' = \emptyset \Big\}. \tag{1.22}$$

The subsets $B \in P$ of X, for a partition $P \in SP(X)$, are often called *blocks*.

2 For a number $K \geq 1$ we simply write $SP(K) = SP\big(\{1, \ldots, K\}\big)$ for the set of partitions of the set $(K] = \{1, \ldots, K\}$ of first K positive natural numbers.

This set of partitions comes with a function $size[K]\colon SP(K) \to (K]$, where $size[K](P) := |P|$ is the number of blocks in a partition $P \in SP(K)$. The parameter K in $size[K]$ is omitted if it is clear from the context.

We shall simply call an element of $SP(K)$ a 'set partition of K', instead of a 'set partition of $\{1,\dots,K\}$'. For instance, the set $SP(3)$ contains the following five partitions of 3.

$$\big\{\{1\},\{2\},\{3\}\big\} \qquad \big\{\{1,2\},\{3\}\big\} \qquad \big\{\{1,3\},\{2\}\big\} \qquad \big\{\{1\},\{2,3\}\big\} \qquad \big\{\{1,2,3\}\big\}.$$

The set partitions of 4 of size 3, that is the partitions in the subset $size[4]^{-1}(3)$, are:

$$\begin{array}{lll} \big\{\{1\},\{2\},\{3,4\}\big\} & \big\{\{1\},\{2,4\},\{3\}\big\} & \big\{\{1\},\{2,3\},\{4\}\big\} \\ \big\{\{1,4\},\{2\},\{3\}\big\} & \big\{\{1,3\},\{2\},\{4\}\big\} & \big\{\{1,2\},\{3\},\{4\}\big\}. \end{array}$$

Stirling numbers of the second kind give the size of such subsets. This may be seen as the defining property of these numbers.

Proposition 1.5.7 *Let $K \geq 1$ and $1 \leq n \leq K$. The number of set partitions in $SP(K)$ of size n is given by the Stirling number $\left\{ {K \atop n} \right\}$ of the second kind:*

$$\Big|size[K]^{-1}(n)\Big| = \left\{ {K \atop n} \right\} \qquad \textit{where} \qquad size[K]\colon SP(K) \longrightarrow \{1,\dots,K\}.$$

Proof. We show, by induction on $K \geq 1$, that the size expression $\Big|size[K]^{-1}(n)\Big|$ satisfies the recurrence equations (1.4), where we concentrate on the situation $K \geq 1$ and $1 \leq n \leq K$. When $K = 1$ we have $n = 1$ and $\{\{1\}\}$ as only partition of $\{1\}$, so that:

$$\Big|size[1]^{-1}(1)\Big| = 1 = \left\{ {1 \atop 1} \right\}.$$

Next we consider a set partition P of $\{1,\dots,K+1\}$. When $size(P) = 1$, it follows that P is the singleton partition containing the whole set $\{1,\dots,K+1\}$. Thus:

$$\Big|size[K+1]^{-1}(1)\Big| = 1 = \left\{ {K+1 \atop 1} \right\}, \qquad \text{by Lemma 1.2.5 (2).}$$

We continue with $n > 1$. For a set partition P of $\{1,\dots,K+1\}$ there are two cases: the number $K+1$ is in a singleton block in P, or it is part of a non-singlteton block. Thus, either:

- $P = Q \cup \{\{K+1\}\}$, where $Q \in SP(K)$; then $size(P) = size(Q) + 1$;
- or $P = (Q \setminus \{B\}) \cup \{B \cup \{K+1\}\}$, where $Q \in SP(K)$ and $B \in Q$. In this case we have $size(P) = size(Q)$. There are $size(P)$ many ways in which this may happen, namely for each block in P.

Thus we can perform the induction step as follows.

$$\begin{aligned} \Big|size[K+1]^{-1}(n)\Big| &= n \cdot \Big|size[K]^{-1}(n)\Big| + \Big|size[K]^{-1}(n-1)\Big| \\ &\overset{\text{(IH)}}{=} n \cdot \left\{ {K \atop n} \right\} + \left\{ {K \atop n-1} \right\} \overset{(1.4)}{=} \left\{ {K+1 \atop n} \right\}. \end{aligned}$$

❑

Lists / sequences / tuples give rise to set partitions via an elementary construction that we call *match*. It turns positions with the same 'matching' elements (in the list) into blocks, see Aldous (1985); Pitman (2006); Jacobs and Stein (2023a).

Definition 1.5.8 Fix an arbitrary set X and a number $K \geq 1$. There is a *match* function $mat[K]\colon X^K \to SP(K)$ defined as:

$$mat[K](x_1, \ldots, x_K) := \Big\{ \{i \mid x_i = x_1\}, \ \ldots, \ \{i \mid x_i = x_K\} \Big\}. \tag{1.23}$$

The parameter K in $mat[K]$ is dropped when it is clear from the context.

We illustrate how blocks are formed from positions with equal elements:

$$mat(b, a, c, b, b, a) = \Big\{ \{1, 4, 5\}, \ \{2, 6\}, \ \{3\} \Big\}.$$

This match function will show up in various places. At this stage we observe the following two basic facts.

Lemma 1.5.9 *Let X be a set and let $K \in \mathbb{N}_{>0}$.*

1 The following rectangle commutes.

$$\begin{array}{ccc} X^K & \xrightarrow{\ mat\ } & SP(K) \\ {\scriptstyle supp}\big\downarrow & & \big\downarrow{\scriptstyle size} \\ \mathcal{P}_{\leq}[K](X) & \xrightarrow{\ size\ } & (K] = \{1, \ldots, K\} \end{array} \tag{1.24}$$

2 Now assume that the set X is finite with $n = |X|$ elements. For a set partition $P \in SP(K)$ with $|P| \leq n$, the number of lists in X^K that match to P is a falling factorial:

$$\Big| mat^{-1}(P) \Big| = (n)_{|P|} = \frac{n!}{(n - |P|)!}. \tag{1.25}$$

Proof. 1 The number of blocks of $mat(\vec{x})$ equals the number of different elements in $\vec{x} \in X^K$. This is what the rectangle expresses. It follows immediately from the formulation of match in (1.23).

2 By an elementary counting argument: given a set partition $P \in SP(K)$ we need to count all lists $\vec{x} \in X^K$ with $size(P) = |P|$ many different elements; the partition prescribes at which positions in the list these elements must be equal. This leaves $n(n-1)(n-2)\cdots(n-|P|+1) = (n)_{|P|}$ many choices. ❑

Set partitions form combinatorially rich structures. We already saw how Stirling numbers of the second kind and falling factorials show up. Also rising factorials and Stirling numbers of first kind appear, in the result below and in Exercise 1.5.8.

Proposition 1.5.10 *For a non-empty finite set X and a real number $r \in \mathbb{R}$,*

$$\sum_{P \in SP(X)} r^{|P|} \cdot \prod_{B \in P} (|B| - 1)! = (r)^{|X|} = \prod_{0 \leq i < |X|} r + i. \tag{1.26}$$

In particular, for $r = 1$,

$$\sum_{P \in SP(X)} \prod_{B \in P} (|B| - 1)! = |X|!.$$

The sum over partitions with a particular size n can be described via Stirling numbers of the first kind, see Exercise 1.5.8.

Proof. We use induction on the size $|X| \geq 1$ of the non-empty set X. When $|X| = 1$, say with $X = \{x\}$, the only set-partition is $\{\{x\}\}$. The left-hand-side of (1.26) is then equal to r. Clearly, also the right-hand-side equals r.

Next, let $X = Y \cup \{x\}$, where $x \notin Y$. We assume that (1.26) holds for Y.

$$\begin{aligned}
&\sum_{P \in SP(X)} r^{|P|} \cdot \prod_{B \in P} (|B|-1)! \\
&= \sum_{Q \in SP(Y)} r^{|Q \cup \{\{x\}\}|} \cdot \prod_{B \in Q \cup \{\{x\}\}} (|B|-1)! \\
&\qquad + \sum_{Q \in SP(Y)} \sum_{C \in Q} r^{|(Q \setminus \{C\}) \cup \{C \cup \{x\}\}|} \cdot \prod_{B \in (Q \setminus \{C\}) \cup \{C \cup \{x\}\}} (|B|-1)! \\
&= \sum_{Q \in SP(Y)} r^{|Q|+1} \cdot \prod_{B \in Q} (|B|-1)! \; + \sum_{Q \in SP(Y)} \sum_{C \in Q} r^{|Q|} \cdot |C| \cdot \prod_{B \in Q} (|B|-1)! \\
&\overset{\text{(IH)}}{=} r \cdot \left(\prod_{0 \leq i < |Y|} r + i \right) + |Y| \cdot \left(\prod_{0 \leq i < |Y|} r + i \right) \qquad \text{since } \sum_{C \in Q} |C| = |Y| \\
&= \big(r + |Y|\big) \cdot \left(\prod_{0 \leq i < |Y|} r + i \right) = \prod_{0 \leq i < |X|} r + i.
\end{aligned}$$

❑

Remark 1.5.11 1 As a consequence of Proposition 1.5.7 we can determine the number of set partitions of K as:

$$\Big| SP(K) \Big| = \sum_{1 \leq n \leq K} \Big| size[K]^{-1}(n) \Big| = \left\{ {K \atop 1} \right\} + \cdots + \left\{ {K \atop K} \right\}.$$

The latter expression forms the K-th Bell number. The sequence of these Bell numbers, starting at 1, is: $1, 2, 5, 15, 52, 203, 877, 4140, \ldots$. These are thus the sizes of $SP(1)$, $SP(2)$, $SP(3)$, $\ldots$, see Conway and Guy (1996).

2 Eearlier we skipped the proof of Lemma 1.4.9 about the number of lists with a given support set. It can now be obtained from Proposition 1.5.7. Let $K \geq 1$ be given together with a set U of size $|U| = n \leq K$. The question is: how many lists $\ell \in U^K$ exist with $supp(\ell) = U$. Well, if ℓ contains n different elements, then the associated set partition $mat(\ell) \in SP(K)$ will have n blocks. Proposition 1.5.7 tells that there are $\left\{ {K \atop n} \right\}$ such set partitions. There are $n!$ many lists ℓ with elements from U that match to a particular set partition. Hence in total we have $n! \cdot \left\{ {K \atop n} \right\} = |U|! \cdot \left\{ {K \atop |U|} \right\}$ many lists, as claimed in Lemma 1.4.9.

Exercises

1.5.1 Continuing Exercise 1.4.1, compute:

- $supp(\ell_1)$
- $supp(\ell_2)$
- $supp(\ell_1 +\!\!+ \ell_2)$
- $supp(\ell_1) \cup supp(\ell_2)$
- $supp(\mathcal{L}(f)(\ell_1))$
- $\mathcal{P}_{fin}(f)(supp(\ell_1))$.

1.5.2 We have used finite unions $(\emptyset, \cup)$ as monoid structure on $\mathcal{P}(X)$ in Lemma 1.5.1 (1). Intersections $(X, \cap)$ give another monoid structure on $\mathcal{P}(X)$.

1 Show that the negation / complement function $\neg\colon \mathcal{P}(X) \to \mathcal{P}(X)$, given by:

$$\neg U = X \setminus U = \{x \in X \mid x \notin U\},$$

is a homomorphism of monoids between $(\mathcal{P}(X), \emptyset, \cup)$ and $(\mathcal{P}, X, \cap)$, in both directions. In fact, it forms an isomorphism of monoids, since $\neg\neg U = U$.

2 Prove that the intersections monoid structure is *not* preserved by maps $\mathcal{P}(f)\colon \mathcal{P}(X) \to \mathcal{P}(Y)$.
Hint: Look at preservation of the unit $X \in \mathcal{P}(X)$.

1.5.3 1 Check that in general,

$$\left|\mathcal{P}(f)(U)\right| \neq \left|U\right|.$$

Conclude that the powerset functor $\mathcal{P}$ does not restrict to a functor $\mathcal{P}[K]$, for $K \in \mathbb{N}$.

2 Show that we do have:

$$\left|\mathcal{P}(f)(U)\right| \leq \left|U\right|.$$

Check that $\mathcal{P}$ does restrict to a functor $\mathcal{P}_{\leq}[K]$.

1.5.4 We consider set partitions on arbitrary sets X, as in Definition 1.5.6 (1).

1 Establish bijective correspondences between:

- set partitions $P \in SP(X)$;
- equivalence relations $E \subseteq X \times X$.
- equivalence classes of surjective functions $X \twoheadrightarrow A$, for some set A where $h\colon X \twoheadrightarrow A$ and $k\colon X \twoheadrightarrow B$ are equivalent if there is an isomorphism $j\colon A \xrightarrow{\cong} B$ with $j \circ h = k$.

2 Let $P, Q \in SP(X)$ be set partitions, with corresponding equivalence relations E_P, E_Q and functions h_P, h_Q, as in the previous item. Prove that that the following points are equivalent:

- $P \sqsubseteq Q$, as defined by: $\forall B \in P.\, \exists C \in Q.\, B \subseteq C$;
- $E_P \subseteq E_Q$;
- $j \circ h_P = h_Q$ for some function j.

3 Check that the order $\sqsubseteq$ on $SP(X)$ defined in the previous item is a partial order. It has a least and greatest element; find out what they are.

4 Let $f\colon X \twoheadrightarrow Y$ be a surjective function. Show that it gives rise to a function:

$$SP(X) \xrightarrow{\mathcal{P}(\mathcal{P}(f))} SP(Y)$$

Can you make this work without the surjectivity requirement? Do you see a functor here? From where to where?

1.5.5 Show that the list-support function can be described as $supp\colon X^K \to \mathcal{P}_{\leq}[K](X)$, for each set X and number $K \geq 1$. Check that these maps are natural in X, in the sense that for each function $f\colon X \to Y$ the following rectangle commutes.

$$\begin{tikzcd} X^K \arrow[r, "supp"] \arrow[d, "f^K"'] & \mathcal{P}_{\leq}[K](X) \arrow[d, "\mathcal{P}(f)"] \\ Y^K \arrow[r, "supp"] & \mathcal{P}_{\leq}[K](X) \end{tikzcd}$$

1.5.6 Let X be a (finite) set with n elements. Check that the binomial coefficient counts the number of subsets of a particular size $K \leq n$, in:

$$\left|\mathcal{P}[K](X)\right| = \binom{n}{K}.$$

1.5.7 Let $1 \leq K \leq N$. Show that:

$$\sum_{P \in SP(K)} \frac{1}{(N - |P|)!} = \frac{N^K}{N!}.$$

1.5.8 (See Jacobs (2022a)) Prove that Stirling numbers of the first kind arise not only via subsets, as in Exercise 1.2.8, but also via set partitions, as in:

$$\left[\begin{matrix} K \\ n \end{matrix}\right] = \sum_{P \in SP(K), |P| = n} \prod_{B \in P} (|B| - 1)!.$$

1.6 Multisets

So far we have discussed two collection data types, namely lists and subsets. In lists, elements occur in a particular order, and may occur multiple times (at different positions). Both properties fail for subsets. In this section we look at multisets, also called bags, which are 'subsets' in which elements may occur multiple times, or alternatively, multisets are 'lists' in which the order of elements does not matter — see Proposition 2.11.3 (1) for a precise formulation of the latter description. Hence multisets are inbetween lists and subsets, since they do involve multiple occurrences, but the order of their elements is irrelevant.

The list and subset examples that we have seen are somewhat remote from probability theory. But multisets are much more directly relevant: they can be used for counting frequencies of elements, as a first step towards determining their probabilities. Indeed, observed data can be organised nicely in terms of multisets. For instance, for topic analysis, a document, as a list of words, is often analysed as a multiset of words, in which one keeps track of the words that occur in the document together with their frequency (multiplicity). For learning the topic of a document, the order of the words is not so relevant, see *e.g.* Blei et al. (2003). Similarly, images may be analysed as multisets of pixels Rubner et al. (2000). Also, tables with observed data can be organised naturally as multisets, see Subsection 1.6.1 below. An elementary form of 'frequentist' learning from such tables will be described in Section 2.1, as a (natural) transformation from multisets to distributions. For more information on multisets, also historical, see Singh et al. (2008).

Despite their importance, multisets do not have a prominent role in computer science, like lists have, or in mathematics. There are obvious applications where multisets are the better data structure, for instance in recording votes in an election. Recording votes as a list, reflecting the order of incoming votes, may leak information about who voted what — and thus compromise confidentiality. Instead, votes are better recorded as multisets, capturing the numbers of votes per candidate (or option). A ballot box is an (imperfect) device for turning a list of votes into a multiset. Later we shall use the name accumulation (*acc*) for the mathematical version of this transformation, from lists to multisets, see Equation (1.38).

Also in mathematics, multisets are not always recognised as such. For instance, eigenvalues

of a matrix form a clear example where the 'multi' aspect is ignored: eigenvalues may occur multiple times, so the proper thing to say is that a matrix has a *multiset* of eigenvalues, instead of a *set*. More generally, solutions of a polynomial form a multiset, since solutions may occur multiple times. For instance, the polynomial $x^3 - 7x^2 + 16x - 12 = (x-2)(x-2)(x-3)$ has a multiset of roots, which we shall write as: $2|2\rangle + 1|3\rangle$.

One reason for the common use of lists (or sequences) instead of multisets may be that there is no established notation for multisets. We shall use a 'ket' notation $|-\rangle$ that is borrowed from quantum theory, but interchangeably also a functional notation[1]. In the next chapter we start using the same notation for discrete probability distributions. Since multisets are not so familiar, we take ample time to introduce the basic definitions and properties, in Sections 1.6 – 1.9.

We start with an introduction about notation, terminology, and conventions for multisets. Consider a set $C = \{R, G, B\}$ for the three colours Red, Green, Blue. An example of a multiset with elements from this set C is:

$$2|R\rangle + 5|G\rangle + 0|B\rangle.$$

In this multiset the element R occurs 2 times, G occurs 5 times, and B occurs 0 times. The latter means that B does not occur, that is, B is not an element of the multiset. From a multiset perspective, we have $2 + 5 + 0 = 7$ elements — and not just 2. A multiset like this may describe an urn containing 2 red balls, 5 green ones, and no blue balls. Such multisets are quite common, in different settings. For instance, the chemical formula $C_2H_4O_2$ for vinegar may be read as a multiset $2|C\rangle + 4|H\rangle + 2|O\rangle$, containing 2 carbon ($C$) atoms, 4 hydrogen ($H$) atoms, and 2 oxygen ($O$) atoms, see also Exercise 1.6.4.

In a situation where we have multiple data items, say arising from successive experiments, a basic question to ask is: does the order of the experiments matter? If so, we need to order the data elements as a list. If the order does not matter we should use a multiset. More concretely, if six successive experiments yield data items d, e, d, f, e, d and their order is relevant we should model the data as the list $[d, e, d, f, e, d]$. When the order is irrelevant, we can capture the data as the multiset $3|d\rangle + 2|e\rangle + 1|f\rangle$.

The special brackets $|-\rangle$ form part of so-called *ket* notation, stemming from quantum theory. These kets are meaningless syntax; they are used to separate the natural numbers, called *multiplicities*, and the elements in the multiset.

We move to a more formal description. Let X be an arbitrary set. In terms of the above ket notation, a (finite) multiset over X is an expression of the form:

$$n_1\big|x_1\big\rangle + \cdots + n_k\big|x_k\big\rangle \qquad \text{where} \qquad n_i \in \mathbb{N} \text{ and } x_i \in X.$$

This expression is a formal sum, not an actual sum (for instance in $\mathbb{R}$). We may write it as $\sum_i n_i|x_i\rangle$. We use the convention:

- $0|x\rangle$ may be omitted; but it may also be written explicitly in order to emphasise that the element x does not occur in a multiset;
- a sum $n|x\rangle + m|x\rangle$ is the same as $(n+m)|x\rangle$;
- the order and any round brackets in a sum do not matter.

[1] A notation for multisets that has become somewhat common uses *monomials*, so that an urn with 2 red and 5 green balls is written as R^2G^5. An advantage of the monomial notation is that identifications like $R^4R^2G^5 = R^6G^5$ are natural. A disadvantge is the clash of notation with powers. Others, for instance Graham et al. (1994), simply repeat elements in subset notation, as in $\{R, R, G, G, G, G, G\}$. This does not really scale.

Thus, for instance, there is an equality of multisets:

$$2|a\rangle + (5|b\rangle + 0|c\rangle) + 4|b\rangle \;=\; 9|b\rangle + 2|a\rangle.$$

There is an alternative, functional description of multisets. A multiset can be defined as a function $\varphi\colon X \to \mathbb{N}$ that has finite support, where the support $supp(\varphi) \subseteq X$ is the subset $supp(\varphi) = \{x \in X \mid \varphi(x) \neq 0\}$ where φ is non-zero. Notice that the set X may be infinite, but each multiset φ over X is required to have finite support $supp(\varphi) \subseteq X$.

For each element $x \in X$ the number $\varphi(x) \in \mathbb{N}$ indicates how often x occurs in the multiset φ. Such a function φ can also be written as a formal sum $\sum_x \varphi(x)|x\rangle$, where x ranges over $supp(\varphi)$.

For instance, the multiset $9|b\rangle + 2|a\rangle$ over $A = \{a, b, c\}$ corresponds to the function $\varphi\colon A \to \mathbb{N}$ given by $\varphi(a) = 2$, $\varphi(b) = 9$, $\varphi(c) = 0$. Its support is thus $\{a, b\} \subseteq A$, with two elements. The number $\|\varphi\|$ of elements in φ is 11.

We shall freely switch back and forth between the ket-description and the function-description of multisets, and use whichever form is most convenient for the goal at hand.

Having said this, we stretch the idea of a multiset and do not only allow natural numbers $n \in \mathbb{N}$ as multiplicities (occurrence frequencies), but also allow non-negative numbers $r \in \mathbb{R}_{\geq 0}$. Thus we can have a multiset of the form $\frac{3}{2}|a\rangle + \pi|b\rangle$ where $\pi \in \mathbb{R}_{\geq 0}$ is the famous constant of Archimedes: the ratio of a circle's circumference to its diameter. This added generality will be useful at times, although many examples of multisets will simply have natural numbers as multiplicities. We call such multisets *natural*.

Definition 1.6.1 Let X be an arbitrary set.

1 We write $\mathcal{M}(X)$ for the set of all *multisets* over X. Thus, using the function approach:

$$\mathcal{M}(X) \;:=\; \{\varphi\colon X \to \mathbb{R}_{\geq 0} \;\big|\; supp(\varphi) \text{ is finite}\}.$$

The functions in $\mathcal{M}(X)$ may be called *mass functions*, as in Ścibior et al. (2018). We shall often write such a function $\varphi \in \mathcal{M}(X)$ as formal sum $\sum_x \varphi(x)|x\rangle$.

2 We write $\mathcal{N}(X) \subseteq \mathcal{M}(X)$ for the subset of *natural* multisets, with natural numbers as multiplicities — also called *bags* or *urns*. Thus, $\mathcal{N}(X)$ contains functions $\varphi \in \mathcal{M}(X)$ with $\varphi(x) \in \mathbb{N}$, for all $x \in X$.

3 We shall write $\mathcal{M}_*(X)$ for the set of *non-empty* multisets. Thus:

$$\begin{aligned}\mathcal{M}_*(X) \;&:=\; \{\varphi \in \mathcal{M}(X) \;\big|\; supp(\varphi) \text{ is non-empty}\}\\ &=\; \{\varphi\colon X \to \mathbb{R}_{\geq 0} \;\big|\; supp(\varphi) \text{ is finite and non-empty}\}.\end{aligned}$$

Similarly, $\mathcal{N}_*(X) \subseteq \mathcal{M}_*(X)$ contains the non-empty natural multisets.

4 For a number K we shall write $\mathcal{M}[K](X) \subseteq \mathcal{M}(X)$ and $\mathcal{N}[K](X) \subseteq \mathcal{N}(X)$ for the subsets of multisets of size K, that is, with K elements. Thus:

$$\mathcal{M}[K](X) \;:=\; \{\varphi \in \mathcal{M}(X) \;\big|\; \|\varphi\| = K\} \qquad \text{where} \qquad \|\varphi\| \;:=\; \textstyle\sum_x \varphi(x). \tag{1.27}$$

This size $\|\varphi\| \in \mathbb{R}_{\geq 0}$ gives the total 'number' of elements in the multiset φ. Sometimes we write $size(\varphi) := \|\varphi\|$, via an explicit function $size\colon \mathcal{M}(X) \to \mathbb{R}_{\geq 0}$; it restricts to $size\colon \mathcal{N}(X) \to \mathbb{N}$.

5 Finally, when the set X is finite, we use special notation for the subsets of (natural) multisets φ with full support, that is with $supp(\varphi) = X$.

$$\begin{aligned}\mathcal{M}_{fs}(X) \;&:=\; \{\varphi \in \mathcal{M}(X) \;\big|\; supp(\varphi) = X\}\\ \mathcal{N}_{fs}(X) \;&:=\; \{\varphi \in \mathcal{N}(X) \;\big|\; supp(\varphi) = X\}.\end{aligned}$$

We have seen in (1.8) that the set $\mathcal{L}(X)$ of lists over a set X can be written as disjoint union $\coprod_n$ of sets X^n of lists of length n. Similarly we can write the set of natural multisets $\mathcal{N}(X)$ as disjoint union of multisets of a fixed size:

$$\mathcal{N}(X) = \coprod_{K \in \mathbb{N}} \mathcal{N}[K](X). \tag{1.28}$$

All of $\mathcal{M}, \mathcal{N}, \mathcal{M}_*, \mathcal{N}_*, \mathcal{M}[K], \mathcal{N}[K]$ are functorial, in the same way. Hence we concentrate on $\mathcal{M}$. For a function $f\colon X \to Y$ we define $\mathcal{M}(f)\colon \mathcal{M}(X) \to \mathcal{M}(Y)$ below, in two equivalent ways. Intuitively, when we see a multiset $\varphi \in \mathcal{M}(X)$ as an urn containing coloured balls, with colours from X, then $\mathcal{M}(f)(\varphi) \in \mathcal{M}(Y)$ is the urn with 'repainted' balls, where the new colours are taken from the set Y. The function $f\colon X \to Y$ defines the transformation of colours. It tells that a ball of colour $x \in X$ in φ will be repainted with colour $f(x) \in Y$. Notice that two different colours $x, x' \in X$ may have the same colour, after repainting, namely when $f(x) = f(x')$.

Definition 1.6.2 For a function $f\colon X \to Y$ and a multiset $\varphi \in \mathcal{M}(X)$ over X there is a multiset $\mathcal{M}(f)(\varphi) \in \mathcal{M}(Y)$ that can be defined in two equivalent ways. First, using ket notation as:

$$\mathcal{M}(f)\Big(\textstyle\sum_i r_i|x_i\rangle\Big) := \textstyle\sum_i r_i|f(x_i)\rangle.$$

Equivalently, using function notation one defines $\mathcal{M}(f)(\varphi)\colon Y \to \mathbb{R}_{\geq 0}$ as:

$$\mathcal{M}(f)(\varphi)(y) := \sum_{x \in f^{-1}(y)} \varphi(x).$$

It may take a bit of effort to see that these two descriptions are the same, see Exercise 1.6.1 below. Notice that in the sum $\sum_i r_i|f(x_i)\rangle$ it may happen that $f(x_{i_1}) = f(x_{i_2})$ for $x_{i_1} \neq x_{i_2}$, so that r_{i_1} and r_{i_2} are added together. Thus, the support of $\mathcal{M}(f)(\sum_i r_i|x_i\rangle)$ may have fewer elements than the support of $\sum_i r_i|x_i\rangle$, but the sum of all multiplicities is the same in $\mathcal{M}(f)(\sum_i r_i|x_i\rangle)$ and $\sum_i r_i|x_i\rangle$, see Exercise 1.7.2 below.

Applying $\mathcal{M}$ to a projection function $\pi_i\colon X_1 \times \cdots \times X_n \to X_i$ yields a function $\mathcal{M}(\pi_i)$, from the set $\mathcal{M}(X_1 \times \cdots \times X_n)$ of multisets over a product to the set $\mathcal{M}(X_i)$ of multisets over a component. This $\mathcal{M}(\pi_i)$ is called a *marginalisation function*. It computes what is 'on the side', in the marginal of a table, as will be illustrated in Subsection 1.6.1 below.

Multisets, like lists and subset form a monoid. In terms of urns with coloured balls, taking the sum of two multisets corresponds to pouring the balls from two urns into a new, single urn.

Lemma 1.6.3 *Let X be an arbitrary set.*

1. *The set $\mathcal{M}(X)$ of multisets over X is a commutative monoid. In functional form, addition and zero (identity) element $\mathbf{0} \in \mathcal{M}(X)$ are defined as:*

$$(\varphi + \psi)(x) := \varphi(x) + \psi(x) \qquad\qquad \mathbf{0}(x) := 0.$$

   These sums restrict to $\mathcal{N}(X)$. In the same way one can also define pointwise products $\varphi \cdot \psi$. They will be used later, but for the time being the emphasis lies on sums of multisets.
2. *The set $\mathcal{M}(X)$ is also a* cone*: it is closed under 'scalar' multiplication with non-negative numbers $r \in \mathbb{R}_{\geq 0}$, via:*

$$(r \cdot \varphi)(x) := r \cdot \varphi(x).$$

   This scalar multiplication $r \cdot (-)\colon \mathcal{M}(X) \to \mathcal{M}(X)$ preserves the sums $(\mathbf{0}, +)$ from the previous item, and is thus a map of monoids.

3 *For each $f\colon X \to Y$, the function $\mathcal{M}(f)\colon \mathcal{M}(X) \to \mathcal{M}(Y)$ is a map of monoids and also of cones. The latter means: $\mathcal{M}(f)(r \cdot \varphi) = r \cdot \mathcal{M}(f)(\varphi)$.*

4 *The support map $\mathrm{supp}\colon \mathcal{M}(X) \to \mathcal{P}_{\mathrm{fin}}(X)$ is a homomorphism of monoids and is natural in X. The latter means that for each function $f\colon X \to Y$ the following rectangle commutes.*

$$\begin{tikzcd} \mathcal{M}(X) \arrow[r, "supp"] \arrow[d, "\mathcal{M}(f)"'] & \mathcal{P}_{\mathrm{fin}}(X) \arrow[d, "\mathcal{P}(f)"] \\ \mathcal{M}(Y) \arrow[r, "supp"] & \mathcal{P}_{\mathrm{fin}}(Y) \end{tikzcd}$$

❑

The fact that $\mathcal{M}(f)$ preserves sums can be understood informally as follows. If we have two urns, we can first combine their contents and then repaint everything. Alternatively, we can first repaint the balls in the two urns separately, and then throw them together. The result is the same, in both cases.

The element $\mathbf{0} \in \mathcal{M}(X)$ used in item (1) is the empty multiset, that is, the urn containing no balls. Similarly, the sum of multisets $+$ is implicit in the ket-notation. The set $\mathcal{N}(X)$ of *natural* multisets is not closed in general under scalar multiplication with $r \in \mathbb{R}_{\geq 0}$. It is closed under scalar multiplication with $n \in \mathbb{N}$, but such multiplications add nothing new since they can also be described via repeated addition.

1.6.1 Tables of data as multisets

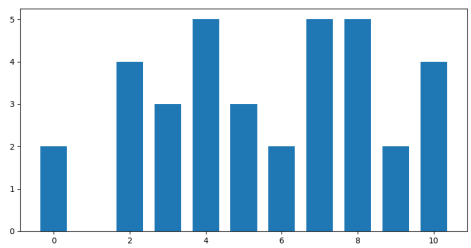
Let's assume that a group of 36 children in the age range $0 - 10$ is participating in some study, where the number of children of each age is given by the following table.

0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	0	4	3	5	3	2	5	5	2	4

We can represent this table as a natural multiset over the set of ages $\{0, 1, \ldots, 10\}$.

$$2|0\rangle + 4|2\rangle + 3|3\rangle + 5|4\rangle + 3|5\rangle + 2|6\rangle + 5|7\rangle + 5|8\rangle + 2|9\rangle + 4|10\rangle.$$

Notice that there is no summand for age 1 because of our convention to ommit expressions like $0|1\rangle$ with multiplicity 0. We can visually represent the above age data / multiset in the form of a histogram:



(1.29)

When such a histogram is given, it is generally easy to see what the underlying multiset is.

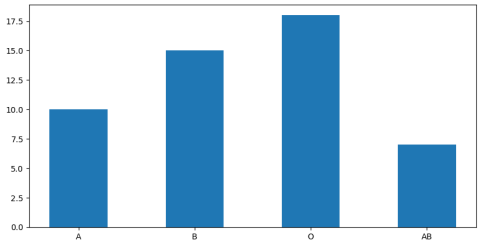
Here is another example, not with numerical data, in the form of ages, but with nominal data, in the form of blood types. Testing the blood type of 50 individuals produced the following table.

A	**B**	**O**	**AB**
10	15	18	7

This corresponds to a (natural) multiset over the set $\{A, B, O, AB\}$ of blood types, namely to:

$$10\big|A\rangle + 15\big|B\rangle + 18\big|O\rangle + 7\big|AB\rangle.$$

It gives rise to the following bar graph, in which there is no particular ordering of elements. For convenience, we follow the order of the above table.



(1.30)

Next, consider the two-dimensional table (1.31) below where we have combined numeric information about blood pressure (either high H, or low L) and certain medicines (either type 1, type 2, or no medicine, indicated as 0). There is data about 100 study participants:

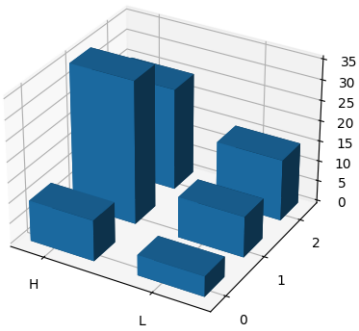
	no medicine	**medicine 1**	**medicine 2**	**totals**
high	10	35	25	70
low	5	10	15	30
totals	15	45	40	100

(1.31)

We claim that we can capture this table as a (natural) multiset. To do so, we first form sets $B = \{H, T\}$ for blood pressure values, and $M = \{0, 1, 2\}$ for types of medicine. The above table can then be described as a natural multiset τ over the product set / space $B \times M$, that is, as an element $\tau \in \mathcal{N}(B \times M)$, namely:

$$\tau = 10\big|H, 0\rangle + 35\big|H, 1\rangle + 25\big|H, 2\rangle + 5\big|L, 0\rangle + 10\big|L, 1\rangle + 15\big|L, 2\rangle.$$

Such a multiset can be plotted in three dimensions as:



We see that Table (1.31) contains 'totals' in its vertical and horizontal margins. They can be obtained from the multiset τ as marginals, using the functoriality of $\mathcal{N}$. This works as follows. Applying the natural multiset functor $\mathcal{N}$ to the two projections $\pi_1 \colon B \times M \to B$ and $\pi_2 \colon B \times M \to M$ yields marginal distributions on B and M, namely:

$$\begin{aligned}
\mathcal{N}(\pi_1)(\tau) &= 10\big|\pi_1(H,0)\big\rangle + 35\big|\pi_1(H,1)\big\rangle + 25\big|\pi_1(H,2)\big\rangle \\
&\qquad + 5\big|\pi_1(L,0)\big\rangle + 10\big|\pi_1(L,1)\big\rangle + 15\big|\pi_1(L,2)\big\rangle \\
&= 10\big|H\big\rangle + 35\big|H\big\rangle + 25\big|H\big\rangle + 5\big|L\big\rangle + 10\big|L\big\rangle + 15\big|L\big\rangle \\
&= 70\big|H\big\rangle + 30\big|L\big\rangle. \\
\mathcal{N}(\pi_2)(\tau) &= (10+5)\big|0\big\rangle + (35+10)\big|1\big\rangle + (25+15)\big|2\big\rangle \\
&= 15\big|0\big\rangle + 45\big|1\big\rangle + 40\big|2\big\rangle.
\end{aligned}$$

The expression 'marginal' is used to describe such totals in the margin of a multidimensional table. In Section 2.1 we describe how to obtain probabilities from tables in a systematic manner.

In general, 'data' may be described as a multiset in a set of the form $\mathcal{N}[K](X_1 \times \cdots \times X_n)$. Such a multiset has K-many n-ary tuples in its support. When one speaks of 'big data', this may mean two things:

- a high size K, so that the data / multisets contains many tuples, as data items;
- a high size n of each tuple, so that there are many components (features) in X_i in each data item in the data / multiset.

Of course, both K and n may be large numbers.

1.6.2 Unit and flatten for multisets

There are unit and flatten maps for multisets too — like for lists and subsets. The function $\mathit{unit} \colon X \to \mathcal{M}(X)$ is simply given by the singleton multiset: $\mathit{unit}(x) := 1|x\rangle$. Flattening involves turning a multiset of multisets into a multiset. Concretely, this is done as:

$$\mathit{flat}\Big(\tfrac{1}{3}\big|\,2|a\rangle + 2|c\rangle\,\big\rangle + 5\big|\,1|b\rangle + \tfrac{1}{6}|c\rangle\,\big\rangle\Big) = \tfrac{2}{3}|a\rangle + 5|b\rangle + \tfrac{3}{2}|c\rangle.$$

More generally, flattening is the function $\mathit{flat} \colon \mathcal{M}(\mathcal{M}(X)) \to \mathcal{M}(X)$ with:

$$\mathit{flat}\Big(\textstyle\sum_i r_i|\varphi_i\rangle\Big) := \displaystyle\sum_{x \in X} \big(\textstyle\sum_i r_i \cdot \varphi_i(x)\big)\big|x\rangle.$$

Notice that the big outer sum $\sum_x$ is a formal one, whereas the inner sum $\sum_i$ is an actual one, in $\mathbb{R}_{\geq 0}$, see the earlier example.

The following result, about unit and flatten for multisets, is formally similar to such results about unit and flatten for lists and subsets, in Lemma 1.4.5 and 1.5.2. We formulate it for general multisets $\mathcal{M}$, but it restricts to natural multisets $\mathcal{N}$.

Lemma 1.6.4 *1 For each function $f \colon X \to Y$ the two rectangles*

$$\begin{tikzcd}
X \arrow[r, "\mathit{unit}"] \arrow[d, "f"'] & \mathcal{M}(X) \arrow[d, "\mathcal{M}(f)"] \\
Y \arrow[r, "\mathit{unit}"'] & \mathcal{M}(Y)
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
\mathcal{M}(\mathcal{M}(X)) \arrow[r, "\mathit{flat}"] \arrow[d, "\mathcal{M}(\mathcal{M}(f))"'] & \mathcal{M}(X) \arrow[d, "\mathcal{M}(f)"] \\
\mathcal{M}(\mathcal{M}(Y)) \arrow[r, "\mathit{flat}"'] & \mathcal{M}(Y)
\end{tikzcd}$$

commute.

2 *The next two diagrams also commute.*

$$\begin{tikzcd} \mathcal{M}(X) \arrow[r, "unit"] \arrow[dr, equal] & \mathcal{M}(\mathcal{M}(X)) \arrow[d, "flat"] & \mathcal{M}(X) \arrow[l, "\mathcal{M}(unit)"'] \arrow[dl, equal] \\ & \mathcal{M}(X) & \end{tikzcd} \qquad \begin{tikzcd} \mathcal{M}(\mathcal{M}(\mathcal{M}(X))) \arrow[r, "flat"] \arrow[d, "\mathcal{M}(flat)"'] & \mathcal{M}(\mathcal{M}(X)) \arrow[d, "flat"] \\ \mathcal{M}(\mathcal{M}(X)) \arrow[r, "flat"'] & \mathcal{M}(X) \end{tikzcd}$$

❑

The next result shows that *natural* multisets are free commutative monoids. Arbitrary multisets are also free, but for other algebraic structures, see Exercise 1.6.13.

Proposition 1.6.5 *Let X be a set and $(M, 0, +)$ a commutative monoid. Each function $f\colon X \to M$ has a unique extension to a homomorphism of monoids $\overline{f}\colon (\mathcal{N}(X), \mathbf{0}, +) \to (M, 0, +)$ with $\overline{f} \circ unit = f$. The diagram below captures the situation, where the dashed arrow is used for uniqueness.*

$$\begin{tikzcd} X \arrow[r, "unit"] \arrow[dr, bend right, "f"'] & \mathcal{N}(X) \arrow[d, dashed, "\overline{f},\ homomorphism"] \\ & M \end{tikzcd} \tag{1.32}$$

Proof. One defines:

$$\overline{f}\big(n_1|x_1\rangle + \cdots + n_k|x_k\rangle\big) := n_1 \cdot f(x_1) + \cdots + n_k \cdot f(x_k).$$

where we write $n \cdot a$ for the n-fold sum $a + \cdots + a$ in a monoid. ❑

The unit and flatten operations for (natural) multisets can be used to capture commutative monoids more precisely, in analogy with Propositions 1.4.6 and 1.5.5

Proposition 1.6.6 *Let X be an arbitrary set.*

1 *A commutative monoid structure $(u, +)$ on X corresponds to an $\mathcal{N}$-algebra $\alpha\colon \mathcal{N}(X) \to X$ making the two diagrams below commute.*

$$\begin{tikzcd} X \arrow[r, "unit"] \arrow[dr, "id"'] & \mathcal{N}(X) \arrow[d, "\alpha"] \\ & X \end{tikzcd} \qquad \begin{tikzcd} \mathcal{N}(\mathcal{N}(X)) \arrow[r, "\mathcal{N}(\alpha)"] \arrow[d, "flat"'] & \mathcal{N}(X) \arrow[d, "\alpha"] \\ \mathcal{N}(X) \arrow[r, "\alpha"] & X \end{tikzcd} \tag{1.33}$$

2 *Let $(M_1, u_1, +_1)$ and $(M_2, u_2, +_2)$ be two commutative monoids, with corresponding $\mathcal{N}$-algebras $\alpha_1\colon \mathcal{N}(M_1) \to M_1$ and $\alpha_2\colon \mathcal{N}(M_2) \to M_2$. A function $f\colon M_1 \to M_2$ is a map of monoids if and only if the rectangle*

$$\begin{tikzcd} \mathcal{N}(M_1) \arrow[r, "\mathcal{N}(f)"] \arrow[d, "\alpha_1"'] & \mathcal{N}(M_2) \arrow[d, "\alpha_2"] \\ M_1 \arrow[r, "f"] & M_2. \end{tikzcd} \tag{1.34}$$

commutes.

Proof. Analogously to the proof Proposition 1.4.6: if $(X, u, +)$ is a commutative monoid, we define $\alpha\colon \mathcal{N}(X) \to X$ by turning formal sums into actual sums: $\alpha(\sum_i n_i|x_i\rangle) := \sum_i n_i \cdot x_i$, see (the proof of) Proposition 1.6.5. In the other direction, given $\alpha\colon \mathcal{N}(X) \to X$ we define a sum as $x + y := \alpha(1|x\rangle + 1|y\rangle)$ with unit $u := \alpha(\mathbf{0})$. Obviously, $+$ is commutative. ❑

1.6.3 Extraction

At the end of the previous section we have seen how to extract a function (channel) from a binary subset, that is, from a relation. It turns out that one can do the same for a binary multiset, that is, for a table. More specifically, in terms of exponents, there are isomorphisms:

$$\mathcal{M}(Y)^X \cong \mathcal{M}(X \times Y) \cong \mathcal{M}(X)^Y. \tag{1.35}$$

This is analogous to (1.21) for powerset.

How does this work in detail? Suppose we have an arbitary multiset / table $\sigma \in \mathcal{M}(X \times Y)$. From σ one can extract a function $extr_1(\sigma)\colon X \to \mathcal{M}(Y)$, and also $extr_2(\sigma)\colon Y \to \mathcal{M}(X)$, via:

$$extr_1(\sigma)(x) = \sum_{y \in Y} \sigma(x,y) \big| y \rangle \qquad\qquad extr_2(\sigma)(y) = \sum_{x \in X} \sigma(x,y) \big| x \rangle.$$

Notice that we are — conveniently — mixing ket and function notation for multisets. Conversely, σ can be reconstructed from $extr_1(\sigma)$, and also from $extr_2(\sigma)$, via $\sigma(x,y) = extr_1(\sigma)(x)(y) = extr_2(\sigma)(y)(x)$.

Functions of the form $X \to \mathcal{M}(Y)$ will also be used as channels from X to Y, see Section 1.11. That's why we often speak about 'channel extraction'.

As illustration, we apply extraction to the medicine - blood pressure Table 1.31, described as the multiset $\tau \in \mathcal{M}(B \times M)$. It gives rise to two channels $extr_1(\tau)\colon B \to \mathcal{M}(M)$ and $extr_2(\tau)\colon M \to \mathcal{M}(B)$. Explicitly:

$$\begin{aligned} extr_1(\tau)(H) &= \sum_{x \in M} \tau(H,x) | x \rangle = 10|0\rangle + 35|1\rangle + 25|2\rangle \\ extr_1(\tau)(L) &= \sum_{x \in M} \tau(L,x) | x \rangle = 5|0\rangle + 10|1\rangle + 15|2\rangle. \end{aligned}$$

We see that this extracted function captures the two rows of Table 1.31. Similarly we get the columns via the second extracted function:

$$extr_2(\tau)(0) = 10|L\rangle + 5|H\rangle \qquad extr_2(\tau)(1) = 35|L\rangle + 10|H\rangle \qquad extr_2(\tau)(2) = 25|L\rangle + 15|H\rangle.$$

Exercises

1.6.1 In the setting of Exercise 1.4.1, consider the multisets $\varphi = 3|a\rangle + 2|b\rangle + 8|c\rangle$ and $\psi = 3|b\rangle + 1|c\rangle$. Compute:

- $\varphi + \psi$
- $\psi + \varphi$
- $\mathcal{M}(f)(\varphi)$, both in ket-formulation and in function-formulation
- idem for $\mathcal{M}(f)(\psi)$
- $\mathcal{M}(f)(\varphi + \psi)$
- $\mathcal{M}(f)(\varphi) + \mathcal{M}(f)(\psi)$.

1.6.2 Consider, still in the context of Exercise 1.4.1, the 'joint' multiset $\varphi \in \mathcal{M}(X \times Y)$ given by $\varphi = 2|a,u\rangle + 3|a,v\rangle + 5|c,v\rangle$. Determine the marginals $\mathcal{M}(\pi_1)(\varphi) \in \mathcal{M}(X)$ and $\mathcal{M}(\pi_2)(\varphi) \in \mathcal{M}(Y)$.

1.6.3 Let $P = \{2, 3, 5, 7, 11, \ldots\} \subseteq \mathbb{N}_{>0}$ be the set of prime numbers. Define prime number factorisation *pnf* as a function:

$$\mathbb{N}_{>0} \xrightarrow{pnf} \mathcal{N}(P)$$

1 Describe concretely *pnf* (2023).
2 Show that *pnf* is a homomorphism of (commutative) monoids.
3 Check that *pnf* is in fact an isomorphism of monoids.

Thus, positive natural numbers can be identified with natural multisets of prime numbers.

1.6.4 Consider the chemical equation for burning methane:

$$CH_4 \;+\; 2\,O_2 \;\longrightarrow\; CO_2 \;+\; 2\,H_2O.$$

Check that there is an underlying equation of multisets:

$$flat\Big(1\Big|\,1|C\rangle + 4|H\rangle\rangle + 2\Big|2|O\rangle\rangle\Big) \;=\; flat\Big(1\Big|\,1|C\rangle + 2|O\rangle\rangle + 2\Big|2|H\rangle + 1|O\rangle\rangle\Big).$$

It expresses the law of conservation of mass.

1.6.5 Recall that we write $\boldsymbol{n} := \{0, \ldots, n-1\}$ so that $\mathbf{0} = \emptyset$, $\mathbf{1} = \{0\}$ and $\mathbf{2} = \{0, 1\}$. Verify that for the list, powerset, natural multiset, multiset and distribution functors satisfy:

$$\begin{array}{lllll}
\mathcal{L}(\mathbf{0}) \cong \mathbf{1} & \mathcal{P}(\mathbf{0}) \cong \mathbf{1} & \mathcal{N}(\mathbf{0}) \cong \mathbf{1} & \mathcal{M}(\mathbf{0}) \cong \mathbf{1} & \mathcal{N}[K](\mathbf{0}) \cong \mathbf{0} \\
\mathcal{L}(\mathbf{1}) \cong \mathbb{N} & \mathcal{P}(\mathbf{1}) \cong \mathbf{2} & \mathcal{N}(\mathbf{1}) \cong \mathbb{N} & \mathcal{M}(\mathbf{1}) \cong \mathbb{R}_{\geq 0} & \mathcal{N}[K](\mathbf{1}) \cong \mathbf{1} \\
 & \mathcal{P}(\boldsymbol{n}) \cong \mathbf{2}^n & \mathcal{N}(\boldsymbol{n}) \cong \mathbb{N}^n & \mathcal{M}(\boldsymbol{n}) \cong \mathbb{R}^n_{\geq 0} & \mathcal{N}[K](\mathbf{2}) \cong \boldsymbol{K}+\mathbf{1}
\end{array}$$

1.6.6 Consider the set $\mathcal{N}(\mathbb{N})$ of multisets over the natural numbers.

1 Identify a multiset $\sum_i n_i|i\rangle$ in this set with a polynomial $\sum_i n_i x^i$ having natural numbers as coefficients — where x is some placeholder variable.
2 Check that addition of multisets corresponds to addition of poynomials.

1.6.7 Check that $\mathcal{M}(id) = id$ and $\mathcal{M}(g \circ f) = \mathcal{M}(g) \circ \mathcal{M}(f)$.

1.6.8 Show that for each natural number K the mappings $X \mapsto \mathcal{M}[K](X)$ and $X \mapsto \mathcal{N}[K](X)$ are functorial. Notice the difference with $\mathcal{P}[K]$, see Exercise 1.5.3.

1.6.9 Consider for a set X the mapping $units\colon \mathcal{P}_{fin}(X) \to \mathcal{N}(X)$ given by $units(\{x_1, \ldots, x_n\}) := \sum_i 1|x_i\rangle$. Check that this mapping *units* is *not* natural in X.

1.6.10 Prove Lemma 1.6.4.

1.6.11 Consider both Lemma 1.6.4 and Proposition 1.6.6.

1 Notice that an abstract way of seeing that $\mathcal{N}(X)$ is a commutative monoid is via the properties of the flatten map $flat\colon \mathcal{N}(\mathcal{N}(X)) \to \mathcal{N}(X)$.
2 Notice also that this flatten map is a homomorphism of monoids.

1.6.12 Verify that the support map $supp\colon \mathcal{M}(X) \to \mathcal{P}_{fin}(X)$ commutes with extraction functions, in the sense that the following diagram commutes.

$$\begin{tikzcd}
\mathcal{M}(X \times Y) \arrow[r, "supp"] \arrow[d, "extr_1"'] & \mathcal{P}_{fin}(X \times Y) \arrow[d, "extr_1"] \\
\mathcal{M}(Y)^X \arrow[r, "supp^X"'] & \mathcal{P}_{fin}(Y)^X
\end{tikzcd}$$

Equationally, this amounts to showing that for $\tau \in \mathcal{M}(X \times Y)$ and $x \in X$ one has:

$$extr_1(supp(\tau))(x) \;=\; supp(extr_1(\tau)(x)).$$

Here we use that $supp^X(f) := supp \circ f$, so that $supp^X(f)(x) = supp(f(x))$.

1.6.13 In Proposition 1.6.5 we have seen that natural multisets $\mathcal{N}(X)$ form free commutative monoids. What about general multisets $\mathcal{M}(X)$? They form free cones. Briefly, a cone is a commutative monoid M with scalar multiplication action $\mathbb{R}_{\geq 0} \times M \to M$ forming for each $r \in \mathbb{R}_{\geq 0}$ a homomorphism of monoids $r \cdot (-) \colon M \to M$. It is like a vector space, not over all reals $\mathbb{R}$, but only over the non-negative reals $\mathbb{R}_{\geq 0}$. Homomorphisms of cones are maps of monoids that preserve such scalar multiplications.

Let X be a set and $(M, 0, +)$ a cone, with a function $f \colon X \to M$. Prove that there is a unique homomorphism of cones $\overline{f} \colon \mathcal{M}(X) \to M$ with $\overline{f} \circ unit = f$.

1.7 Multisets in summations

The current and next section will dive deeper into the use of natural multisets in combinatorics, as preparation for later use. This section will focus on the use of multisets in summations, such as the multinomial theorem — which is probably best known in binary form, as the binomial theorem for expanding sums of the form $(a + b)^n$. We shall describe various extensions, both for finite and infinite sums. All material in this section is standard, but its presentation in terms of multisets is not.

This section also introduces the 'accumulation' operation *acc* that sends a list to a multiset by forgetting the order but keeping the multiplicities of elements. This *acc* turns out to be a natural transformation, see the diagram (1.45). We shall encounter the multinomial coefficient of a multiset as the number of lists that accumulate to it.

In this section we focus on counting *with* multisets, in particular in (infinite) sums of powers. The next section focuses on counting multisets themselves, where we ask, for instance, how many multisets of size K are there over a set with n elements?

There are several ways to associate a natural number with a multiset φ. For instance, we can look at the size of its support $|supp(\varphi)| \in \mathbb{N}$, or at its size, as total number of elements $\|\varphi\| = \sum_x \varphi(x) \in \mathbb{R}_{>0}$. This size is a natural number when φ is a natural multiset. Below we will introduce several more such numbers for natural multisets φ, namely $\varphi!_\circ$ and (φ), and later on also a binomial coefficient $\binom{\psi}{\varphi}$.

Definition 1.7.1 Let X be an arbitrary set.

1 For two multisets $\varphi, \psi \in \mathcal{M}(X)$ we write:

$$\varphi \leq \psi \iff \forall x \in X.\, \varphi(x) \leq \psi(x).$$

When $\varphi \leq \psi$ we define subtraction $\psi - \varphi$ of multisets as the obvious multiset, defined pointwise as: $(\psi - \varphi)(x) = \psi(x) - \varphi(x)$.

2 For an arbitrary number $K \in \mathbb{N}$ and natural multisets $\varphi, \psi \in \mathcal{N}(X)$ we use an order $\leq_K$ with K as subscript for:

$$\varphi \leq_K \psi \iff \|\varphi\| = K \text{ and } \varphi \leq \psi.$$

3 For a collection of numbers $r = (r_x)_{x \in X}$ and a natural multiset $\varphi \in \mathcal{N}(X)$ we write:

$$r^{\varphi} \coloneqq \prod_{x \in X} r_x^{\varphi(x)} = \vec{r}^{\,\varphi}.$$

The latter vector notation is appropriate in a situation with a particular order.

4 The factorial $\varphi\!\!\!\!\,!$ of a natural multiset $\varphi \in \mathcal{N}(X)$ is the product of the factorial of its multiplicities:

$$\varphi\!\!\!\!\,! \coloneqq \prod_{x \in supp(\varphi)} \varphi(x)! \tag{1.36}$$

5 The *multiset coefficient* $\binom{\varphi}{}$ is defined as:

$$\binom{\varphi}{} \coloneqq \frac{\|\varphi\|!}{\varphi\!\!\!\!\,!} = \frac{\|\varphi\|!}{\prod_x \varphi(x)!} = \binom{\|\varphi\|}{\varphi(x_1), \ldots, \varphi(x_n)}.$$

The multinomial coefficient from Definition 1.2.1 (3) is used in the latter formulation; it assumes that the support of φ is somehow ordered as a list $[x_1, \ldots, x_n]$.

For instance,

$$(3|R\rangle + 2|B\rangle)\!\!\!\!\,! = 3! \cdot 2! = 12 \qquad \text{and} \qquad \binom{3|R\rangle + 2|B\rangle}{} = \frac{5!}{12} = 10.$$

In Proposition 1.7.2 below we make precise how the multiset coefficient $\binom{\varphi}{}$ in item (5) counts the number of lists that 'accumulate to' the multiset φ.

The traditional notation $\binom{m}{m_1, \ldots, m_k}$ for multinomial coefficients, that we have seen in Definition 1.2.1 (3), is suboptimal for two reasons: first, the number m at the top is superflous, since it is determined by the m_i as $m = \sum_i m_i$; second, the order of the m_i is irrelevant. These disadvantages are resolved by the multiset variant $\binom{\varphi}{}$. It has our preference.

For a multiset φ we have already used the 'support' definition $supp(\varphi) = \{x \mid \varphi(x) \neq 0\}$. This yields a map $supp \colon \mathcal{M}(X) \to \mathcal{P}_{fin}(X)$, which is well-behaved, in the sense that it is natural and preserves the monoid structures on $\mathcal{M}(X)$ and $\mathcal{P}_{fin}(X)$, see Lemma 1.6.3 (4). We have also seen a support map from lists $\mathcal{L}$ to finite powerset $\mathcal{P}_{fin}$. This support map factorises through multisets, as described with a new function *acc* in the following triangle.

$$\begin{tikzcd} \mathcal{L}(X) \arrow[rr, "supp"] \arrow[dr, bend right, "acc"'] & & \mathcal{P}_{fin}(X) \\ & \mathcal{N}(X) \arrow[ur, bend right, "supp"'] & \end{tikzcd} \tag{1.37}$$

The 'accumulator' map $acc \colon \mathcal{L}(X) \to \mathcal{N}(X)$ plays an important role in the rest of this book; it can be traced back to early work in formal lanuage theory, see *e.g.* (Parikh, 1966, Defn. 13). This map *acc* counts (accumulates) how many times an element occurs in a list, while ignoring the order of occurrences. Thus, for a list $\ell \in \mathcal{L}(X)$,

$$acc(x_1, \ldots, x_n) \coloneqq 1|x_1\rangle + \cdots + 1|x_n\rangle. \tag{1.38}$$

Since multiple terms $1|x\rangle$ add up we have $acc(\ell)(x) = n$ if and only if the element x occurs n times in the list ℓ.

In an example:

$$\begin{aligned} acc(a, b, a, b, c, b, b) &= 1|a\rangle + 1|b\rangle + 1|a\rangle + 1|b\rangle + 1|c\rangle + 1|b\rangle + 1|b\rangle \\ &= 2|a\rangle + 4|b\rangle + 1|c\rangle. \end{aligned}$$

The above diagram (1.37) reflects an earlier informal statement from (1.1), namely that multisets are inbetween lists and subsets, see also diagram (1.45) below for a more elaborate description.

We shall use a K-ary version of accumulation, for $K \in \mathbb{N}$, restricted to K-many elements. It then becomes a mapping:

$$X^K \xrightarrow{acc[K]} \mathcal{N}[K](X). \tag{1.39}$$

The parameter K will often be omitted from $acc[K]$ when it is clear from the context.

A basic question is: how many lists of coloured balls give rise to a specific urn with balls? More technically, given a natural multiset φ, how many lists ℓ statisfy $acc(\ell) = \varphi$? In yet another form, what is the size $\left|acc^{-1}(\varphi)\right|$ of the inverse image? This is where the multiset coefficient $\mathbf{(}\varphi\mathbf{)}$ from Definition 1.7.1 (5) comes in.

Proposition 1.7.2 *For $\varphi \in \mathcal{N}[K](X)$ one has:*

$$\mathbf{(}\varphi\mathbf{)} = \left|acc^{-1}(\varphi)\right| = \textit{the number of lists } \ell \in X^K \textit{ with } acc(\ell) = \varphi.$$

Proof. We use induction on the number of elements of the support $supp(\varphi)$ of the multiset φ. If this number is 0, then $\varphi = \mathbf{0}$, with $\mathbf{(0)} = 1$. And indeed, there is precisely one list ℓ with $acc(\ell) = \mathbf{0}$, namely the empty list [].

Next suppose that $supp(\varphi) = \{x_1, \ldots, x_n, x_{n+1}\}$. Take $m := \varphi(x_{n+1})$ and $\varphi' := \varphi - m|x_{n+1}\rangle$ so that $\|\varphi'\| = K - m$ and $supp(\varphi') = \{x_1, \ldots, x_n\}$. By the induction hypothesis there are $\mathbf{(}\varphi'\mathbf{)}$-many lists $\ell' \in X^{K-m}$ with $acc(\ell') = \varphi'$. Each such list ℓ' can be extended to a list ℓ with $acc(\ell) = \varphi$ by m times adding x_{n+1} to ℓ'. How many such additions are there? It is not hard to see that this number of additions is $\binom{K}{m}$. Thus:

$$\begin{aligned}
\left|acc^{-1}(\varphi)\right| &= \binom{K}{m} \cdot \mathbf{(}\varphi'\mathbf{)} \\
&= \frac{K!}{m! \cdot (K-m)!} \cdot \frac{(K-m)!}{\prod_{i \leq n} \varphi'(x_i)!} \\
&= \frac{K!}{\prod_{i \leq n+1} \varphi(x_i)!} \qquad \text{since } m = \varphi(x_{n+1}) \text{ and } \varphi'(x_i) = \varphi(x_i) \\
&= \mathbf{(}\varphi\mathbf{)}.
\end{aligned}$$

❑

Multinomial coefficients satisfy the following *recurrence relations*.

$$\binom{K-1}{k_1 - 1, \ldots, k_n} + \cdots + \binom{K-1}{k_1, \ldots, k_n - 1} = \binom{K}{k_1, \ldots, k_n} \tag{1.40}$$

A snappy re-formulation, for a natural multiset φ, is:

$$\sum_{x \in supp(\varphi)} \mathbf{(}\varphi - 1|x\rangle\mathbf{)} = \mathbf{(}\varphi\mathbf{)}. \tag{1.41}$$

Multinomial coefficients, see Definition 1.2.1 (3), are useful, for instance in the *Multinomial Theorem* (see *e.g.* Ross (2018)):

$$(r_1 + \cdots + r_n)^K = \sum_{k_i, \sum_i k_i = K} \binom{K}{k_1, \ldots, k_n} \cdot r_1^{k_1} \cdot \ldots \cdot r_n^{k_n}. \tag{1.42}$$

An equivalent formulation using multisets and multiset coefficients is:

$$(r_1 + \cdots + r_n)^K = \sum_{\varphi \in \mathcal{M}[K](\{1,\ldots,n\})} \binom{\varphi}{} \cdot \vec{r}^{\,\varphi} = \sum_{\varphi \in \mathcal{M}[K](\{1,\ldots,n\})} \binom{\varphi}{} \cdot \prod_{1 \le i \le n} r_i^{\varphi(i)}. \tag{1.43}$$

Remark 1.7.3 The above formulation (1.43) of the Multinomial Theorem involves multisets $\varphi \in \mathcal{M}[K](\{1,\ldots,n\})$ over the set $\{1,\ldots,n\}$. The numbers in this set are merely placeholders. The set $\{1,\ldots,n\}$ can thus be replaced by an arbitrary set X, giving another level of abstraction.

Let $\mathcal{M}_\pm(X)$ be the set of 'multisets' with both positive and negative multiplicities. Thus, we write $\mathcal{M}_\pm(X) = \{\psi \colon X \to \mathbb{R} \mid \psi \text{ has finite support}\}$. We extend the size formulation $\|\psi\| = \sum_{x \in X} \psi(x)$ to such $\psi \in \mathcal{M}_\pm(X)$. The Multinomial Theorem can now be formulated purely in terms of multisets: for $\psi \in \mathcal{M}_\pm(X)$,

$$\|\psi\|^K = \sum_{\varphi \in \mathcal{M}[K](X)} \binom{\varphi}{} \cdot \prod_{x \in X} \psi(x)^{\varphi(x)}. \tag{1.44}$$

There is an 'infinite' version of the Multinomial Theorem, known as the (Binomial) Series Theorem. It holds more generally than formulated in the first item below, for complex numbers, with adapted meaning of the binomial coefficient, but that's beyond the current scope.

Theorem 1.7.4 *Fix a natural number K.*

1 For a real number $r \in [0,1)$,

$$\sum_{n \ge 0} \binom{n+K}{K} \cdot r^n = \frac{1}{(1-r)^{K+1}}.$$

When $K \ge 1$ this can be reformulated as:

$$\sum_{n \ge 0} \left(\!\binom{K}{n}\!\right) \cdot x^n = \frac{1}{(1-x)^K}$$

2 As a special case,

$$\sum_{n \ge 0} r^n = \frac{1}{1-r}.$$

3 Another consequence is, still for $r \in [0,1)$,

$$\sum_{n \ge 1} n \cdot r^n = \frac{r}{(1-r)^2}.$$

4 For $r_1, \ldots, r_m \in [0,1]$ with $\sum_i r_i < 1$,

$$\sum_{n_1,\ldots,n_m \ge 0} \binom{K + \sum_i n_i}{K, n_1, \ldots, n_m} \cdot \prod_i r_i^{n_i} = \frac{1}{(1 - \sum_i r_i)^{K+1}}.$$

Equivalently,

$$\sum_{\varphi \in \mathcal{N}(\{1,\ldots,m\})} \binom{K + \|\varphi\|}{K} \cdot \binom{\varphi}{} \cdot \vec{r}^{\,\varphi} = \frac{1}{(1 - \sum_i r_i)^{K+1}}.$$

Proof. 1 The equation arises as the Taylor series $f(x) = \sum_n \frac{f^{(n)}(0)}{n!} \cdot x^n$ of the function $f(x) = \frac{1}{(1-x)^{K+1}}$. One can show, by induction on n, that the n-th derivative of f is:

$$f^{(n)}(x) = \frac{(n+K)!}{K!} \cdot \frac{1}{(1-x)^{n+K+1}}.$$

2 The second equation is a special case of the first one, for $K = 0$. There is also a simple direct proof. Define $s_n = r^0 + r^1 + \cdots + r^n$. Then $s_n - r \cdot s_n = 1 - r^{n+1}$, so that $s_n = \frac{1-r^{n+1}}{1-r}$. Hence $s_n \to \frac{1}{1-r}$ as $n \to \infty$.

3 We choose to use the first item, but there are other ways to prove this result, see Exercise 1.7.20.

$$\begin{aligned}\frac{r}{(1-r)^2} &= r \cdot \sum_{n\geq 0} \binom{n+1}{1} \cdot r^n \qquad \text{by item (1), with } K = 1\\ &= \sum_{n\geq 0} (n+1) \cdot r^{n+1} = \sum_{n\geq 1} n \cdot r^n.\end{aligned}$$

4 The trick is to turn the multiple sums into a single 'leading' one, in:

$$\begin{aligned}\sum_{n_1,\ldots,n_m\geq 0} \binom{K+\sum_i n_i}{K, n_1, \ldots, n_m} \cdot \prod_i r_i^{n_i} &= \sum_{n\geq 0} \sum_{n_1,\ldots,n_m,\sum_i n_i = n} \binom{K+n}{K, n_1, \ldots, n_m} \cdot \prod_i r_i^{n_i}\\ &= \sum_{n\geq 0} \sum_{n_1,\ldots,n_m,\sum_i n_i = n} \binom{K+n}{K} \cdot \binom{n}{n_1, \ldots, n_m} \cdot \prod_i r_i^{n_i}\\ &= \sum_{n\geq 0} \binom{K+n}{K} \cdot \sum_{n_1,\ldots,n_m,\sum_i n_i = n} \binom{n}{n_1, \ldots, n_m} \cdot \prod_i r_i^{n_i}\\ &\overset{(1.42)}{=} \sum_{n\geq 0} \binom{K+n}{K} \cdot \left(\sum_i r_i\right)^n\\ &= \frac{1}{(1-\sum_i r_i)^{K+1}}, \qquad \text{by item (1).}\end{aligned}$$

❑

Exercises

1.7.1 Consider the function $f\colon \{a,b,c\} \to \{0,1\}$ given by $f(a) = f(b) = 1$ and $f(c) = 0$.

1 Take the natural multiset $\varphi = 1|a\rangle + 3|b\rangle + 1|c\rangle \in \mathcal{N}(\{a,b,c\})$ and compute consecutively:

- $\mathbf{(}\varphi\mathbf{)}$
- $\mathcal{N}(f)(\varphi)$
- $\mathbf{(}\mathcal{N}(f)(\varphi)\mathbf{)}$.

Conclude that $\mathbf{(}\varphi\mathbf{)} \neq \mathbf{(}\mathcal{N}(f)(\varphi)\mathbf{)}$, in general.

2 Now take $\psi = 2|0\rangle + 3|1\rangle \in \mathcal{N}(\{0,1\})$.

- Compute $\mathbf{(}\psi\mathbf{)}$.
- Show that there are four multisets $\varphi_1, \varphi_2, \varphi_3, \varphi_4 \in \mathcal{N}(\{a,b,c\})$ with $\mathcal{M}(f)(\varphi_i) = \psi$, for each i.
- Check that $\mathbf{(}\psi\mathbf{)} \neq \mathbf{(}\varphi_1\mathbf{)} + \mathbf{(}\varphi_2\mathbf{)} + \mathbf{(}\varphi_3\mathbf{)} + \mathbf{(}\varphi_4\mathbf{)}$.

What is the general formulation $\neq$ now?

1.7.2 Check that:

1 the size map $\|-\|\colon \mathcal{M}(X) \to \mathbb{R}_{\geq 0}$ is a homomorphism of monoids, preserving rescaling — and thus a homomorphism of cones, see Exercise 1.6.13;

2 $\|\mathcal{M}(f)(\varphi)\| = \|\varphi\|$.

1.7.3 Check that the equation (1.38) that defines the accumulation function *acc* corresponds to the following commuting diagram, involving the unit and sum of natural multisets.

$$\begin{tikzcd} X^n \arrow[r, "unit^n"] \arrow[dr, "acc"'] & \mathcal{N}(X)^n \arrow[d, "+"] \\ & \mathcal{N}(X) \end{tikzcd}$$

1.7.4 Show that for natural multisets $\varphi, \psi \in \mathcal{N}(X)$,

$$\varphi \leq \psi \iff \exists \varphi' \in \mathcal{N}(X).\, \varphi + \varphi' = \psi.$$

1.7.5 Recall the obvious fact that for each natural number $N \in \mathbb{N}$ the subset $\{n \in \mathbb{N} \mid n \leq N\}$ has $N + 1$ elements. Show now that for each natural multiset $\psi \in \mathcal{N}(X)$,

$$\Big|\{\varphi \in \mathcal{N}(X) \mid \varphi \leq \psi\}\Big| = \prod_{x \in supp(\psi)} (\psi(x) + 1).$$

1.7.6 Let $\Psi \in \mathcal{N}(\mathcal{N}(X))$ be given with $\psi = flat(\Psi)$. Show that for $\varphi \in supp(\Psi)$ one has $\varphi \leq \psi$ and $flat(\Psi - 1|\varphi\rangle) = \psi - \varphi$.

1.7.7 Let φ be a natural multiset.

1 Show that:

$$\sum_{x \in supp(\varphi)} \frac{\varphi!}{(\varphi - 1|x\rangle)!} = \|\varphi\|.$$

2 Derive the recurrence relation (1.41) from this equation.

1.7.8 Prove the multiset formulation (1.43) of the Multinomial Theorem, by induction on K.

1.7.9 Let $K \geq 0$ and let set X have $n \geq 1$ elements.

1 Prove that:

$$\sum_{\varphi \in \mathcal{N}[K](X)} \binom{\varphi}{} = n^K.$$

This generalises the sum-formula for binomial coefficients from Exercise 1.2.4.
Hint: Use $n = 1 + \cdots + 1$ in (1.43).

2 Show also that for a permutation $\pi \colon X \xrightarrow{\cong} X$ one has $\binom{\mathcal{M}(\pi)(\varphi)}{} = \binom{\varphi}{}$, for each multiset $\varphi \in \mathcal{N}[K](X)$.

1.7.10 Let set X have $N > 0$ elements. Show that for $K \in \mathbb{N}_{>0}$ and $x \in X$ one has:

$$\sum_{\varphi \in \mathcal{N}[K](X)} \binom{\varphi}{} \cdot \varphi(x) = K \cdot N^{K-1}.$$

1.7.11 Fix $N \geq 0$ and consider the obvious addition function $sum_N \colon \mathbb{N}^N \to \mathbb{N}$ given by $sum_N(k_1, \ldots, k_N) = \sum_i k_i$.

1 Use the previous exercise to show that for each $m \in \mathbb{N}$,

$$\Big|(sum_N)^{-1}(m)\Big| = N^m.$$

2 Now restrict addition to non-negative numbers $\mathbb{N}_{>0}$, for which we use the *ad hoc* notation $sum^*_N \colon (\mathbb{N}_{>0})^N \to \mathbb{N}_{>0}$. Show now that for $m \geq N$,

$$\Big|(sum^*_N)^{-1}(m)\Big| = \binom{m-1}{N-1}.$$

1.7.12 Let φ be a natural multiset. Show that:

1 $(\varphi) = 1 \iff supp(\varphi)$ is a singleton;
2 $(\varphi) = \|\varphi\|! \iff \forall x \in supp(\varphi).\, \varphi(x) = 1 \iff \varphi$ consists of singletons.

1.7.13 Show that $\|acc(\ell)\| = \|\ell\|$, using the length $\|\ell\|$ of a list ℓ from Exercise 1.4.4.

1.7.14 Consider the K-ary accumulator function (1.39), for $K > 0$. Check that *acc* is invariant under transposition, in the sense that for each permutation π of the positions $\{1, \ldots, K\}$ one has:

$$acc(x_1, \ldots, x_K) \;=\; acc(x_{\pi(1)}, \ldots, x_{\pi(K)}).$$

1.7.15 1 Check that the accumulator map $acc\colon \mathcal{L}(X) \to \mathcal{N}(X)$ is a homomorphism of monoids.

2 Check also that it arises via freeness of lists, as extension of the function $unit^{\mathcal{N}}\colon X \to \mathcal{N}(X)$, following Proposition 1.4.3:

$$\begin{tikzcd} X \arrow[r, "unit^{\mathcal{L}}"] \arrow[dr, bend right, "unit^{\mathcal{N}}"'] & \mathcal{L}(X) \arrow[d, dashed, "acc = \overline{unit^{\mathcal{N}}}"] \\ & \mathcal{N}(X) \end{tikzcd}$$

3 Prove that the accumulator map is natural: for an arbitrary function $f\colon X \to Y$ the rectangle below commutes.

$$\begin{tikzcd} \mathcal{L}(X) \arrow[r, "acc"] \arrow[d, "\mathcal{L}(f)"'] & \mathcal{N}(X) \arrow[d, "\mathcal{M}(f)"] \\ \mathcal{L}(Y) \arrow[r, "acc"'] & \mathcal{N}(Y) \end{tikzcd}$$

1.7.16 The diagram below displays the fundamental relationships between lists, multisets and subsets via natural transformations, see also diagram (1.1) and (1.37).

$$\begin{tikzcd} \mathcal{L} \arrow[r, Rightarrow, "acc", "\text{forget order}"'] & \mathcal{N} \arrow[r, Rightarrow, "supp", "\text{forget multiplicity}"'] & \mathcal{P} \end{tikzcd} \tag{1.45}$$

It is a natural to ask: is there also a functor that captures order but not multiplicity. This could be achieved via lists in which elements occur (at most) once. Explicitly, for a set X, we define:

$$\mathcal{L}_{\leq 1}(X) \coloneqq \{\langle x_1, \ldots, x_n\rangle \in \mathcal{L}(X) \mid \forall 1 \leq i, j \leq n.\, x_i = x_j \Rightarrow i = j\}.$$

Give an example demonstrating that $\mathcal{L}_{\leq 1}$ is *not* a functor **Sets** $\to$ **Sets**, when $\mathcal{L}_{\leq 1}$ is the same as $\mathcal{L}$ on functions.

1.7.17 Convince yourself that the following composite

$$\mathcal{N}[K](X)^L \xrightarrow{acc} \mathcal{N}[L]\big(\mathcal{N}[K](X)\big) \xrightarrow{flat} \mathcal{N}[L{\cdot}K](X)$$

is the L-fold sum of multisets.

1.7.18 Prove that the following diagram commutes, where *swap* is the obvious isomorphism, comparable to transposition of a matrix: $swap(h)(i)(j) = h(j)(i)$.

$$\begin{tikzcd}
(X^K)^L \arrow[r, "acc^L"] \arrow[d, "swap"', "\cong"] & \mathcal{N}[K](X)^L \arrow[r, "acc"] \arrow[dr, bend left, "+"] & \mathcal{N}[L](\mathcal{N}[K](X)) \arrow[d, "flat"] \\
 & & \mathcal{N}[L{\cdot}K](X) \\
(X^L)^K \arrow[r, "acc^K"] & \mathcal{N}[L](X)^K \arrow[r, "acc"'] \arrow[ur, bend right, "+"'] & \mathcal{N}[K](\mathcal{N}[L](X)) \arrow[u, "flat"']
\end{tikzcd}$$

The two triangles on the right commute by the the previous exercise. For commutation of the subdiagram on the left one can use that accumulation is a homorphism of monoids, see Exercise 1.7.15 (1).

1.7.19 Let $n \geq 1$ and $r \in (0, 1)$. Show that:

$$\sum_{k \geq n} \binom{k-1}{n-1} \cdot r^n \cdot (1-r)^{k-n} \;=\; 1.$$

1.7.20 Elaborate the details of the following two (alternative) proofs of the equation in Theorem 1.7.4 (3).

1 Use the derivate $\frac{d}{dr}$ on both sides of Theorem 1.7.4 (2).

2 Write $s := \sum_{n \geq 1} n \cdot r^n$ and exploit the following recursive equation.

$$\begin{aligned}
s \;&=\; r + 2r^2 + 3r^3 + 4r^4 + \cdots \\
&=\; r + (1+1)r^2 + (1+2)r^3 + (1+3)r^4 + \cdots \\
&=\; (r + r^2 + r^3 + r^4 + \cdots) \;+\; r \cdot (r + 2r^2 + 3r^3 + \cdots) \\
&=\; \frac{r}{1-r} \;+\; r \cdot s, \qquad \text{by Theorem 1.7.4 (2).}
\end{aligned}$$

1.7.21 In the proof of Theorem 1.7.4 we have used Taylor's formula for a single-variable function. For multi-variable functions we can use multisets for a compact, 'multi-index' formulation. For an an n-ary function $f(x_1, \ldots, x_n)$ and a natural multiset $\varphi \in \mathcal{N}(\{1, \ldots, n\})$ write:

$$\partial^\varphi f \;:=\; (\partial x_1)^{\varphi(1)} \cdots (\partial x_n)^{\varphi(n)} f.$$

Check that Taylor's expansion formula (around $0 \in \mathbb{R}^n$) then becomes:

$$f(\vec{x}) \;=\; \sum_{\varphi \in \mathcal{N}(\{1,\ldots,n\})} \frac{(\partial^\varphi f)(0)}{\varphi!} \cdot \vec{x}^{\,\varphi}.$$

1.8 Coefficients of multisets

Binomial coefficients $\binom{n}{k}$ for numbers $n \geq k$ are a standard tool in many areas of (discrete) mathematics, see Section 1.2 for the definition and the most basic properties. Here we extend binomial coefficients from numbers to natural multisets: we define $\binom{\psi}{\varphi}$ for natural multisets ψ, φ with $\psi \geq \varphi$. In the next section we shall also look at the extension of the less familiar 'multichoose' coefficients $\left(\!\binom{n}{m}\!\right)$ to multiset form $\left(\!\binom{\psi}{\varphi}\!\right)$. The coefficients $\binom{\psi}{\varphi}$ and $\left(\!\binom{\psi}{\varphi}\!\right)$ will play an important role in (multivariate) hypergeometric and Pólya distributions.

Definition 1.8.1 Let X be an arbitrary set.

1 For two natural multisets $\varphi, \psi \in \mathcal{N}(X)$ with $\varphi \leq \psi$, the *multiset binomial* is defined as:

$$\begin{aligned}\binom{\psi}{\varphi} &:= \frac{\psi\,\underset{\circ}{!}}{\varphi\,\underset{\circ}{!} \cdot (\psi - \varphi)\,\underset{\circ}{!}} \\ &= \frac{\prod_x \psi(x)!}{(\prod_x \varphi(x)!) \cdot (\prod_x (\psi(x) - \varphi(x))!)} = \prod_{x \in supp(\psi)} \binom{\psi(x)}{\varphi(x)}.\end{aligned} \tag{1.46}$$

2 For a multiset $\varphi \in \mathcal{N}(X)$ and a number $n \geq \|\varphi\|$ we define a 'mixed number & multiset' binomial coefficient:

$$\binom{n}{\varphi} := \frac{n!}{\varphi\,\underset{\circ}{!} \cdot (n - \|\varphi\|)!}. \tag{1.47}$$

The second coefficient $\binom{n}{\varphi}$ will occur only occasionally — for instance in Lemma 1.9.6 — but the first one $\binom{\psi}{\varphi}$ with two multisets will play a central role. Here is an illustration:

$$\binom{3|R\rangle + 2|B\rangle}{2|R\rangle + 1|B\rangle} = \frac{3! \cdot 2!}{(2! \cdot 1!) \cdot (1! \cdot 1!)} = 6 = 3 \cdot 2 = \binom{3}{2} \cdot \binom{2}{1}.$$

The following result is known as Vandermonde's formula, see *e.g.* Graham et al. (1994); Benjamin and Quinn (2003). We reformulate it for numbers $B, G \in \mathbb{N}$ and $K \leq B + G$. It says:

$$\binom{B+G}{K} = \sum_{b \leq B,\, g \leq G,\, b+g=K} \binom{B}{b} \cdot \binom{G}{g}. \tag{1.48}$$

Intuitively: if you select K children out of B boys and G girls, the number of options is given by the sum over the options for $b \leq B$ boys times the options for $g \leq G$ girls, with $b + g = K$.

Lemma 1.8.2 *Let $\psi \in \mathcal{N}(X)$ be a multiset of size $L = \|\psi\|$, with a number $K \leq L$. Then:*

$$\sum_{\varphi \leq_K \psi} \binom{\psi}{\varphi} = \binom{L}{K} \qquad \textit{so that} \qquad \sum_{\varphi \leq_K \psi} \frac{\binom{\psi}{\varphi}}{\binom{L}{K}} = 1.$$

We recall that $\varphi \leq_K \psi$ means that $\varphi \leq \psi$ and $\|\varphi\| = K$.

These fractions adding up to one will form the probabilities of the so-called hypergeometric distribution, see Subsection 2.6.1 and Section 3.5 later on.

Proof. We use induction on the number of elements in $supp(\psi)$. We go through some initial values explicitly. If the number of elements in $supp(\psi)$ is 0, then $\psi = \mathbf{0}$ and so $L = 0 = K$ and $\varphi \leq_K \psi$ means $\varphi = \mathbf{0}$, so that the result holds. Similarly, if $supp(\psi)$ is a singleton, say $\{x\}$, then $L = \psi(x)$. For $K \leq L$ and $\varphi \leq_K \psi$ we have $\varphi = K|x\rangle$ as the sole draw. The result then obviously holds.

The binary case of the lemma, where $supp(\psi) = \{x, y\}$, corresponds the ordinary form of Vandermonde's formula (1.48). We briefly show how this equation (1.48) can be proven by induction on $G \in \mathbb{N}$. When $G = 0$ both sides amount to $\binom{B}{K}$ so we quickly proceed to the induction step.

The case $K = 0$ is trivial, so we may assume $K > 0$.

$$\begin{aligned}
\sum_{b\le B,\, g\le G+1,\, b+g=K} \binom{B}{b}\cdot\binom{G+1}{g} &= \binom{B}{K}\cdot\binom{G+1}{0}+\binom{B}{K-1}\cdot\binom{G+1}{1} \\
&\quad + \cdots + \binom{B}{K-G}\cdot\binom{G+1}{G}+\binom{B}{K-G-1}\cdot\binom{G+1}{G+1} \\
&\overset{(1.2)}{=} \binom{B}{K}\cdot\binom{G}{0}+\binom{B}{K-1}\cdot\binom{G}{1}+\binom{B}{K-1}\cdot\binom{G}{0} \\
&\quad + \cdots + \binom{B}{K-G}\cdot\binom{G}{G}+\binom{B}{K-G}\cdot\binom{G}{G-1}+\binom{B}{K-G-1}\cdot\binom{G}{G} \\
&= \sum_{b\le B,\, g\le G,\, b+g=K} \binom{B}{b}\cdot\binom{G}{g} + \sum_{b\le B,\, g\le G,\, b+g=K-1} \binom{B}{b}\cdot\binom{G}{g} \\
&\overset{(\mathrm{IH})}{=} \binom{B+G}{K}+\binom{B+G}{K-1} \\
&\overset{(1.2)}{=} \binom{B+G+1}{K}.
\end{aligned}$$

For the induction step, let $supp(\psi) = \{x_1, \ldots, x_n, y\}$, for $n \ge 2$. Writing $\ell = \psi(y)$, $L' = L - \ell$ and $\psi' = \psi - \ell|y\rangle \in \mathcal{N}[L'](\{x_1, \ldots, x_n\})$ gives:

$$\begin{aligned}
\sum_{\varphi\le_K\psi} \binom{\psi}{\varphi} = \sum_{\varphi\le_K\psi} \prod_x \binom{\psi(x)}{\varphi(x)} &= \sum_{n\le\ell} \sum_{\varphi\le_{K-n}\psi'} \binom{\ell}{n}\cdot\prod_i \binom{\psi'(x_i)}{\varphi(x_i)} \\
&\overset{(\mathrm{IH})}{=} \sum_{n\le\ell,\, K-n\le L-\ell} \binom{\ell}{n}\cdot\binom{L-\ell}{K-n} \overset{(1.48)}{=} \binom{L}{K}.
\end{aligned}$$

❑

The next result connects multinomial and binomial coefficients of multisets.

Lemma 1.8.3 *Let $\psi \in \mathcal{N}[L](X)$ be a natural multiset and let $K \le L = \|\psi\|$.*

1 For $\varphi \le_K \psi$ one has:

$$\frac{(\varphi)\cdot(\psi-\varphi)}{(\psi)} = \frac{\binom{\psi}{\varphi}}{\binom{L}{K}}.$$

2 Now:

$$(\psi) = \sum_{\varphi\le_K\psi} (\varphi)\cdot(\psi-\varphi).$$

The earlier equation (1.41) is a special case, for $K = 1$.

Proof. 1 Because:

$$\frac{(\varphi)\cdot(\psi-\varphi)}{(\psi)} = \frac{K!}{\varphi!}\cdot\frac{(L-K)!}{(\psi-\varphi)!}\cdot\frac{\psi!}{L!} = \frac{K!\cdot(L-K)!}{L!}\cdot\frac{\psi!}{\varphi!\cdot(\psi-\varphi)!} = \frac{\binom{\psi}{\varphi}}{\binom{L}{K}}.$$

2 By the previous item and Vandermonde's formula from Lemma 1.8.2:

$$\sum_{\varphi\le_K\psi} (\varphi)\cdot(\psi-\varphi) = (\psi)\cdot\frac{\sum_{\varphi\le_K\psi}\binom{\psi}{\varphi}}{\binom{L}{K}} = (\psi)\cdot\frac{\binom{L}{K}}{\binom{L}{K}} = (\psi).$$

❑

1.8.1 Multichoose coefficents

We proceed with another counting challenge. Let X be a finite set, say with n elements. How many elements does the set of multisets $\mathcal{N}[K](X)$ have? That is, how many multisets of size K are there over n elements? This is sometimes formulated informally as: how many ways are there to divide K balls over n urns? It is the multiset-analogue of Exercise 1.5.6, where the number of *subsets* of size K of an n-element set is identified as $\binom{n}{K}$. Below we show that the answer for multisets is given by the *multiset number* or *multichoose number* $\left(\!\binom{n}{K}\!\right)$, see *e.g.* Ross (2018). We have introduce this multichoose number in Section 1.2 as $\left(\!\binom{n}{K}\!\right) = \binom{n+K-1}{K}$; here we extend it from numbers to multisets, in analogy with the extension of (ordinary) binomials to multisets in Definition 1.7.1.

Definition 1.8.4 Let ψ, φ be natural multisets over the same set X, where ψ is non-zero and $supp(\varphi) \subseteq supp(\psi)$.

$$\left(\!\binom{\psi}{\varphi}\!\right) := \prod_{x \in supp(\psi)} \left(\!\binom{\psi(x)}{\varphi(x)}\!\right) = \prod_{x \in supp(\psi)} \binom{\psi(x) + \varphi(x) - 1}{\varphi(x)}. \tag{1.49}$$

Consider the set $\mathcal{N}[3](\{a, b, c\})$ of multisets of size 3 over $\{a, b, c\}$. It has $\left(\!\binom{3}{3}\!\right) = \binom{5}{3} = \frac{4 \cdot 5}{2} = 10$ elements, namely:

$$\begin{array}{cccccc} 3|a\rangle & 3|b\rangle & 3|c\rangle & 2|a\rangle + 1|b\rangle & 2|a\rangle + 1|c\rangle & 1|a\rangle + 2|b\rangle, \\ 2|b\rangle + 1|c\rangle & 1|a\rangle + 2|c\rangle & 1|b\rangle + 2|c\rangle & 1|a\rangle + 1|b\rangle + 1|c\rangle. & & \end{array}$$

Below, Proposition 1.8.7 states in general that $\left(\!\binom{n}{K}\!\right)$ is the number of multisets of size K over a non-empty n-element set. We first need to obtain a multichoose analogue of Vandermonde's (binary) formula (1.48). It may be found in (Rota and Baclawski, 1979, II.5).

Lemma 1.8.5 *Fix numbers $B \geq 1$ and $G \geq 1$. For all K on has:*

$$\left(\!\binom{B+G}{K}\!\right) = \sum_{0 \leq k \leq K} \left(\!\binom{B}{k}\!\right) \cdot \left(\!\binom{G}{K-k}\!\right). \tag{1.50}$$

In particular:

$$\binom{B+K}{K} = \left(\!\binom{B+1}{K}\!\right) = \sum_{0 \leq k \leq K} \left(\!\binom{B}{k}\!\right) = \sum_{0 \leq k \leq K} \left(\!\binom{B}{K-k}\!\right). \tag{1.51}$$

Proof. The second equation (1.51) easily follows from the first one by taking $G = 1$ and using that $\left(\!\binom{1}{n}\!\right) = \binom{n}{n} = 1$.

We shall make frequent use of the following equation, whose proof is left to the reader (in Exercise 1.8.12) below.

$$\left(\!\binom{n}{K+1}\!\right) + \left(\!\binom{n+1}{K}\!\right) = \left(\!\binom{n+1}{K+1}\!\right). \tag{$*$}$$

We shall prove the first equation (1.50) in the lemma by induction on $B \geq 1$. In both the base case $B = 1$ and the induction step we shall use induction on K. We shall try to keep the structure clear by using nested bullets.

- We first prove Equation (1.50) for $B = 1$, by induction on K.
  - When $K = 0$ both sides in (1.50) are equal to 1.

– Assume Equation (1.50) holds for K (and $B = 1$).

$$\begin{aligned}\sum_{0\le k\le K+1}\left(\!\!\binom{1}{k}\!\!\right)\cdot\left(\!\!\binom{G}{(K+1)-k}\!\!\right) &= \sum_{0\le k\le K+1}\left(\!\!\binom{G}{K-(k-1)}\!\!\right)\\ &= \left(\!\!\binom{G}{K+1}\!\!\right)+\sum_{0\le \ell\le K}\left(\!\!\binom{G}{K-\ell}\!\!\right)\\ &\overset{\text{(IH)}}{=} \left(\!\!\binom{G}{K+1}\!\!\right)+\left(\!\!\binom{G+1}{K}\!\!\right) \overset{(*)}{=} \left(\!\!\binom{G+1}{K+1}\!\!\right).\end{aligned}$$

- Now assume Equation (1.50) holds for B (for all G, K). In order to show that it then also holds for $B + 1$ we use induction on K.

  – When $K = 0$ both sides in (1.50) are equal to 1.

  – Now assume that Equation (1.50) holds for K, and for B. Then:

$$\begin{aligned}&\sum_{0\le k\le K+1}\left(\!\!\binom{B+1}{k}\!\!\right)\cdot\left(\!\!\binom{G}{(K+1)-k}\!\!\right)\\ &= \left(\!\!\binom{G}{K+1}\!\!\right)+\sum_{0\le k\le K}\left(\!\!\binom{B+1}{k+1}\!\!\right)\cdot\left(\!\!\binom{G}{K-k}\!\!\right)\\ &\overset{(*)}{=} \left(\!\!\binom{G}{K+1}\!\!\right)+\sum_{0\le k\le K}\left[\left(\!\!\binom{B}{k+1}\!\!\right)+\left(\!\!\binom{B+1}{k}\!\!\right)\right]\cdot\left(\!\!\binom{G}{K-k}\!\!\right)\\ &= \left(\!\!\binom{G}{K+1}\!\!\right)+\sum_{0\le k\le K}\left(\!\!\binom{B}{k+1}\!\!\right)\cdot\left(\!\!\binom{G}{K-k}\!\!\right)+\sum_{0\le k\le K}\left(\!\!\binom{B+1}{k}\!\!\right)\cdot\left(\!\!\binom{G}{K-k}\!\!\right)\\ &\overset{\text{(IH, K)}}{=} \sum_{0\le k\le K+1}\left(\!\!\binom{B}{k}\!\!\right)\cdot\left(\!\!\binom{G}{(K+1)-k}\!\!\right)+\left(\!\!\binom{(B+1)+G}{K}\!\!\right)\\ &\overset{\text{(IH, B)}}{=} \left(\!\!\binom{B+G}{K+1}\!\!\right)+\left(\!\!\binom{(B+1)+G}{K}\!\!\right) \overset{(*)}{=} \left(\!\!\binom{(B+1)+G}{K+1}\!\!\right).\end{aligned}$$

❑

We now get the double-bracket analogue of Lemma 1.8.2.

Proposition 1.8.6 *Let ψ be a non-empty natural multiset. Write $X = supp(\psi)$ and $L = \|\psi\|$. Then, for each $K \in \mathbb{N}$,*

$$\sum_{\varphi\in\mathcal{N}[K](X)}\left(\!\!\binom{\psi}{\varphi}\!\!\right) = \left(\!\!\binom{L}{K}\!\!\right) \qquad \textit{so} \qquad \sum_{\varphi\in\mathcal{N}[K](X)}\frac{\left(\!\!\binom{\psi}{\varphi}\!\!\right)}{\left(\!\!\binom{L}{K}\!\!\right)} = 1.$$

The fractions in this equation will show up later in so-called Pólya distributions, see Subsection 2.6.2 and Section 3.6. These fractions capture the probability of drawing a multiset φ from an urn ψ when for each drawn ball an extra ball is added to the urn (of the same colour).

Proof. We use induction on the number of elements in the support X of ψ, like in the proof of Lemma 1.8.2. By assumption X cannot be empty, so the induction starts when X is a singleton, say $X = \{x\}$. But then $\psi(x) = \|\psi\| = L$ and $\varphi(x) = \|\varphi\| = K$, so the result obviously holds.

Now let $supp(\psi) = X \cup \{y\}$ where $y \notin X$ and X is not empty. Write:

$$L = \|\psi\| \qquad \ell = \psi(y) > 0 \qquad \psi' = \psi - \ell|y\rangle \qquad L' = L - \ell > 0.$$

By construction $X = supp(\psi')$ and $L' = \|\psi'\|$. Now:

$$\begin{aligned}
\sum_{\varphi\in\mathcal{N}[K](X\cup\{y\})} \left(\!\!\binom{\psi}{\varphi}\!\!\right) &\overset{(1.49)}{=} \sum_{\varphi\in\mathcal{N}[K](X\cup\{y\})} \prod_{x\in X\cup\{y\}} \left(\!\!\binom{\psi(x)}{\varphi(x)}\!\!\right) \\
&= \sum_{0\le k\le K} \sum_{\varphi\in\mathcal{N}[K-k](X)} \left(\!\!\binom{\psi(y)}{k}\!\!\right) \cdot \prod_{x\in X} \left(\!\!\binom{\psi(x)}{\varphi(x)}\!\!\right) \\
&= \sum_{0\le k\le K} \left(\!\!\binom{\ell}{k}\!\!\right) \cdot \sum_{\varphi\in\mathcal{N}[K-k](X)} \left(\!\!\binom{\psi}{\varphi}\!\!\right) \\
&\overset{(\mathrm{IH})}{=} \sum_{0\le k\le K} \left(\!\!\binom{\ell}{k}\!\!\right) \cdot \left(\!\!\binom{L'}{K-k}\!\!\right) \overset{(1.50)}{=} \left(\!\!\binom{\ell+L'}{K}\!\!\right) = \left(\!\!\binom{L}{K}\!\!\right).
\end{aligned}$$

❑

We finally come to our multiset counting result. It is the multiset-analogue of Exercise 1.5.6 for subsets, saying that $\binom{n}{K}$ is the number of *subsets* of size K of a set with n elements.

Proposition 1.8.7 *Let X be a non-empty set with $n \ge 1$ elements. The number of natural multisets of size K over X is $\left(\!\!\binom{n}{K}\!\!\right)$, that is:*

$$\left|\mathcal{N}[K](X)\right| = \left(\!\!\binom{n}{K}\!\!\right) = \binom{n+K-1}{K}.$$

Proof. The statement holds for $K = 0$ since there is precisely $1 = \binom{n-1}{0} = \left(\!\!\binom{n}{0}\!\!\right)$ multiset set of 0, namely the empty multiset $\mathbf{0}$. Hence we may assume $K \ge 1$, so that Lemma 1.8.5 can be used.

We proceed by induction on $n \ge 1$. For $n = 1$ the statement holds since there is only $1 = \binom{K}{K} = \left(\!\!\binom{1}{K}\!\!\right)$ multiset of size K over $1 = \{0\}$, namely $K|0\rangle$.

The induction step works as follows. Let the set X have n elements, say $X = \{x_1, \ldots, x_n\}$ and let $y \notin X$. For a multiset $\varphi \in \mathcal{N}[K](X\cup\{y\})$ there are $K+1$ possible multiplicities $\varphi(y)$. If $\varphi(y) = k$, then the number possibilities for $\varphi(x_1), \ldots, \varphi(x_n)$ is the number of multisets in $\mathcal{N}[K-k](X)$. Thus:

$$\begin{aligned}
\left|\mathcal{N}[K](X\cup\{y\})\right| &= \sum_{0\le k\le K} \left|\mathcal{N}[K-k](X)\right| \overset{(\mathrm{IH})}{=} \sum_{0\le k\le K} \left(\!\!\binom{n}{K-k}\!\!\right) \\
&= \left(\!\!\binom{n+1}{K}\!\!\right), \qquad \text{by Lemma 1.8.5.}
\end{aligned}$$

❑

There is also a visual proof of this result, described in terms of stars and bars, see *e.g.* (Feller, 1970, II, proof of (5.2)), where multiplicities of multisets are described in terms of 'occupancy numbers'.

An associated question is: given a fixed element a in an n-element set X, what is the sum of all multiplicities $\varphi(a)$, for multisets φ over X with size K?

Lemma 1.8.8 *For an element $a \in X$, where X has $n \ge 1$ elements,*

$$\sum_{\varphi\in\mathcal{N}[K](X)} \varphi(a) = \frac{K}{n} \cdot \left(\!\!\binom{n}{K}\!\!\right).$$

Proof. When we sum over a we get by Proposition 1.8.7:

$$\sum_{a\in X} \sum_{\varphi\in\mathcal{N}[K](X)} \varphi(a) = \sum_{\varphi\in\mathcal{N}[K](X)} \sum_{a\in X} \varphi(a) = \sum_{\varphi\in\mathcal{N}[K](X)} K = K \cdot \left(\!\!\binom{n}{K}\!\!\right).$$

Since $a \in X$ is arbitrary, the outcome should be the same for a different $b \in X$. Hence we have to divide by n, giving the equation in the lemma. ❑

Exercises

1.8.1 Generalise the familiar equation $\sum_{0 \leq k \leq K} \binom{K}{k} = 2^K$ to:

$$\sum_{\varphi \leq \psi} \binom{\psi}{\varphi} = 2^{\|\psi\|}.$$

1.8.2 Let X be a finite set with $n \geq 1$ elements and let $K \geq n$. Prove:

$$\Big|\{\varphi \in \mathcal{N}[K](X) \;\big|\; \varphi \text{ has full support}\}\Big| = \left(\!\!\binom{n}{K-n}\!\!\right).$$

1.8.3 In analogy with the powerset operator, with type $\mathcal{P}\colon \mathcal{P}(X) \to \mathcal{P}(\mathcal{P}(X))$, a *powerbag* operator $\mathsf{PB}\colon \mathcal{N}(X) \to \mathcal{N}(\mathcal{N}(X))$ is introduced in Libkin and Wong (1993) (see also Dash and Staton (2021)). It can be defined as:

$$\mathsf{PB}(\psi) := \sum_{\varphi \leq \psi} \binom{\psi}{\varphi} \big|\varphi\rangle.$$

1 Take $X = \{a, b\}$ and show that:

$$\begin{aligned}\mathsf{PB}(1|a\rangle + 3|b\rangle) = {} & 1\big|\mathbf{0}\rangle + 1\big|1|a\rangle\rangle + 3\big|1|b\rangle\rangle + 3\big|1|a\rangle + 1|b\rangle\rangle + 3\big|2|b\rangle\rangle \\ & + 3\big|1|a\rangle + 2|b\rangle\rangle + 1\big|3|b\rangle\rangle + 1\big|1|a\rangle + 3|b\rangle\rangle.\end{aligned}$$

2 Check that one can compute the powerbag of ψ as follows. Take a list of elements that accumulate to ψ, such as $[a, b, b, b]$ in the previous item. Take the accumulation of all subsequences.

1.8.4 For $N \geq 2$ many natural multisets $\varphi_1, \ldots, \varphi_N \in \mathcal{N}(X)$, with $\psi := \sum_i \varphi_i$, define a multinomial coefficient of multisets as:

$$\binom{\psi}{\varphi_1, \ldots, \varphi_N} := \frac{\psi!}{\varphi_1! \cdot \ldots \cdot \varphi_N!}.$$

1 Check that for $N \geq 3$, in analogy with Exercise 1.2.6,

$$\binom{\psi}{\varphi_1, \ldots, \varphi_N} = \binom{\psi}{\varphi_1} \cdot \binom{\psi - \varphi_1}{\varphi_2, \ldots, \varphi_N}$$

2 For $K_1, \ldots, K_N \in \mathbb{N}$ write $K = \sum_i K_i$ and assume that $\psi \in \mathcal{N}[K](X)$ is given. Show that:

$$\sum_{\substack{\varphi_1 \leq_{K_1} \psi, \ldots, \varphi_N \leq_{K_N} \psi, \\ \sum_i \varphi_i = \psi}} \binom{\psi}{\varphi_1, \ldots, \varphi_N} = \binom{K}{K_1, \ldots, K_N}.$$

1.8.5 Prove the following multiset versions of Vandermonde's formulas (1.48) and (1.50). Let $\psi_1, \psi_2, \varphi \in \mathcal{N}(X)$ be given.

1 When $\varphi \leq \psi_1 + \psi_2$,

$$\binom{\psi_1 + \psi_2}{\varphi} = \sum_{\varphi_1 \leq \psi_1, \varphi_2 \leq \psi_2, \varphi_1 + \varphi_2 = \varphi} \binom{\psi_1}{\varphi_1} \cdot \binom{\psi_2}{\varphi_2}.$$

2 When $\psi_1 \geq \mathbf{1}$ and $\psi_2 \geq \mathbf{1}$, then:

$$\left(\!\!\binom{\psi_1+\psi_2}{\varphi}\!\!\right) = \sum_{\chi\leq\varphi} \left(\!\!\binom{\psi_1}{\chi}\!\!\right)\cdot\left(\!\!\binom{\psi_2}{\varphi-\chi}\!\!\right).$$

1.8.6 Let $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[L](X)$ be given.

1 Prove that:

$$(\varphi+\psi) = \frac{\binom{K+L}{K}}{\binom{\varphi+\psi}{\varphi}}\cdot(\varphi)\cdot(\psi).$$

2 Now assume that φ, ψ have disjoint supports, that is, $supp(\varphi) \cap supp(\psi) = \emptyset$. Show that now:

$$(\varphi+\psi) = \binom{K+L}{K}\cdot(\varphi)\cdot(\psi),$$

1.8.7 Let φ, ψ be natural multiset on the same finite set X, where $\psi \geq \mathbf{1}$.

1 Show that one has:

$$\left(\!\!\binom{\psi}{\varphi}\!\!\right) = \binom{\psi+\varphi-\mathbf{1}}{\varphi} = \frac{(\psi+\varphi-\mathbf{1})!}{\varphi!\cdot(\psi-\mathbf{1})!}.$$

2 Conclude, analogously to Lemma 1.8.3 (1), that:

$$\frac{(\varphi)\cdot(\psi-\mathbf{1})}{(\psi+\varphi-\mathbf{1})} = \frac{\left(\!\!\binom{\psi}{\varphi}\!\!\right)}{\left(\!\!\binom{L}{K}\!\!\right)}.$$

1.8.8 1 Show that for numbers $m \leq n-1$,

$$n\cdot\binom{n-1}{m} = (m+1)\cdot\binom{n}{m+1} = (n-m)\cdot\binom{n}{m}.$$

2 Show similarly that for natural multisets φ, ψ with $x \in supp(\psi)$ and $\varphi \leq \psi - 1|x\rangle$,

$$\psi(x)\cdot\binom{\psi-1|x\rangle}{\varphi} = (\varphi(x)+1)\cdot\binom{\psi}{\varphi+1|x\rangle} = (\psi(x)-\varphi(x))\cdot\binom{\psi}{\varphi}.$$

3 For $x \in supp(\varphi) \subseteq supp(\psi)$,

$$\varphi(x)\cdot\left(\!\!\binom{\psi}{\varphi}\!\!\right) = \psi(x)\cdot\left(\!\!\binom{\psi+1|x\rangle}{\varphi-1|x\rangle}\!\!\right).$$

4 For $supp(\varphi) \subseteq supp(\psi)$ and $x \in supp(\psi)$,

$$\psi(x)\cdot\left(\!\!\binom{\psi+1|x\rangle}{\varphi}\!\!\right) = (\varphi(x)+\psi(x))\cdot\left(\!\!\binom{\psi}{\varphi}\!\!\right) = (\varphi(x)+1)\cdot\left(\!\!\binom{\psi}{\varphi+1|x\rangle}\!\!\right).$$

1.8.9 Let X be a non-empty finite set, say with $N = |X| > 0$ elements.

1 Let $\varphi \in \mathcal{N}[K](X)$ be a fixed multiset of size K. Use Proposition 1.8.6 to prove, for arbitrary $K \in \mathbb{N}$,

$$\sum_{\psi\in\mathcal{N}[L](X)} \binom{\varphi+\psi}{\psi} = \left(\!\!\binom{K+N}{L}\!\!\right).$$

2 Show in the setting of the previous point that, for an element $x \in X$,

$$\sum_{\psi \in \mathcal{N}[L](X)} \psi(x) \cdot \binom{\varphi + \psi}{\psi} = \left(\!\!\binom{L + N + 1}{K - 1}\!\!\right) \cdot (\varphi(x) + 1) = \left(\!\!\binom{L}{K + N}\!\!\right) \cdot (\varphi(x) + 1).$$

3 Now let $L \geq K$ and let $\varphi \in \mathcal{N}[K](X)$ be given. Use the first point to show that:

$$\sum_{\psi \in \mathcal{N}[L](X),\, \varphi \leq_K \psi} \binom{\psi}{\varphi} = \binom{L + N - 1}{K + N - 1}.$$

1.8.10 Let X be a finite set with $N := |X| > 0$ elements and let $L \geq N$ be given. Fix a multiset $\varphi \in \mathcal{N}[K]$.

1 Use Proposition 1.8.6 to prove that:

$$\sum_{\psi \in \mathcal{N}_{fs}[L](X)} \left(\!\!\binom{\psi}{\varphi}\!\!\right) = \left(\!\!\binom{K + N}{L - N}\!\!\right).$$

2 Show via Exercise 1.8.9 (1), (2) that for $x \in X$,

$$\sum_{\psi \in \mathcal{N}_{fs}[L](X)} \psi(x) \cdot \left(\!\!\binom{\psi}{\varphi}\!\!\right) = \left(\!\!\binom{L - N}{K + N}\!\!\right) \cdot \varphi(x) + \binom{L + K}{L - N}.$$

1.8.11 Let $N \geq 0$ and $M \geq m \geq 1$ and be given. Use the multichoose Vandermonde Equation (1.51) to prove:

$$\sum_{0 \leq i \leq N} \left(\!\!\binom{m}{i}\!\!\right) \cdot \binom{M - m + N - i}{N - i} = \binom{N + M}{N}.$$

1.8.12 Prove the following properties.

1 $\left(\!\!\binom{n-k}{m-(k+1)}\!\!\right) = \left(\!\!\binom{m-k}{n-(k+1)}\!\!\right)$

2 $\left(\!\!\binom{n}{m}\!\!\right) + \left(\!\!\binom{m}{n}\!\!\right) = \binom{n+m}{n}$

3 $m \cdot \left(\!\!\binom{n}{m}\!\!\right) = n \cdot \left(\!\!\binom{n+1}{m-1}\!\!\right)$.

4 $n \cdot \left(\!\!\binom{n+1}{m}\!\!\right) = (n+m) \cdot \left(\!\!\binom{n}{m}\!\!\right) = (m+1) \cdot \left(\!\!\binom{n}{m+1}\!\!\right)$.

1.8.13 Let $n \geq 1$ and $m \geq 1$.

1 Show that:

$$\sum_{0 \leq j < m} \left(\!\!\binom{n}{j}\!\!\right) = \left(\!\!\binom{m}{n}\!\!\right).$$

2 Deduce that for $n \geq 1$ and $k \geq 0$,

$$\sum_{0 \leq j \leq k} \left(\!\!\binom{n}{j}\!\!\right) = \binom{n + k}{n}.$$

The second equation in Exercise 1.2.4 is a special case.

3 Prove next, for $n \geq 1$ and $m \geq 1$,

$$\sum_{i < n} \left(\!\!\binom{m}{i}\!\!\right) + \sum_{j < m} \left(\!\!\binom{n}{j}\!\!\right) = \binom{n + m}{n}.$$

1.8.14 Use the first item in the previous exercise to prove, for $n \geq 1$, $m \geq 2$ and $k \geq 3$

$$\sum_{0 \leq j < m} \left(\!\!\binom{n}{j}\!\!\right) \cdot j = n \cdot \left(\!\!\binom{m-1}{n+1}\!\!\right)$$
$$\sum_{0 \leq j < k} \left(\!\!\binom{n}{j}\!\!\right) \cdot j^2 = n \cdot (n+1) \cdot \left(\!\!\binom{k-2}{n+2}\!\!\right) + n \cdot \left(\!\!\binom{k-1}{n+1}\!\!\right).$$

1.8.15 (From Jacobs and Széles (2025)) Show that the sizes of supports of natual multisets can also counted, in the following way. Let X have $N > 0$ elements and let $K \in \mathbb{N}$ be the size of the multisets under consideration.

1 (From Jacobs and Széles (2025)) Show first that:

$$\sum_{\varphi \in \mathcal{N}[K](X)} \left|supp(\varphi)\right| = \begin{cases} 0 & \text{if } K = 0 \\ N \cdot \left(\!\!\binom{N}{K-1}\!\!\right) & \text{otherwise.} \end{cases}$$

2 Show that for $0 \leq i \leq K$ and $x \in X$ one has, when $K > 0$,

$$\sum_{\varphi \in \mathcal{N}[K](X),\, \varphi(x) = i} \left|supp(\varphi)\right| = \begin{cases} (N-1) \cdot \left(\!\!\binom{N-1}{K-1}\!\!\right) & \text{if } i = 0 \\ (N-1) \cdot \left(\!\!\binom{N-1}{K-i-1}\!\!\right) + \left(\!\!\binom{N-1}{K-i}\!\!\right) & \text{if } i > 0. \end{cases}$$

1.9 Multiset partitions and the triangular prism

So far we have seen lists, subsets, set partitions and multisets with various maps between them. In this section we add one more collection type, namely multiset partitions over a number K, written as $MP(K)$. The following triangular prism then gives an overview.

$$\begin{tikzcd}
 & & \mathcal{N}[K](X) \arrow[rr, "mc"] \arrow[dl, "mat"] & & MP(K) \arrow[dl, "size"] \\
X^K \arrow[urr, "acc"] \arrow[rr, "mat"] \arrow[dr, "supp"'] & & SP(K) \arrow[urr, "sc"] \arrow[dr, "size"] & & \\
 & \mathcal{P}_{\leq}[K](X) \arrow[ur, "supp"] \arrow[rr, "size"] & & (K] &
\end{tikzcd} \tag{1.52}$$

We have already seen the triangle on the left — but not for a fixed size K — namely in Diagram (1.37). The front rectangle appeared in Diagram (1.24). The parts of the prism involving the set $MP(K)$ of 'multiset partitions' is new and will be discussed below.

This set $MP(K)$ contains special multisets over the positive natural numbers with sum equal to K. They will be called multiset partitions. The operations of multiplicity count mc and size count sc in the above prism produce such multiset partitions, see Jacobs (2022b); Jacobs and Stein (2023a); Jacobs (2022a) for further information.

Definition 1.9.1 For a positive number $K \in \mathbb{N}_{>0}$ we write:

$$MP(K) := \{\alpha \in \mathcal{N}(\mathbb{N}_{>0}) \mid sum(\alpha) = K\},$$

where $sum(\alpha) := \sum_n \alpha(n) \cdot n$.

For $\alpha \in MP(K)$ we write $size(\alpha) = \|\alpha\| = \sum_n \alpha(n)$ for its size, as ordinary multiset. Since $size(\alpha) \leq sum(\alpha)$, this size forms a function $MP(K) \to (K]$, as in the prism (1.52).

It is not hard to see that the support of a multiset partition $\alpha \in MP(K)$ is contained in $(K] = \{1, \ldots, K\}$. For instance, the set $MP(4)$ contains the five multiset partitions:

$$4|1\rangle \qquad 2|1\rangle + 1|2\rangle \qquad 1|1\rangle + 1|3\rangle \qquad 2|2\rangle \qquad 1|4\rangle.$$

These can be understood as the different ways to 'break' the value 4 into coins. In Subsection 1.4.3 we have looked at *lists* of coins with a fixed sum. Clearly, the order matters in such lists, see especially (1.13). Here we use *multisets* of coins, where the order is irrelevant. When asked to break 4, people will reply with, for instance: I can give you two of 1 and one of 2. This answer corresponds to a multiset $2|1\rangle + 1|2\rangle$. People will typically not take the different orders $[1, 1, 2]$, $[1, 2, 1]$ and $[2, 1, 1]$ into account.

The sizes of the sets $MP(K)$ of multiset partitions, for $K = 1, 2, \ldots$, are given by the partition numbers:

$$\begin{array}{llll} \big|MP(1)\big| = 1 & \big|MP(2)\big| = 2 & \big|MP(3)\big| = 3 & \big|MP(4)\big| = 5 \\ \big|MP(5)\big| = 7 & \big|MP(6)\big| = 11 & \big|MP(7)\big| = 15 & \big|MP(8)\big| = 22 \quad \cdots \end{array}$$

No closed formula is known for these partition numbers 1, 2, 3, 5, 7, 11, 15, 22, 30, 42, 56, 77, 101, 135, 176, 231, 297, ... see Andrews (1998).

We turn to the maps labeled *mc* and *sc* in (1.52).

Definition 1.9.2 Fix a number $K \geq 1$.

1 For a set X we define the *multiplicity count* function $mc[K]\colon \mathcal{N}[K](X) \to MP(K)$ on a multiset $\varphi \in \mathcal{N}[K](X)$ as:

$$mc[K]\Big(\textstyle\sum_i n_i|x_i\rangle\Big) := \textstyle\sum_i 1|n_i\rangle \qquad \text{that is} \qquad mc[K](\varphi) := \displaystyle\sum_{x \in supp(\varphi)} 1\Big|\varphi(x)\Big\rangle.$$

In the first formulation it is assumed implicitly that $n_i > 0$ for each i. To be precise and explicit, the multiplicity count function *mc* counts *non-zero* multiplicities.

2 The *size count* function $sc[K]\colon SP(K) \to MP(K)$ is defined on a set partition $P \in SP(K)$ as:

$$sc[K](P) := \sum_{B \in P} 1\Big||B|\Big\rangle.$$

It thus counts the sizes of blocks B in the set partition P.

The parameter K in multiplicity / size count expressions $mc[K]$ and $sc[K]$ is dropped when it is clear from the context.

For instance, for $X = \{a, b, c\}$ and $K = 10$,

$$\begin{aligned} mc\Big(2|a\rangle + 6|b\rangle + 2|c\rangle\Big) &= 1|2\rangle + 1|6\rangle + 1|2\rangle = 2|2\rangle + 1|6\rangle \\ mc\Big(2|a\rangle + 3|b\rangle + 5|c\rangle\Big) &= 1|2\rangle + 1|3\rangle + 1|5\rangle. \end{aligned}$$

In these cases the multiplicities of elements a, b, c are counted, resulting in multiset partitions. By counting the sizes of blocks in set partitions we also get multiset partitions:

$$\begin{aligned} sc\Big(\big\{\{1,2\},\{3,4,5,6,7,8\},\{9,10\}\big\}\Big) &= 1|2\rangle + 1|6\rangle + 1|2\rangle = 2|2\rangle + 1|6\rangle \\ sc\Big(\big\{\{1,10\},\{2,5,9\},\{3,4,6,7,8\}\big\}\Big) &= 1|2\rangle + 1|3\rangle + 1|5\rangle. \end{aligned}$$

We collect some basic facts about size and multiplicity count.

Lemma 1.9.3 *1 The back-rectangle in (1.52) commutes: size ∘ mc = size ∘ supp;*

2 The right-triangle in (1.52) commutes: size ∘ sc = size;

3 When $|X| \geq K$, the function $\mathrm{mc}\colon \mathcal{N}[K](X) \to MP(K)$ is surjective;

Now let $\varphi \in \mathcal{N}[K](X)$ a natural multiset of size $K \geq 1$.

4 Multiplicity count is invariant under substitution: for each isomorphism $\pi\colon X \xrightarrow{\cong} X$,

$$mc\Big(\mathcal{N}(\pi)(\varphi)\Big) = mc(\varphi).$$

5 For an arbitrary element $x \in X$,

$$mc\Big(\varphi + 1|x\rangle\Big) = mc(\varphi) + 1\Big|\varphi(x)+1\,\rangle - 1\Big|\varphi(x)\rangle.$$

6 Similarly, for $x \in supp(\varphi)$,

$$mc\Big(\varphi - 1|x\rangle\Big) = \begin{cases} mc(\varphi) - 1\Big|1\,\rangle & \text{if } \varphi(x) = 1 \\ mc(\varphi) + 1\Big|\varphi(x)-1\,\rangle - 1\Big|\varphi(x)\rangle & \text{otherwise.} \end{cases}$$

7 $\varphi\!\underset{\circ}{!} = \prod_{1\leq n\leq K} (n!)^{mc(\varphi)(n)}$.

Proof. 1 For $\varphi \in \mathcal{N}[K](X)$,

$$(size \circ mc)(\varphi) = size\left(\sum_{x\in supp(\varphi)} 1\Big|\varphi(x_i)\rangle\right) = \sum_{x\in supp(\varphi)} 1 = \Big|supp(\varphi)\Big| = (size \circ supp)(\varphi).$$

2 For a set partition $P \in SP(K)$,

$$(size \circ sc)(P) = size\left(\sum_{B\in P} 1\Big||B|\Big\rangle\right) = \sum_{B\in P} 1 = \Big|P\Big| = size(P).$$

3 Let $\alpha = \sum_{1\leq k\leq K} n_k|k\rangle \in MP(K)$ be given. We can construct a multiset $\varphi_K \in \mathcal{N}[K](X)$ with $mc(\varphi_K) = \alpha$ as follows. We do so by constructing consecutively multisets $\varphi_0, \varphi_1, \ldots, \varphi_K$. We start with the empty multiset $\varphi_0 = \mathbf{0}$.

For each number m with $1 \leq m \leq K$ we set $\varphi_m = \varphi_{m-1} + m|x_1\rangle + \cdots m|x_{n_m}\rangle$, where $x_1, \ldots, x_{n_m}$ are freshly chosen elements from X, not occurring in φ_{m-1}. This guarantees that φ_m has n_m many elements occurring m times. By construction $mc(\varphi_m) = \sum_{1\leq k\leq m} n_k|k\rangle$. In particular, $mc(\varphi_K) = \sum_{1\leq k\leq K} n_k|k\rangle = \alpha$. This construction uses $\|\alpha\|$ many elements from X. Since $K = sum(\alpha) = \sum_m n_m \cdot m \geq \sum_m n_m = \|\alpha\|$, the construction always works for a set X with at least K elements.

4 Obvious, since swapping the elements of a multiset does not change its multiplicities.

5 Let $k = \varphi(x)$. If we add another element x to the multiset φ, then the number $mc(\varphi)(k)$ of elements in φ occuring k times is decreased by one, and the number $mc(\varphi)(k+1)$ of elements occurring $k+1$ times is increased by one.

6 By a similar argument.

7 By induction on $K \geq 1$. The statement clearly holds when $K = 1$. Next, by the using item (5),

$$\begin{aligned}
\prod_{1\leq m\leq K+1} (k!)^{mc(\varphi+1|x\rangle)(m)} &= ((K+1)!)^{mc(\varphi+1|x\rangle)(K+1)} \cdot \prod_{1\leq m\leq K} (k!)^{mc(\varphi+1|x\rangle)(m)} \\
&= \begin{cases} (K+1)! & \text{if } \varphi = K|x\rangle \\ \displaystyle\prod_{1\leq m\leq K} (k!)^{mc(\varphi)(m)} \cdot \frac{(\varphi(x)+1)!}{\varphi(x)!} & \text{otherwise} \end{cases} \\
&\overset{\text{(IH)}}{=} \begin{cases} (\varphi + 1|x\rangle)\mathbin{\underset{\circ}{!}} & \text{if } \varphi = K|x\rangle \\ \varphi\mathbin{\underset{\circ}{!}} \cdot (\varphi(x)+1) & \text{otherwise} \end{cases} \\
&= (\varphi + 1|x\rangle)\mathbin{\underset{\circ}{!}}.
\end{aligned}$$

The last item in the above lemma gives rise to a separate 'open' factorial $\mathbin{\underset{\circ}{!}}$ for multiset partitions.

Definition 1.9.4 Let $\alpha \in MP(K)$ be a multiset partition with sum K.

1 The factorial $\alpha\mathbin{\underset{\circ}{!}}_p \in \mathbb{N}$ is defined as:

$$\alpha\mathbin{\underset{\circ}{!}}_p := \prod_{1\leq n\leq K} (n!)^{\alpha(n)}$$

2 Along the same lines we put:

$$(\alpha)_p := \frac{sum(\alpha)!}{\alpha\mathbin{\underset{\circ}{!}}_p} = \frac{K!}{\alpha\mathbin{\underset{\circ}{!}}_p}.$$

These definitions make sense in the light of the following result.

Lemma 1.9.5 *For $\varphi \in \mathcal{N}[K](X)$,*

$$\varphi\mathbin{\underset{\circ}{!}} = mc(\varphi)\mathbin{\underset{\circ}{!}}_p \qquad \textit{and} \qquad (\varphi) = (mc(\varphi))_p. \tag{1.53}$$

Proof. By Lemma 1.9.3 (7). ❑

Several times in this chapter we have looked at sizes of inverse images. We shall also do so for multiplicity count and size count.

Lemma 1.9.6 *Let X be a finite set with $N = |X|$ elements and let $\alpha \in MP(K)$ be a multiset partition with $\|\alpha\| \leq N$. Then:*

$$\left|mc[K]^{-1}(\alpha)\right| = \frac{N!}{\alpha\mathbin{\underset{\circ}{!}} \cdot (N-\|\alpha\|)!} \overset{(1.47)}{=} \binom{N}{\alpha}.$$

Proof. Let $\alpha = \sum_{1\leq i\leq K} n_i|i\rangle \in MP(K)$. If $\varphi \in \mathcal{N}[K](X)$ satisfies $mc(\varphi) = \alpha$, then φ must have n_i

elements of X occurring i times, for each i. The number of choices for these elements is:

$$\begin{aligned}
&\binom{N}{n_1}\cdot\binom{N-n_1}{n_2}\cdot\ldots\cdot\binom{N-n_1-\cdots-n_{K-1}}{n_K}\\
&= \frac{N!\cdot(N-n_1)!\cdot\ldots\cdot(N-n_1-\cdots-n_{K-1})!}{n_1!\cdot(N-n_1)!\cdot n_2!\cdot(N-n_1-n_2)!\cdot\ldots\cdot n_K!\cdot(N-n_1-\cdots-n_K)!}\\
&= \frac{N!}{n_1!\cdot n_2!\cdot\ldots\cdot n_K!\cdot(N-n_1-\cdots-n_K)!}\\
&= \frac{N!}{\alpha\llparenthesis\cdot(N-\|\alpha\|)!}.
\end{aligned}$$

❑

We turn to size count, from set partitions to multiset partitions.

Lemma 1.9.7 *For $\alpha\in MP(K)$,*

$$\Big|sc[K]^{-1}(\alpha)\Big| = \frac{(\alpha)_p}{\alpha\llparenthesis} = \frac{K!}{\alpha\llparenthesis_p\cdot\alpha\llparenthesis}.$$

Proof. Let $\alpha=\sum_i n_i|i\rangle$ with $\sum_i n_i\cdot i=K$. We need to find the number of set partitions of the set $(K)=\{1,\ldots,K\}$ whose size count equals α. We switch to lists ℓ of length K over $(K]$ with $supp(\ell)=(K]$. There are $K!$ many such lists. Given such a list ℓ there are $\prod_i n_i! = \alpha\llparenthesis$ many ways to break up the list into sublists, with n_i-many subslists of length i. This gives $\frac{K!}{\alpha\llparenthesis}$ many set partitions. We still have to divide by $\alpha\llparenthesis_p$ because within each block of size i we have to take the $i!$ many permutations of the elements into account. ❑

We include one more combinatorial result.

Lemma 1.9.8 *Let X be a finite set with $|X|\geq K\geq 1$. Consider a multiset $\varphi\in\mathcal{N}[K](X)$ and a set partition $P\in SP(K)$ with,*

$$\begin{aligned}
mc(\varphi) &= sc(P)\\
&= \alpha\in MP(K),\ \textit{say.}
\end{aligned}$$

Then there are $\alpha\llparenthesis$ many sequences $\vec{x}\in X^K$ with $acc(\vec{x})=\varphi$ and $mat(\vec{x})=P$, that is:

$$\Big|\{\vec{x}\in X^K \mid acc(\vec{x})=\varphi \textit{ and } mat(\vec{x})=P\}\Big| = \alpha\llparenthesis.$$

Proof. Let $\vec{x}\in X^K$ be a candidate sequence with $acc(\vec{x})=\varphi$ and $mat(\vec{x})=P$. The multiset φ tells us that an element $y\in supp(\varphi)$ must occur $\varphi(y)$ many times in $\vec{x}$. The set partition P tells where in $\vec{x}$ elements must be equal. Suppose that there is only one element $y\in supp(\varphi)$ with $\varphi(y)=3$, so occurring three times in $\vec{x}$; then we know precisely where to put y in the candidate sequence $\vec{x}$, since there are precisely three equal elements in $\vec{x}$. More generally, if $\alpha(i)=n$, where $\alpha=mc(\varphi)=sc(P)$, then there are n elements occurring i-many times in φ. There are then $n!$ many ways to place these n elements in $\vec{x}$, namely at the n subsequences of $\vec{x}$ with i-many equal elements. This works for each i, so that in total there are $\prod_i\alpha(i)! = \alpha\llparenthesis$ many sequences $\vec{x}$ with $acc(\vec{x})=\varphi$ and $mat(\vec{x})=P$. ❑

We conclude with a multiset analogue of Theorem 1.4.8. The proof that we use allows us to prove also the as yet unproven part of this earlier theorem.

Theorem 1.9.9 *For each $K \in \mathbb{N}_{>0}$,*

$$\sum_{\alpha \in MP(K)} \frac{1}{\alpha\llparenthesis \cdot prod(\alpha)} = 1 \quad \textit{where} \quad prod(\alpha) := \sum_{n \in supp(\alpha)} n^{\alpha(n)}. \tag{1.54}$$

Proof. Since:

$$\begin{aligned}
\sum_{\alpha \in MP(K)} \frac{1}{\alpha\llparenthesis \cdot prod(\alpha)} &= \sum_{1 \le n \le K} \sum_{\alpha \in MP(K),\, \|\alpha\| = n} \frac{1}{\alpha\llparenthesis \cdot prod(\alpha)} \\
&= \frac{1}{K!} \cdot \sum_{1 \le n \le K} \sum_{\alpha \in MP(K),\, \|\alpha\| = n} \frac{K!}{\alpha\llparenthesis \cdot prod(\alpha)} \\
&= \frac{1}{K!} \cdot \sum_{1 \le n \le K} \begin{bmatrix} K \\ n \end{bmatrix} && \text{by Exercise 1.9.5} \\
&= \frac{1}{K!} \cdot K! && \text{by Lemma 1.2.4 (5)} \\
&= 1.
\end{aligned}$$

❑

In a similar way we can now resolve an outstanding issue.

Proof. (of Theorem 1.4.8 (1)) For $1 \le n \le K$, one has:

$$\begin{aligned}
\sum_{\ell \in sum^{-1}(K),\, \|\ell\| = n} \frac{1}{prod(\ell)} &= \sum_{\alpha \in MP(K),\, \|\alpha\| = n} \sum_{\ell \in acc^{-1}(\alpha)} \frac{1}{prod(\ell)} \\
&= \sum_{\alpha \in MP(K),\, \|\alpha\| = n} \sum_{\ell \in acc^{-1}(\alpha)} \frac{1}{prod(\alpha)} \\
&= \sum_{\alpha \in MP(K),\, \|\alpha\| = n} \frac{\mathbf{(}\alpha\mathbf{)}}{prod(\alpha)} && \text{by Proposition 1.7.2} \\
&= \frac{n!}{K!} \cdot \sum_{\alpha \in MP(K),\, \|\alpha\| = n} \frac{K!}{\alpha\llparenthesis \cdot prod(\alpha)} \\
&= \frac{n!}{K!} \cdot \begin{bmatrix} K \\ n \end{bmatrix} && \text{by Exercise 1.9.5.}
\end{aligned}$$

❑

Exercises

1.9.1 1 Double-check that all subdiagrams of Diagram (1.52) commute.

2 Check in detail that starting from the list $[b, a, b, c]$ gives the following outcomes.

$$\begin{tikzcd}
 & 1|a\rangle + 2|b\rangle + 1|c\rangle \arrow[rr, mapsto, "mc"] & & 2|1\rangle + 1|2\rangle \arrow[dd, mapsto, "size"] \\
{[b,a,b,c]} \arrow[ur, mapsto, "acc"] \arrow[rr, mapsto, "mat"] \arrow[dr, mapsto, "supp"'] & & \{\{1,3\},\{2\},\{4\}\} \arrow[ur, mapsto, "sc"] \arrow[dr, mapsto, "size"] & \\
 & \{a,b,c\} \arrow[ur, mapsto, "supp"] \arrow[rr, mapsto, "size"] & & 3
\end{tikzcd}$$

1.9.2 Fix $K \ge 1$.

1 Let $\ell \in X^K$ be a sequence. Show that

$$supp(\ell) = supp(acc(\ell)) \cong mat(\ell).$$

2 Let $P \in SP(K)$ be a set partition. Show that there is a sequence ℓ with $mat(\ell) = P = supp(\ell)$.

1.9.3 Let $\alpha \in MP(K)$ and $\beta \in MP(L)$ be multiset partitions.

1 Write $\alpha + \beta \in \mathcal{N}(\mathbb{N}_{>0})$ for the pointwise sum of multisets, so that $(\alpha + \beta)(i) = \alpha(i) + \beta(i)$. Check that this $\alpha + \beta$ is an element of $MP(K + L)$.

2 Calculate: $\big(2|1\rangle + 1|2\rangle + 2|3\rangle\big) + \big(2|2\rangle + 3|4\rangle\big)$.

3 Write $\alpha \star \beta \in \mathcal{N}(\mathbb{N}_{>0})$ for $\sum_{i,j} \alpha(i) \cdot \beta(j) \big| i \cdot j \rangle$ for the multiplicative convolution of α, β. Show that $\alpha \star \beta$ is in $MP(K \cdot L)$.

4 Calculate: $\big(2|1\rangle + 1|2\rangle + 2|3\rangle\big) \star \big(2|2\rangle + 3|4\rangle\big)$.

5 Check that $+$ and $\star$ on multiset partitions are commutative.

6 Prove that multiplication $\star$ distributes over sum $+$ in:

$$(\alpha_1 + \alpha_2) \star \beta = (\alpha_1 \star \beta) + (\alpha_2 \star \beta).$$

1.9.4 Let natural multisets $\varphi, \psi \in \mathcal{N}(X)$ have disjoint supports: $supp(\varphi) \cap supp(\psi) = \emptyset$. Show that then:

$$mc(\varphi + \psi) = mc(\varphi) + mc(\psi).$$

Give an example that demonstrates that the disjointness requirement is necessary.

1.9.5 (See Jacobs (2022a)) In Exercises 1.2.8 and 1.5.8 we have seen how Stirling numbers of the first kind arise via sums over subsets and via sums over sub partitions. This also works with multiset partitions:

$$\begin{bmatrix} K \\ n \end{bmatrix} = \sum_{\alpha \in MP(K),\, \|\alpha\| = n} \frac{K!}{\alpha! \cdot prod(\alpha)}.$$

Prove this equation.

1.9.6 Check that the binomial coefficient (1.47) satisfies the following formula.

$$\sum_{\alpha \in MP(K),\, \|\alpha\| \leq n} (\alpha)_p \cdot \binom{n}{\alpha} = n^K \qquad \text{for all } K, n \in \mathbb{N}_{>0}.$$

Hint: Combine Exercise 1.7.9 with Lemmas 1.9.5 and 1.9.7.

1.10 Combinatorics

Briefly, combinatorics is the area of mathematics that concentrates on counting and arranging. It forms a basis for many results in (discrete) probability theory, where the likelihood of certain events depends on the number of possible options. In the previous sections we have reviewed many elementary facts about lists, subsets, and multisets. In this very short section we relate these facts to some of the classical problems in combinatorics.

1.10.1 Sorting

Sorting involves imposing an order on objects. For instance, one may ask how many words can we form out of the first four letters a, b, c, d of the alphabet? The answer is of course $4! = 24$. It is the number of permutations of these four letters.

Things get a bit more interesting if we add the condition that the letters a, b, c, d may be used respectively $2, 3, 4, 5$ times. The number of options is now:

$$\frac{(2+3+4+5)!}{2! \cdot 3! \cdot 4! \cdot 5!} = \frac{14!}{2! \cdot 3! \cdot 4! \cdot 5!} = 2.522.520.$$

Such sorting problems have occurred in Section 1.7 in terms of accumulation: how many lists of length 14 over the set $\{a, b, c, d\}$ accumulate to the multiset $\varphi = 2|a\rangle + 3|b\rangle + 4|c\rangle + 5|d\rangle$? Proposition 1.7.2 provides the answer in terms of the multiset coefficient:

$$\mathbf{(}\varphi\mathbf{)} = \frac{\|\varphi\|!}{\prod_x \varphi(x)!} = 2.522.520.$$

In general, the number of sortings of N elements that may occur respectively $m_1, \ldots, m_N$ many times is given by the multiset coefficient $\mathbf{(}\varphi\mathbf{)}$ of the (natural) multiset $\varphi = m_1|1\rangle + \cdots + m_N|N\rangle$. In the special case where each element occurs precisely once, that is, when $m_1 = \cdots = m_N = 1$, the outcome is $\frac{N!}{1! \cdot \ldots \cdot 1!} = N!$, as in the original question.

1.10.2 Allocation / distribution

A classical question is: in how many ways can we distribute K balls over $n \geq 1$ buckets? This amounts to: how many (natural) multisets exist of size K over a set X with n elements? The answer is given in Proposition 1.8.7 in terms of the multichoose number:

$$\left(\!\!\binom{n}{K}\!\!\right) = \frac{(n+K-1)!}{(n-1)! \cdot K!}.$$

As an aside: in the above formulation of the question it is relevant that all balls are the same and indistinguishable. If we have K different balls, say numbered $1, \ldots, K$, and we wish to distribute them over n buckets, then there are n^K allocations, since for each ball we can choose each of the n buckets.

Further restrictions may be imposed: suppose we have four buckets, labeled a, b, c, d and we wish to distribute 8 (indistinguishable) balls over them such that two buckets have three balls and one bucket has two. In how many ways can this be done? We list the options below — and count 12 possible allocations.

$3\|a\rangle + 3\|b\rangle + 2\|c\rangle$	$3\|a\rangle + 2\|b\rangle + 3\|c\rangle$	$2\|a\rangle + 3\|b\rangle + 3\|c\rangle$
$3\|a\rangle + 3\|b\rangle + 2\|d\rangle$	$3\|a\rangle + 3\|c\rangle + 2\|d\rangle$	$3\|b\rangle + 3\|c\rangle + 2\|d\rangle$
$3\|a\rangle + 2\|b\rangle + 3\|d\rangle$	$2\|a\rangle + 3\|b\rangle + 3\|d\rangle$	$3\|a\rangle + 2\|c\rangle + 3\|d\rangle$
$3\|b\rangle + 2\|c\rangle + 3\|d\rangle$	$2\|a\rangle + 3\|c\rangle + 3\|d\rangle$	$2\|b\rangle + 3\|c\rangle + 3\|d\rangle$

How does this number 12 arise? The requirement "two buckets have three balls and one bucket has two" corresponds to a multiset partition $\alpha = 1|2\rangle + 2|3\rangle$ with sum 8, so that $\alpha \in MP(8)$. A multiset over $\{a, b, c, d\}$ matches this requirement if its multiplicity count mc is equal to α.

Lemma 1.9.6 gives a formula for the number of multisets in the inverse image, with $N = 4$ buckets:

$$\left|mc^{-1}(\alpha)\right| = \frac{N!}{\alpha! \cdot (N - \|\alpha\|)!} = \frac{4!}{1! \cdot 2! \cdot 1!} = 12.$$

In general, when we wish to distribute K balls over N buckets, where n_i buckets get m_i balls, we use the multiset partition $\alpha = \sum_i n_i|m_i\rangle$. The allocation is possible if $K = sum(\alpha) = \sum_i n_i \cdot m_i$ and if $N \geq \|\alpha\| = \sum_i n_i$. In that case the number of options is $\binom{N}{\alpha} = \frac{N!}{\alpha! \cdot (N-\|\alpha\|)!}$, see Lemma 1.9.6.

1.10.3 Choosing / partitioning

A basic choice question is: given 10 people, in how many ways can we choose 3 of them. The answer is given by the binomial coefficient $\binom{10}{3} = \frac{10!}{7! \cdot 3!} = 120$.

A more abstract way to formulate this problem is: given a 10-element set, how many ways are there to partition this set into two blocks, where one block has size 3, so that the other block has size 7. These two sizes form a multiset partition $\alpha = 1|3\rangle + 1|7\rangle$. It has sum 10, so that $\alpha \in MP(10)$. The question thus is: how many set partitions in $SP(10)$ have α as size count? The answer is given by Lemma 1.9.7 and takes the form:

$$\left|sc^{-1}(\alpha)\right| = \frac{(\alpha)_p}{\alpha!} = \frac{10!}{\alpha!_p \cdot \alpha!} = \frac{10!}{(3!)^1 \cdot (7!)^1 \cdot 1! \cdot 1!} = 120.$$

The general choice question is of the form: given K many elements, the number of ways to partition them into subsets / blocks of size m_i, where blocks of size m_i may occur n_i many times, is determined by the multiset partition $\alpha = \sum_i n_i|m_i\rangle$ with sum K. The number of such set partitions is then described in Lemma 1.9.7 as $\frac{(\alpha)_p}{\alpha!}$.

In the special case where each subset / block size is allowed to occur only once, all multiplicities of the multiset partition α are equal to 1, that is, α is of the form $\sum_i 1|m_i\rangle$. The general expression $\frac{(\alpha)_p}{\alpha!}$ then equals $\frac{K!}{\prod_i m_i!}$, where $K = \sum_i m_i$. This expression is not the same as the multiset coefficent $\binom{\alpha}{}$.

For instance, the number of ways to choose two subsets of size 3 and one subset of size 2 from an 8-element set is:

$$\left|sc^{-1}(1|2\rangle + 2|3\rangle)\right| = \frac{8!}{(2!)^1 \cdot (3!)^2 \cdot 1! \cdot 2!} = 280.$$

1.10.4 Nomials

We have seen binomial coefficients $\binom{N}{i}$. There are also trinomial, quadrinomials, *etc.*. These generalisations are described in a systematic manner in Jacobs (2025b), in the context of statistical mechanics. The description uses sequences and multisets and fits well in what we have seen so far. For later usage we collect the basics.

We recall that the binomial coefficient $\binom{N}{i} \in \mathbb{N}$ gives the number of subsets of size i, of a (finite) set of size N. Such a subset of size i, say of a set $\{x_1, \ldots, x_N\}$ of N elements, can be identified with sequence of binary numbers $(b_1, \ldots, b_N) \in \{0, 1\}^N$ of length N with numbers $b_i \in \{0, 1\}$ satisfying $i = sum(b_1, \ldots, b_N) = b_1 + \cdots + b_N$. Each number b_i then tells if the element x_i is in the subset (when $b_i = 1$) or not (when $b_i = 0$). This description is generalised below.

Definition 1.10.1 Fix a natural number $N \geq 1$ and write, as usual, $\boldsymbol{N} = \{0, 1, \ldots, N-1\} \subseteq \mathbb{N}$. For numbers $K \in \mathbb{N}$ and $0 \leq i \leq (N-1) \cdot K$ we define the *N-nomial coefficient* $C_N(K, i) \in \mathbb{N}$ as the number of sequences that sum to i.

$$\begin{aligned} C_N(K, i) &:= \Big|\{\vec{n} \in \boldsymbol{N}^K \mid sum(\vec{n}) = i\}\Big| \\ &= \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{N}),\, sum(\varphi)=i} (\varphi). \end{aligned} \tag{1.55}$$

The last equation depends on the fact that for $\vec{n} \in acc^{-1}(\varphi)$ one has $sum(\vec{n}) = i$ iff $sum(\varphi) = \sum_j \varphi(j) \cdot j = i$, and of course on Proposition 1.7.2.

As argued before this definition, $C_2(K, i) = \binom{K}{i}$. For $N = 3$, one gets the following numbers, presented as a pyramid:

K	$C_3(K, -)$								
0					1				
1				1	1	1			
2			1	2	3	2	1		
3		1	3	6	7	6	3	1	
4	1	4	10	16	19	16	10	4	1
$\vdots$									

The last line says, for instance $C_3(4, 0) = 1$, $C_3(4, 1) = 4$, $C_3(4, 2) = 10$, *etc.*

Lemma 1.10.2 *Let numbers $N \geq 1$, $K \geq 0$ be given.*

1 N-nomials add up to N^K in the following way.

$$\sum_{0 \leq i \leq (N-1) \cdot K} C_N(K, i) = N^K.$$

This generalises the familiar equation $\sum_{0 \leq i \leq K} \binom{K}{i} = 2^K$ in the binary case.

2 N-nomials satisfy a Vandermonde property: for each $0 \leq i \leq (N-1) \cdot K$, if $K = K_1 + K_2$, then:

$$C_N(K, i) = \sum_{0 \leq i_1 \leq (N-1) \cdot K_1,\, 0 \leq i_2 \leq (N-1) \cdot K_2,\, i_1 + i_2 = i} C_N(K_1, i_1) \cdot C_N(K_2, i_2).$$

The Vandermonde property (1.48) for binomial coefficients is a special case, for $N = 2$.

3 N-nomials are closed under reversal:

$$C_N(K, i) = C_N\Big(K,\ (N-1) \cdot K - i\Big).$$

This generalises $\binom{K}{i} = \binom{K}{K-i}$ for $N = 2$.

Proof. 1 Since:

$$\begin{aligned} \sum_{0 \leq i \leq (N-1) \cdot K} C_N(K, i) &= \sum_{0 \leq i \leq (N-1) \cdot K} \Big|\{\vec{n} \in \boldsymbol{N}^K \mid sum(\vec{n}) = i\}\Big| \\ &= \Big|\{\vec{n} \in \boldsymbol{N}^K \mid 0 \leq sum(\vec{n}) \leq (N-1) \cdot K\}\Big| = \Big|\boldsymbol{N}^K\Big| = N^K. \end{aligned}$$

2 We use that sequences of length $K = K_1 + K_2$ can be written as concatenations $+\!\!+$ of sequences of length K_1 and K_2.

$$\begin{aligned}
C_N(K,i) &\overset{(1.55)}{=} \left|\{\vec{n} \in \boldsymbol{N}^K \mid sum(\vec{n}) = i\}\right| \\
&= \sum_{0 \le i_1 \le (N-1)\cdot K_1,\, 0 \le i_2 \le (N-1)\cdot K_2,\, i_1+i_2=i} \left|\{\vec{n_1} +\!\!+ \vec{n_2} \mid \vec{n_1} \in \boldsymbol{N}^{K_1}, \vec{n_2} \in \boldsymbol{N}^{K_2}, \right. \\
&\qquad\qquad \left. sum(\vec{n_1}) = i_1, sum(\vec{n_2}) = i_2\}\right| \\
&= \sum_{0 \le i_1 \le (N-1)\cdot K_1,\, 0 \le i_2 \le (N-1)\cdot K_2,\, i_1+i_2=i} \left|\{\vec{n_1} \in \boldsymbol{N}^{K_1} \mid sum(\vec{n_1}) = i_1\}\right| \\
&\qquad\qquad \cdot \left|\{\vec{n_2} \in \boldsymbol{N}^{K_2} \mid sum(\vec{n_2}) = i_2\}\right| \\
&= \sum_{0 \le i_1 \le (N-1)\cdot K_1,\, 0 \le i_2 \le (N-1)\cdot K_2,\, i_1+i_2=i} C_N(K_1, i_1) \cdot C_N(K_2, i_2).
\end{aligned}$$

3 Consider the reversal self-isomorphism $rev_L \colon \{0, \ldots, L\} \xrightarrow{\cong} \{0, \ldots, L\}$ by $rev(i) = L - i$. The following properties then hold.

(a) The multiset coefficient of a reversed multiset is the same as the coefficient of the original: $\boldsymbol{(}\mathcal{N}(rev_{N-1})(\varphi)\boldsymbol{)} = \boldsymbol{(}\varphi\boldsymbol{)}$.

(b) There is a commuting diagram:

$$\begin{tikzcd}
\mathcal{N}[K](\boldsymbol{N}) \arrow[r, "sum"] \arrow[d, "\mathcal{N}(rev_{N-1})"', "\cong"] & \{0, 1, \ldots, (N-1)\cdot K\} \arrow[d, "\cong"', "rev_{(N-1)\cdot K}"] \\
\mathcal{N}[K](\boldsymbol{N}) \arrow[r, "sum"] & \{0, 1, \ldots, (N-1)\cdot K\}
\end{tikzcd} \tag{1.56}$$

In particular, this gives: $sum(\varphi) = i \iff sum(\mathcal{N}(rev_{N-1})(\varphi)) = (N-1)\cdot K - i$.

Now we can prove the closure of N-nomials under reversal:

$$\begin{aligned}
C_N(K,i) &\overset{(??)}{=} \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{N}),\, sum(\varphi)=i} \boldsymbol{(}\varphi\boldsymbol{)} \\
&\overset{(a)}{=} \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{N}),\, sum(\mathcal{N}(rev_{N-1})(\varphi))=i} \boldsymbol{(}\mathcal{N}(rev_{N-1})(\varphi)\boldsymbol{)} \\
&\overset{(b)}{=} \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{N}),\, sum(\varphi)=(N-1)\cdot K-i} \boldsymbol{(}\varphi\boldsymbol{)} \overset{(??)}{=} C_N\Big(K,\, (N-1)\cdot K - i\Big).
\end{aligned}$$

❑

There is a description / definition of N-nomials in terms of polynomial expressions, occurring on the Online Encyclopedia of Integer Sequences (OEIS)[2]. We show how this matches the approach used here, using the Multinomial Theorem.

Theorem 1.10.3 *For $N \ge 1$ and $K \ge 0$ one has, for an arbitrary variable x,*

$$\left(\sum_{0 \le j < N} x^j\right)^K = \sum_{0 \le i \le (N-1)\cdot K} C_N(K,i) \cdot x^i.$$

[2] See oeis.org/wiki/Multinomial_coefficients, accessed on 2025-06-25.

Proof. We use the Multinomial Theorem (1.43) in terms of multisets in the first step in:

$$\begin{aligned}\left(\sum_{0\le j<N} x^j\right)^K &= \sum_{\varphi\in\mathcal{N}[K](N)} \binom{\varphi}{} \cdot \prod_{0\le j<N} (x^j)^{\varphi(j)} \\ &= \sum_{\varphi\in\mathcal{N}[K](N)} \binom{\varphi}{} \cdot x^{\sum_j \varphi(j)\cdot j} \\ &= \sum_{\varphi\in\mathcal{N}[K](N)} \binom{\varphi}{} \cdot x^{sum(\varphi)} \\ &= \sum_{0\le i\le (N-1)\cdot K} \sum_{\varphi\in\mathcal{N}[K](N),\, sum(\varphi)=i} \binom{\varphi}{} \cdot x^i \overset{(??)}{=} \sum_{0\le i\le (N-1)\cdot K} C_N(K,i)\cdot x^i. \end{aligned}$$ ❑

The next result is rather technical in nature. It is a preparation for the mean and variance of a distribution associated with N-nomials, see Proposition 5.1.6.

Lemma 1.10.4 *Consider numbers $N \ge 1$ and $K \ge 0$.*

1 One has:

$$\sum_{0\le i\le (N-1)\cdot K} C_N(K,i)\cdot i = \frac{(N-1)\cdot K}{2}\cdot N^K.$$

2 And:

$$\sum_{0\le i\le (N-1)\cdot K} C_N(K,i)\cdot i^2 = \left(\frac{(N-1)\cdot K}{2} + \frac{N+1}{6}\right)\cdot \frac{(N-1)\cdot K}{2}\cdot N^K.$$

Proof. 1 We abbreviate $M := (N-1)\cdot K$ and $c_i := C_N(K,i)$. The reversal $C_N(K,i) = C_N(K,M-i)$ from Lemma 1.10.2 (3) gives $c_i = c_{M-i}$. Then:

$$\begin{aligned}\sum_{0\le i\le M} c_i\cdot i &= \tfrac{1}{2}\cdot \sum_{0\le i\le M} c_i\cdot i + c_{M-i}\cdot (M-i) = \tfrac{1}{2}\cdot \sum_{0\le i\le M} c_i\cdot i + c_i\cdot (M-i) \\ &= \tfrac{1}{2}\cdot \sum_{0\le i\le M} c_i\cdot M = \frac{M}{2}\cdot N^K.\end{aligned}$$

The last equation follows from Lemma 1.10.2 (1).

2 This is more work. We first rewrite:

$$\begin{aligned}C_N(K,i)\cdot i &= \sum_{\varphi\in\mathcal{N}[K](N),\, sum(\varphi)=i} \binom{\varphi}{} \cdot sum(\varphi) = \sum_{\varphi\in\mathcal{N}[K](N),\, sum(\varphi)=i} \sum_{0\le j<N} \binom{\varphi}{} \cdot \varphi(j)\cdot j \\ &= \sum_{0\le j<N} j\cdot K\cdot \sum_{\varphi\in\mathcal{N}[K-1](N),\, sum(\varphi)=i-j} \binom{\varphi}{} \qquad (*) \\ &= \sum_{0\le j<N} j\cdot K\cdot C_N(K-1,i-j).\end{aligned}$$

Now we get, via the previous point and Proposition 1.2.6,

$$\begin{aligned}
&\sum_{0\le i\le (N-1)\cdot K} C_N(K,i)\cdot i^2 \\
&\overset{(*)}{=} \sum_{0\le i\le (N-1)\cdot K}\ \sum_{0\le j<N} j\cdot K\cdot C_N(K-1,i-j)\cdot i \\
&= K\cdot \sum_{0\le i\le (N-1)\cdot (K-1)}\ \sum_{0\le j<N} C_N(K-1,i)\cdot (i+j)\cdot j \\
&= K\cdot \sum_{0\le i\le (N-1)\cdot (K-1)} \left(C_N(K-1,i)\cdot i\cdot \left(\sum_{0\le j<N} j\right) + C_N(K-1,i)\cdot \left(\sum_{0\le j<N} j^2\right)\right) \\
&= K\cdot \left(\frac{(N-1)\cdot (K-1)}{2}\cdot N^{K-1}\cdot \frac{N\cdot (N-1)}{2} + N^{K-1}\cdot \frac{(N-1)\cdot N\cdot (2N-1)}{6}\right) \\
&= \left(\frac{(N-1)\cdot (K-1)}{2} + \frac{2N-1}{6}\right)\cdot \frac{(N-1)\cdot K}{2}\cdot N^K \\
&= \left(\frac{(N-1)\cdot K}{2} + \frac{N+1}{6}\right)\cdot \frac{(N-1)\cdot K}{2}\cdot N^K.
\end{aligned}$$

❑

In Lemma 1.8.2 and Proposition 1.8.6 we have seen generalisations of the Vandermonde summations to multisets. A similar generalisation exists for N-nomials.

Proposition 1.10.5 *Let a number $N \ge 1$ and a multiset $\psi \in \mathcal{N}[K](X)$ of size K be given. For a number $0 \le i \le (N-1)\cdot K$ and a multiset $\varphi \in \mathcal{N}[i](X)$ with $\varphi \le (N-1)\cdot \psi$, we define a product of N-nomials:*

$$C_N(\psi,\varphi) \coloneqq \prod_{x\in X} C_N(\psi(x),\varphi(x)). \tag{1.57}$$

Then:

$$\sum_{\varphi\le_i (N-1)\cdot\psi} C_N(\psi,\varphi) = C_N(K,i).$$

Proof. We use induction on the size of the support of ψ. If this size is one, there is nothing to prove, since ψ and φ can be identified with the numbers K and i.

Now let $supp(\psi) = \{x_1,\dots,x_n,y\} = X\cup\{y\}$ for $X = \{x_1,\dots,x_n\}$ be of size $n+1$. Write $\psi(y) = L \le K$ and $\psi' = \psi - 1|L\rangle \in \mathcal{N}[L-K](X)$. Then by Lemma 1.10.2 (2):

$$\begin{aligned}
\sum_{\varphi\le_i (N-1)\cdot\psi} C_N(\psi,\varphi) &= \sum_{0\le j\le (N-1)\cdot L,\, 0\le k\le (N-1)\cdot (K-L),\, j+k=i}\ \sum_{\varphi\le_k \psi'} C_N(\varphi,\psi')\cdot C_N(L,j) \\
&\overset{\text{(IH)}}{=} \sum_{0\le j\le (N-1)\cdot L,\, 0\le k\le (N-1)\cdot (K-L),\, j+k=i} C_N(K-L,k)\cdot C_N(L,j) \\
&= C_N(K,i).
\end{aligned}$$

❑

So far we used a constraint on the number i of the form $0 \le i \le (N-1)\cdot K$. It turns out that the constructions that we used for N-nomials can be simplified when we assume $i < N$. This is a crucial observation, elaborated in the theorem below. The assumption $i < N$ makes sense, when N is seen as a number that can become arbitrarily big. This is relevant in particle physics, see Jacobs (2025b).

Proposition 1.10.6 *Consider numbers $N, K \ge 1$.*

1 For $0 \le n \le N$,

$$\sum_{0 \le i < n} C_N(K, i) = \frac{n}{K} \cdot \left(\!\!\binom{K}{n}\!\!\right).$$

The right-hand-side does not depend on N, intuitively because if $\mathrm{sum}(\vec{m})$ *equals i, for* $\vec{m} \in \boldsymbol{N}^K$ *where* $i \le n < N$, *then the number N is big enough to so that the requirement* $m_j \in \boldsymbol{N}$ *does not impose any restriction.*

2 When $i < N$, *there is a simple formula for the N-nomial, namely:*

$$C_N(K, i) = \left(\!\!\binom{K}{i}\!\!\right).$$

Proof. 1 By induction on $K \ge 1$. First,

$$\sum_{0 \le i \le n} C_N(1, i) \overset{(1.55)}{=} \sum_{0 \le i \le n} \left|\{\vec{n} \in \boldsymbol{N}^1 \mid sum(\vec{n}) = i\}\right| = \sum_{0 \le i \le n} 1 = n = \frac{n}{1} \cdot \left(\!\!\binom{1}{n}\!\!\right).$$

Next we use that we sum over $i < n \le N$ in the first line below. When we count sequences of length $K+1$ with sum $i < n$, then all elements of the sequence fit in the set $\boldsymbol{N} = \{0, \ldots, N-1\}$; hence we can express the number $C_N(K+1, i)$ of sequences of length $K+1$ as sum of numbers of sequences of length K.

$$\begin{aligned}
\sum_{0 \le i < n} C_N(K+1, i) &= \sum_{0 \le i < n} \sum_{0 \le j \le i} C_N(K, i-j) \\
&= \sum_{0 \le i < n} \sum_{0 \le j \le i} C_N(K, j) \\
&\overset{(\mathrm{IH})}{=} \sum_{0 \le i < n} \frac{i+1}{K} \cdot \left(\!\!\binom{K}{i+1}\!\!\right) \\
&= \sum_{0 \le i < n} \frac{i+1}{K} \cdot \frac{(K+i)!}{(K-1)! \cdot (i+1)!} \\
&= \sum_{1 \le i < n} \left(\!\!\binom{K+1}{i}\!\!\right) \\
&= \left(\!\!\binom{n}{K+1}\!\!\right) \qquad \text{by Exercise 1.8.13 (1)} \\
&= \frac{(K+n)!}{(n-1)! \cdot (K+1)!} = \frac{n}{K+1} \cdot \left(\!\!\binom{K+1}{n}\!\!\right).
\end{aligned}$$

2 By the previous item, for $i < N$, so that $i+1 \le N$,

$$\begin{aligned}
C_N(K, i) &= \left(\sum_{0 \le j < i+1} C_N(K, j)\right) - \left(\sum_{0 \le j < i} C_N(K, j)\right) \\
&= \frac{i+1}{K} \cdot \left(\!\!\binom{K}{i+1}\!\!\right) - \frac{i}{K} \cdot \left(\!\!\binom{K}{i}\!\!\right) \\
&= \frac{(K+i)!}{K! \cdot i!} - \frac{(K+i-1)!}{K! \cdot (i-1)!} \\
&= \frac{(K+i-1)!}{K! \cdot (i-1)!} \cdot \left(\frac{K+i}{i} - 1\right) = \frac{(K+i-1)!}{(K-1)! \cdot i!} = \left(\!\!\binom{K}{i}\!\!\right).
\end{aligned}$$

❑

Exercises

1.10.1 Consider the general formulation at the end of Subsection 1.10.2, for the special case where there is precisely one bucket for each size m_i, so that the relevent multiset partition is $\alpha = \sum_i 1|m_i\rangle$. Check that the number of options, as given by Lemma 1.9.6, reduces to a falling factorial.

1.10.2 Describe all set partitions of a four element set, with two blocks of size 1 and one block of size 2. Calculate the number of such partitions via Lemma 1.9.7.

1.10.3 Verify yourself points (a) and (b) in the proof of Lemma 1.10.2 (3).

1.10.4 Let X be a finite set with $N := |X| > 0$ elements, with a size number $K \geq 1$ and a bound $b \geq 1$. Check that the number of multiset $\varphi \in \mathcal{N}[K](X)$ with $\varphi \leq b \cdot \mathbf{1}$, that is, with $\varphi(x) \leq b$ for each $x \in X$, is equal to the nomial number $C_{b+1}(N, K)$.

1.11 Channels

The previous sections covered the collection types of lists, subsets, and multisets, with much emphasis on the similarities between them. In this section we will exploit these similarities in order to introduce the concept of channel, in a uniform approach, for all of these collection types at the same time. This will illustrate how these data types are not only used for certain types of collections, but also for certain types of computation. Much of the rest of this book builds on the concept of a channel, especially for probabilistic computations, which are introduced in the next chapter. The same general approach to channels that will be described in this section will work for probability distributions.

Let T be one of the collection functors $\mathcal{L}, \mathcal{P}$, or $\mathcal{M}$, respectively for list, powerset and multiset. What we call a *state* of type T on Y is an element $\omega \in T(Y)$, for some set Y; it collects a number of elements of Y in a particular manner. In this section we abstract away from the particular type of collection. A *channel*, or sometimes more explicitly, a T*-channel*, is a collection of states, parameterised by a set. Thus, a channel is a function of the form $c\colon X \to T(Y)$. Such a channel turns an element $x \in X$ into a certain collection $c(x)$ of elements of Y. An ordinary function $f\colon X \to Y$ can be seen as a deterministic computation, giving a single outcome $f(x) \in Y$ for each input $x \in X$. A T-channel, in contrast, is a computation *of type* T. For instance, for $T = \mathcal{P}$, a channel $X \to \mathcal{P}(Y)$ is a non-deterministic computation; for $T = \mathcal{M}$, a channel $X \to \mathcal{M}(Y)$ is a 'weighted' computation that can produce multiply occurring elements.

When it is clear from the context what T is, we often write a channel using functional notation, as $c\colon X \twoheadrightarrow Y$, with a special arrow $\twoheadrightarrow$ that carries a circle on its shaft.

Definition 1.11.1 Let $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$. We use that T is functorial and comes with its own unit and flatten operations, as described in previous sections.

1 A T*-channel* from a set X to a set Y is a function of the form $c\colon X \to T(Y)$. It will be written as $c\colon X \twoheadrightarrow Y$, when T is clear from the context, and then simply called a *channel*. The set X is called the *domain* and Y is called the *codomain* of this channel.

2 For a state $\omega \in T(X)$ on the domain of a channel $c\colon X \to T(Y)$ we can form a new state $c_*(\omega)$ in $T(Y)$, on the codomain. It is defined as:

$$c_*(\omega) := (\mathit{flat} \circ T(c))(\omega) \qquad \text{where} \qquad T(X) \xrightarrow{T(c)} T(T(Y)) \xrightarrow{\mathit{flat}} T(Y).$$

This operation $\omega \mapsto c_*(\omega)$ is called *state tranformation*, sometimes with additional clarification *along the channel* c. It may also be called *pushforward*. In functional programming it is commonly called *bind*.

3 Let $c\colon X \rightsquigarrow Y$ and $d\colon Y \rightsquigarrow Z$ be two channels. Then we can compose them and get a new channel $d \odot c\colon X \rightsquigarrow Z$ via:

$$(d \odot c)(x) := d_*(c(x)) \qquad \text{so that} \qquad d \odot c = \mathit{flat} \circ T(d) \circ c.$$

Notice that we use special notation $\odot$ for composition of channels, different from standard composition $\circ$ for ordinary functions.

We first look at some examples of state transformation.

Example 1.11.2 Take $X = \{a, b, c\}$ and $Y = \{u, v\}$.

1 For $T = \mathcal{L}$ an example of a state $\omega \in \mathcal{L}(X)$ is $\omega = [c, b, b, a]$. An $\mathcal{L}$-channel $f\colon X \to \mathcal{L}(Y)$ can for instance be of the form:

$$f(a) = [u, v] \qquad f(b) = [u, u] \qquad f(c) = [v, u, v].$$

State transformation $f_*(\omega)$ amounts to 'map list' with f and then flattening. It turns a list of lists into a list, as in:

$$\begin{aligned} f_*(\omega) = \mathit{flat}\big(\mathcal{L}(f)(\omega)\big) &= \mathit{flat}\big([f(c), f(b), f(b), f(a)]\big) \\ &= \mathit{flat}\big([[v, u, v], [u, u], [u, u], [u, v]]\big) \\ &= [v, u, v, u, u, u, u, u, v]. \end{aligned}$$

2 We consider the analogous example for $T = \mathcal{P}$. We thus take as state $\sigma = \{a, b, c\}$ and as channel $g\colon X \to \mathcal{P}(Y)$, now given by subsets:

$$g(a) = \{u, v\} \qquad g(b) = \{u\} \qquad g(c) = \{u, v\}.$$

Then:

$$g_*(\sigma) = \mathit{flat}\big(\mathcal{P}(f)(\sigma)\big) = \bigcup \big\{ g(a), g(b), g(c) \big\} = \bigcup \big\{ \{u, v\}, \{u\}, \{u, v\} \big\} = \{u, v\}.$$

3 For multisets, a state in $\mathcal{M}(X)$ could be of the form $\tau = 3|a\rangle + 2|b\rangle + 5|c\rangle$ and a channel $h\colon X \to \mathcal{M}(Y)$ could have:

$$h(a) = 10|u\rangle + 5|v\rangle \qquad h(b) = 1|u\rangle \qquad h(c) = 4|u\rangle + 1|v\rangle.$$

We then get as state transformation:

$$\begin{aligned} h_*(\tau) = \mathit{flat}\big(\mathcal{M}(h)(\tau)\big) &= \mathit{flat}\big(3\big|h(a)\rangle + 2\big|h(b)\rangle + 5\big|h(c)\rangle\big) \\ &= \mathit{flat}\big(3\big|10|u\rangle + 5|v\rangle\rangle + 2\big|1|u\rangle\rangle + 5\big|4|u\rangle + 1|v\rangle\rangle\big) \\ &= 30|u\rangle + 15|v\rangle + 2|u\rangle + 20|u\rangle + 5|v\rangle = 52|u\rangle + 20|v\rangle. \end{aligned}$$

We shall mostly be using multiset — and probabilistic channels, as special case — and so we explicitly describe state transformation in these cases. So let $c\colon X \to \mathcal{M}(Y)$ be an $\mathcal{M}$-channel. Transformation of a state ω on X can be described as:

$$(c_*(\omega))(y) = \sum_{x \in X} \omega(x) \cdot c(x)(y). \tag{1.58}$$

Equivalently, we can describe the transformed state $c_*(\omega)$ as a formal sum:

$$c_*(\omega) \;=\; \sum_{y \in Y} \left(\sum_{x \in X} c(x)(y) \cdot \omega(x) \right) \big| y \rangle. \tag{1.59}$$

We now prove some general properties about state transformation and about composition of channels, demonstrating that they behave well. The proofs are based on the abstract description in Definition 1.11.1 and use the 'monad' properties of flatten and unit.

Lemma 1.11.3 *1 Channel composition $\odot$ has an identity channel, namely* $\mathit{unit} \colon Y \rightsquigarrow Y$*, so that:*

$$\mathit{unit} \odot c \;=\; c \qquad \text{and} \qquad d \odot \mathit{unit} \;=\; d,$$

for all channels $c \colon X \rightsquigarrow Y$ and $d \colon Y \rightsquigarrow Z$. Another way to write the second equation is: $d_*(\mathit{unit}(y)) = d(y)$.

2 Channel composition $\odot$ is associative:

$$e \odot (d \odot c) \;=\; (e \odot d) \odot c,$$

for all channels $c \colon X \rightsquigarrow Y$, $d \colon Y \rightsquigarrow Z$ and $e \colon Z \rightsquigarrow W$.

3 State tranformation via a composite channel is the same as two consecutive transformations:

$$(d \odot c)_*(\omega) \;=\; d_* c_*(\omega).$$

4 Each ordinary function $f \colon Y \to Z$ gives rise to a 'trivial' or 'deterministic' or 'single-valued' channel $\langle f \rangle := \mathit{unit} \circ f \colon Y \rightsquigarrow Z$. This construction $\langle - \rangle$ satisfies:

$$\langle f \rangle_*(\omega) \;=\; T(f)(\omega),$$

where T is the type of channel involved. Moreover:

$$\langle g \rangle \odot \langle f \rangle \;=\; \langle g \circ f \rangle \qquad \langle f \rangle \odot c \;=\; T(f) \circ c \qquad d \odot \langle f \rangle \;=\; d \circ f,$$

for all functions $g \colon Z \to W$ and channels $c \colon X \rightsquigarrow Y$ and $d \colon Y \rightsquigarrow W$.

Proof. We can give generic proofs, without knowing the type $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$ of the channel, by using uniform results in Lemma 1.4.5, 1.5.2, and 1.6.4 about unit and flatten. In the calculations below we carefully distinguish channel composition $\odot$ and ordinary function composition $\circ$.

1 Both equations follow from the *flat-unit* law. By Definition 1.11.1 (3):

$$\mathit{unit} \odot c \;=\; \mathit{flat} \circ T(\mathit{unit}) \circ c \;=\; \mathit{id} \circ c \;=\; c.$$

For the second equation we use naturality of *unit* in:

$$d \odot \mathit{unit} \;=\; \mathit{flat} \circ T(d) \circ \mathit{unit} \;=\; \mathit{flat} \circ \mathit{unit} \circ d \;=\; \mathit{id} \circ d \;=\; d.$$

2 The proof of associativity uses naturality and also the commutation of flatten with itself (the

'*flat*-*flat* law'), expressed as $\mathit{flat} \circ \mathit{flat} = \mathit{flat} \circ T(\mathit{flat})$.

$$
\begin{aligned}
e \circledcirc (d \circledcirc c) &= \mathit{flat} \circ T(e) \circ (d \circledcirc c) \\
&= \mathit{flat} \circ T(e) \circ \mathit{flat} \circ T(d) \circ c \\
&= \mathit{flat} \circ \mathit{flat} \circ T(T(e)) \circ T(d) \circ c && \text{by naturality of } \mathit{flat} \\
&= \mathit{flat} \circ T(\mathit{flat}) \circ T(T(e)) \circ T(d) \circ c && \text{by the } \mathit{flat}\text{-}\mathit{flat} \text{ law} \\
&= \mathit{flat} \circ T\big(\mathit{flat} \circ T(e) \circ d\big) \circ c && \text{by functoriality of } T \\
&= \mathit{flat} \circ T\big(e \circledcirc d\big) \circ c \\
&= (e \circledcirc d) \circledcirc c
\end{aligned}
$$

3 Along the same lines:

$$
\begin{aligned}
(d \circledcirc c)_*(\omega) &= (\mathit{flat} \circ T(d \circledcirc c))(\omega) \\
&= (\mathit{flat} \circ T(\mathit{flat} \circ T(d) \circ c))(\omega) \\
&= (\mathit{flat} \circ T(\mathit{flat}) \circ T(T(d)) \circ T(c))(\omega) && \text{by functoriality of } T \\
&= (\mathit{flat} \circ \mathit{flat} \circ T(T(d)) \circ T(c))(\omega) && \text{by the } \mathit{flat}\text{-}\mathit{flat} \text{ law} \\
&= (\mathit{flat} \circ T(d) \circ \mathit{flat} \circ T(c))(\omega) && \text{by naturality of } \mathit{flat} \\
&= (\mathit{flat} \circ T(d))\Big((\mathit{flat} \circ T(c))(\omega)\Big) \\
&= (\mathit{flat} \circ T(d))(c_*(\omega)) \\
&= d_*(c_*(\omega)).
\end{aligned}
$$

4 All these properties follow from elementary facts that we have seen before:

$$
\begin{aligned}
\langle f \rangle_*(\omega) &= (\mathit{flat} \circ T(\mathit{unit} \circ f))(\omega) \\
&= (\mathit{flat} \circ T(\mathit{unit}) \circ T(f))(\omega) && \text{by functoriality of } T \\
&= T(f)(\omega) && \text{by a } \mathit{flat}\text{-}\mathit{unit} \text{ law} \\
\langle g \rangle \circledcirc \langle f \rangle &= \mathit{flat} \circ T(\mathit{unit} \circ g) \circ \mathit{unit} \circ f \\
&= \mathit{flat} \circ \mathit{unit} \circ (\mathit{unit} \circ g) \circ f && \text{by naturality of } \mathit{unit} \\
&= \mathit{unit} \circ g \circ f && \text{by a } \mathit{flat}\text{-}\mathit{unit} \text{ law} \\
&= \langle g \circ f \rangle \\
\langle f \rangle \circledcirc c &= \mathit{flat} \circ T(\mathit{unit} \circ f) \circ c \\
&= \mathit{flat} \circ T(\mathit{unit}) \circ T(f) \circ c && \text{by functoriality of } T \\
&= T(f) \circ c && \text{by a } \mathit{flat}\text{-}\mathit{unit} \text{ law} \\
d \circledcirc \langle f \rangle &= \mathit{flat} \circ T(d) \circ \mathit{unit} \circ f \\
&= \mathit{flat} \circ \mathit{unit} \circ d \circ f && \text{by naturality of } \mathit{unit} \\
&= d \circ f && \text{by a } \mathit{flat}\text{-}\mathit{unit} \text{ law.}
\end{aligned}
$$

❑

In the sequel we often omit writing the brackets $\langle - \rangle$ that turn an ordinary function $f \colon X \to Y$ into a channel $\langle f \rangle$. For instance, in a state transformation $f_*(\omega)$, it is clear that we use f as a channel, so that the expression should be read as $\langle f \rangle_*(\omega)$.

Exercises

1.11.1 For a function $f\colon X \to Y$ define an *inverse image* (or *preimage*) $\mathcal{P}$-channel $f^{-1}\colon Y \rightsquigarrow X$ by:

$$f^{-1}(y) \coloneqq \{x \in X \mid f(x) = y\}.$$

Prove that:

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1} \qquad \text{and} \qquad id^{-1} = unit.$$

1.11.2 Notice that a state of type X can be identified with a channel $\mathbf{1} \to T(Y)$ with singleton set $\mathbf{1} = \{0\}$ as domain. Check that under this identification, state transformation $c_*(\omega)$ corresponds to channel composition $c \circ \omega$.

1.11.3 Let $f\colon X \rightsquigarrow Y$ be a channel.

1 Prove that if f is a $\mathcal{P}_{fin}$-channel, then the state transformation function $f_*\colon \mathcal{P}_{fin}(X) \to \mathcal{P}_{fin}(Y)$ can also be defined via freeness, namely as the unique function $\overline{f}$ in Proposition 1.5.4.

2 Similarly, show that $f_* = \overline{f}$ when f is an $\mathcal{M}$-channel, as in Exercise 1.6.13.

1.11.4 1 Describe how (non-deterministic) powerset channels can be reversed, via a bijective correspondence between functions:

$$\frac{X \longrightarrow \mathcal{P}(Y)}{\overline{Y \longrightarrow \mathcal{P}(X)}}$$

(A description of this situation in terms of 'daggers' will appear in Example 7.11.1.)

2 Show that for finite sets X, Y there is a similar correspondence for multiset channels.

1.12 The role of category theory

The previous sections have highlighted several structural properties of, and similarities between, the collection types list, subset, multiset. Later on we can add probability distributions as such a collection type. By now readers may ask: what is the underlying structure? Surely someone must have axiomatised what makes all of this work!

Indeed, this axiomatisation is part of the field of *category theory*. It provides a foundational language for mathematics, which was first formulated in the 1950s by Saunders Mac Lane and Samuel Eilenberg, see the first overview book Mac Lane (1971); Leinster (2014). Category theory focuses on the structural aspects of mathematics and shows that many mathematical constructions have the same underlying pattern. It emphasises similarities between different areas (see *e.g.* Mac Lane (1986)). Category theory has become very useful in (theoretical) computer science too, since it involves a clear distinction between specification and implementation, see books like Awodey (2006); Barr and Wells (1990); Leinster (2014); Pierce (1991); Simons (2011); Cheng (2022); Perrone (2024b). We refer to those sources for more information.

The role of category theory in capturing the mathematical essentials and estabilishing connections also applies to probability theory. William Lawvere, another founding father of the area, first worked in this direction. Lawvere himself published little on this approach to probability theory, but his ideas can be found in *e.g.* the early notes Lawvere (1962). This line of work was picked up, extended, and published by his PhD student Michèle Giry. Her name continues in the

'Giry monad' $\mathcal{G}$ of continuous probability distributions, see Section **??**. The precise source of the distribution monad $\mathcal{D}$ for discrete probability theory, that will be introduced in Section 2.1 in the next chapter, is less clear, but it can be regarded as the discrete version of $\mathcal{G}$. Probabilistic automata have been studied in categorical terms as coalgebras, see Chapter 10, and *e.g.* Sokolova (2011) and Jacobs (2016a) for general background information on the area of coalgebra. There is a recent surge in interest in more foundational, semantically oriented studies in probability theory, through the rise of probabilistic programming languages (see *e.g.* Gordon et al. (2014); Staton (2021)), probabilistic Bayesian reasoning Coecke and Spekkens (2012); Fong (2012); Culbertson and Sturtz (2014); Jacobs and Zanasi (2021), and category theory Fritz (2020). Probabilistic methods have received wider attention, for instance, via the current interest in data analytics (see the essay Anderson (2008)), in quantum probability Nielsen and Chuang (2000); Coecke and Kissinger (2016), and in cognition theory Hohwy (2013); Sloman (2005) and in computational neuroscience Griffiths et al. (2024); Parr et al. (2022).

Readers who know category theory will have recognised its implicit use in earlier sections. For readers who are not familiar (yet) with category theory, some basic concepts will be explained informally in this section. This is in no way a serious introduction to the area, for instance because the categorical notion of an adjunction is not covered. The remainder of this book will continue to make implicit use of category theory, but will make this usage increasingly explicit. Hence it is useful to know the basic concepts of category, functor, natural transformation, and monad. Category theory is sometimes seen as a difficult area to get into. But our experience is that it is easiest to learn category theory by recognising its concepts in constructions that one already knows. That is why this chapter started with concrete descriptions of various collections and their use in channels. For more solid expositions of category theory we refer to the sources listed above.

1.12.1 Categories

A category is a mathematical structure given by a collection of 'objects' with 'morphisms' between them. The requirements are that these morphisms are closed under (associative) composition and that there is an identity morphism on each object. Morphisms are also called 'maps' or 'arrows', and are written as $f\colon X \to Y$, where X, Y are objects and f is a homomorphism from X to Y. It is tempting to think of morphisms in a category as actual functions, but there are plenty of examples where this is not the case.

A category is like an abstract universe of discourse, giving a setting in which one is working, with properties of that setting depending on the category at hand. We shall give a number of illustrations. They re-appear later on in the book.

Examples 1.12.1 1 There is the category **Sets**, whose objects are sets and whose morphisms are ordinary functions between them. This is a standard example.

2 One can also restrict to finite sets as objects, in the category **FinSets**, with functions between them. This category is more restrictive, since for instance it contains objects $\boldsymbol{n} = \{0, 1, \ldots, n-1\}$ for each $n \in \mathbb{N}$, but not $\mathbb{N}$ itself. Also, in **Sets** one can take arbitrary products $\prod_{i \in I} X_i$ of objects X_i, over arbitrary index sets I, whereas in **FinSets** only finite products exist. Hence **FinSets** is a more restrictive world.

3 Monoids and monoid maps have been mentioned in Definition 1.4.1. They can be organised in

a category **Mon**, whose objects are monoids, and whose homomorphisms are monoid maps. We now have to check that monoid maps are closed under composition and that identity functions are monoid maps; this is easy. Many mathematical structures can be organised into categories in this way, where the morphisms preserve the relevant structure. For instance, one can form a category **PoSets**, with partially ordered sets (posets) as objects, and monotone functions between them as morphisms (also closed under composition, with identity).

4 For a fixed monoid M we can form the category $\mathbf{Act}_M$ of M-actions. Its objects are sets X with a monoid action $\alpha\colon M \times X \to X$. Its morphisms from $(M \times X \xrightarrow{\alpha} X)$ to $(M \times Y \xrightarrow{\beta} Y)$ are homomorphisms $f\colon X \to Y$ of actions. This means that f commutes with the action maps, see Definition 1.4.4 for details. In order to make sure that $\mathbf{Act}_M$ is indeed a category we have to check that such homomorphisms of actions are closed under composition, and that identity functions are homomorphisms of actions. This is easy.

5 For $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$ we can form the category $\mathbf{Chan}(T)$. Its objects are arbitrary sets X, but its morphisms X to Y are T-channels, $X \to T(Y)$, written as $X \twoheadrightarrow Y$. We have already seen that channels are closed under composition $\circ$ and have *unit* as identity, see Lemma 1.11.3. We can now say that $\mathbf{Chan}(T)$ is a category.

These categories of channels form good examples of the idea that a category forms a universe of discourse. For instance, in $\mathbf{Chan}(\mathcal{P})$ we are in the world of non-deterministic computation, whereas $\mathbf{Chan}(\mathcal{M})$ is the world of weighted computation in which resources are counted.

We will encounter several more examples of categories later on in the book. Occasionally, the following construction will be used. Given a category $\mathbb{C}$, a new 'opposite' category $\mathbb{C}^{\mathrm{op}}$ can be formed. It has the same objects as $\mathbb{C}$, but its morphisms are reversed. Thus $f\colon Y \to X$ in $\mathbb{C}^{\mathrm{op}}$ means $f\colon X \to Y$ in $\mathbb{C}$.

Also, given two categories $\mathbb{C}$, $\mathbb{D}$ one can form a product category $\mathbb{C} \times \mathbb{D}$. Its objects are pairs (X, A) with X an object in $\mathbb{C}$ and A an object in $\mathbb{D}$. Similarly, an arrow $(X, A) \to (Y, B)$ in $\mathbb{C} \times \mathbb{D}$ is given by a pair (f, g) of arrows $f\colon X \to Y$ in $\mathbb{C}$ and $g\colon A \to B$ in $\mathbb{D}$.

1.12.2 Functors

Category theorists like abstraction, hence the question arises: if categories are so important, then why not organise them as objects themselves in a superlarge category **Cat**, with morphisms between them preserving the relevant structure? The latter morphisms between categories are called 'functors'. More precisely, given categories $\mathbb{C}$ and $\mathbb{D}$, a functor $F\colon \mathbb{C} \to \mathbb{D}$ between them consists of two mappings, both written F, sending an object X in $\mathbb{C}$ to an object $F(X)$ in $\mathbb{D}$, and a morphism $f\colon X \to Y$ in $\mathbb{C}$ to a morphism $F(f)\colon F(X) \to F(Y)$ in $\mathbb{D}$. This mapping F should preserve composition and identities, as in: $F(g \circ f) = F(g) \circ F(f)$ and $F(id_X) = id_{F(X)}$.

Earlier we have already called some operations 'functorial' for the fact that they preserve composition and identities. We can now be a bit more precise.

1 Each $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{P}_{fin}, \mathcal{P}_*, \mathcal{P}_{\leq}[K], \mathcal{M}, \mathcal{M}_*, \mathcal{N}, \mathcal{N}_*, \mathcal{N}[K]\}$ is a functor $T\colon \mathbf{Sets} \to \mathbf{Sets}$. This has been described in the beginning of each of the sections 1.4 – 1.6.

2 Taking lists is also a functor $\mathcal{L}\colon \mathbf{Sets} \to \mathbf{Mon}$. This is in essence the content of Lemma 1.4.2. One can also view $\mathcal{P}, \mathcal{P}_{fin}$ and $\mathcal{M}$ as functors $\mathbf{Sets} \to \mathbf{Mon}$, see Lemmas 1.5.1 and 1.6.3.

Moreover, one can describe $\mathcal{P}, \mathcal{P}_{fin}$ as a functor **Sets** $\to$ **PoSets**, by considering each set of subsets $\mathcal{P}(X)$ and $\mathcal{P}_{fin}(X)$ with its subset relation $\subseteq$ as partial order. In order to verify this claim one has to check that $\mathcal{P}(f)\colon \mathcal{P}(X) \to \mathcal{P}(Y)$ is a morphism of posets, that is, forms a monotone function. But that is easy.

3 There is also a functor $J\colon$ **Sets** $\to$ **Chan**(T), for each T. It is the identity on sets / objects: $J(X) := X$. But it sends a function $f\colon X \to Y$ to the channel $J(f) := \langle f \rangle = \mathit{unit} \circ f\colon X \nrightarrow Y$. We have seen, in Lemma 1.11.3 (4), that $J(g \circ f) = J(g) \circ J(f)$ and that $J(\mathit{id}) = \mathit{id}$, where the latter identity id is unit in the category **Chan**(T). This functor J shows how to embed the world of ordinary computations (functions) into the world of compuations of type T (channels).

4 Taking the product of two sets can be described as a functor $\times\colon$ **Sets** $\times$ **Sets** $\to$ **Sets**. Its action on morphisms was already described at the end of Subsection 1.3.1, see also Exercise 1.3.3.

5 For each category $\mathbb{C}$ there is an identity functor $\mathit{id}\colon \mathbb{C} \to \mathbb{C}$, which is the identity both on objects and on morphisms. Given functors $F\colon \mathbb{C} \to \mathbb{D}$ and $G\colon \mathbb{D} \to \mathbb{E}$ we can form a composite functor $G \circ F\colon \mathbb{C} \to \mathbb{E}$. It is given in the obvious way, both on objects and on morphisms: $(G \circ F)(X) = G(F(X))$ and $(G \circ F)(f) = G(F(f))$. Sometimes we simply write GF for $G \circ F$. If F is an 'endo' functor $\mathbb{C} \to \mathbb{C}$ we write $F^2 = F \circ F\colon \mathbb{C} \to \mathbb{C}$, and more generally $F^n = F \circ \cdots \circ F$.

6 If we have functors $F_i\colon \mathbb{C}_i \to \mathbb{D}_i$, for $i = 1, 2$, then we also have a product functor $F_1 \times F_2\colon \mathbb{C}_1 \times \mathbb{C}_2 \to \mathbb{D}_1 \times \mathbb{D}_2$ between product categories, simply by $(F_1 \times F_2)(X_1, X_2) = (F_1(X_1), F_2(X_2))$, and similarly for morphisms.

1.12.3 Natural transformations

Let us move one further step up the abstraction ladder and look at morphisms between functors. These are called *natural transformations*. We have already seen examples of those as well. Given two functors $F, G\colon \mathbb{C} \to \mathbb{D}$, a natural transformation α from F to G is a collection of maps $\alpha_X\colon F(X) \to G(X)$ in $\mathbb{D}$, indexed by objects X in $\mathbb{C}$. Naturality means that α works in the same way on all objects and is expressed as follows: for each morphism $f\colon X \to Y$ in $\mathbb{C}$, the rectangle

$$\begin{array}{ccc} F(X) & \xrightarrow{\alpha_X} & G(X) \\ {\scriptstyle F(f)}\downarrow & & \downarrow{\scriptstyle G(f)} \\ F(Y) & \xrightarrow[\alpha_Y]{} & G(Y) \end{array}$$

in $\mathbb{D}$ commutes.

Such a natural transformation is often denoted by a double arrow $\alpha\colon F \Rightarrow G$. We briefly review some of the examples of natural transformations that we have seen.

1 The support maps $\mathit{supp}\colon \mathcal{L} \Rightarrow \mathcal{P}_{fin}$, $\mathit{supp}\colon \mathcal{M} \Rightarrow \mathcal{P}_{fin}$ for lists and multisets are natural transformations, see the overview Diagram 1.37. Also accumulation forms a natural transformation $\mathit{acc}\colon \mathcal{L} \Rightarrow \mathcal{N}$.

2 For each $T \in \{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$ we have described maps $\mathit{unit}\colon X \to T(X)$ and $\mathit{flat}\colon T(T(X)) \to T(X)$ and have seen naturality results about them. We can now state more precisely that they are natural transformations $\mathit{unit}\colon \mathit{id} \Rightarrow T$ and $\mathit{flat}\colon (T \circ T) \Rightarrow T$. Here we have used id as the identity functor **Sets** $\to$ **Sets**, and $T \circ T$ as the composite of T with itself, also as a functor **Sets** $\to$ **Sets**.

1.12.4 Monads

A monad on a category $\mathbb{C}$ is a functor $T\colon \mathbb{C} \to \mathbb{C}$ that comes with two natural transformations $\mathit{unit}\colon \mathit{id} \Rightarrow T$ and $\mathit{flat}\colon (T \circ T) \Rightarrow T$ satisfying:

$$\mathit{flat} \circ \mathit{unit} \;=\; \mathit{id} \;=\; \mathit{flat} \circ T(\mathit{unit}) \qquad\qquad \mathit{flat} \circ \mathit{flat} \;=\; \mathit{flat} \circ T(\mathit{flat}). \tag{1.60}$$

All the collection functors $\mathcal{L}, \mathcal{P}, \mathcal{P}_*, \mathcal{P}_{\mathit{fin}}, \mathcal{M}, \mathcal{M}_*, \mathcal{N}, \mathcal{N}_*$ that we have seen so far are monads, see *e.g.*, Lemma 1.4.5, 1.5.2, or 1.6.4. For each monad T we can form a category $\mathbf{Chan}(T)$ of T-channels, that capture computations of type T, see Subsection 1.12.1. In category theory this is called the Kleisli category of T. Composition in this category $\mathbf{Chan}(T)$ is called Kleisli composition. In this book it is written as $\circledcirc$, where the context should make clear what the monad T at hand is.

Monads have become popular in functional programming Moggi (1991) as mechanisms for including special effects (*e.g.*, for input-output, writing, side-effects, continuations) into a functional programming language[3]. The structure of probabilistic computation is also given by monads, namely by the discrete distribution monads $\mathcal{D}, \mathcal{D}_\infty, \mathcal{D}_{\leq 1}$ and by the continuous distribution monad $\mathcal{G}$.

We thus associate the (Kleisli) category $\mathbf{Chan}(T)$ of channels with a monad T. A second category is associated with a monad T, namely the category $\mathcal{EM}(T)$ of "Eilenberg-Moore" algebras. The objects of $\mathcal{EM}(T)$ are algebras $\alpha\colon T(X) \to X$, satisfying $\alpha \circ \mathit{unit} = \mathit{id}$ and $\alpha \circ \mathit{flat} = \alpha \circ T(\alpha)$. We have seen algebras for the monads $\mathcal{L}, \mathcal{P}_{\mathit{fin}}$, and $\mathcal{M}$ in Propositions 1.4.6, 1.5.5, and 1.6.6. They capture monoids, commutative idempotent monoids, and commutative monoids respectively. A morphism in $\mathcal{EM}(T)$ is a morphism of algebras, given by a commuting rectangle, as described in these propositions. In general, algebras of a monad capture algebraic structure in a uniform manner.

Here is an easy result that describes so-called *writer* monads.

Lemma 1.12.2 *Let $M = (M, +, 0)$ be an arbitrary monoid. The mapping $X \mapsto M \times X$ forms a monad on the category* **Sets**.

Proof. Let us write $T(X) = M \times X$. For a function $f\colon X \to Y$ we define $T(f)\colon M \times X \to M \times Y$ by $T(f)(m, x) = (m, f(x))$. There is a unit map $\mathit{unit}\colon X \to M \times X$, namely $\mathit{unit}(x) = (0, x)$ and a flattening map $\mathit{flat}\colon M \times (M \times X) \to M \times X$ by $\mu(m, m', x) = (m + m', x)$. We skip naturality and concentrate on the monad equations (1.60). First, for $(m, x) \in T(X) = M \times X$,

$$\begin{aligned}(\mathit{flat} \circ \mathit{unit})(m, x) \;&=\; \mathit{flat}(0, m, x) \;=\; (0 + m, x) \;=\; (m, x)\\ (\mathit{flat} \circ T(\mathit{unit}))(m, x) \;&=\; \mathit{flat}(m, \mathit{unit}(x)) \;=\; \mathit{flat}(m, 0, x) \;=\; (m, x).\end{aligned}$$

Next, the flatten-equation holds by associativity of the monoid addition $+$. This is left to the reader. ❑

We have seen natural transformations as maps between functors. In the special case where the functors involved are monads, these natural transformations can be called maps of monads if they additionally commute with the unit and flatten maps.

Definition 1.12.3 Let $T_1 = (T_1, \mathit{unit}_1, \mathit{flat}_1)$ and $T_2 = (T_2, \mathit{unit}_2, \mathit{flat}_2)$ be two monads (on

[3] See the online overview https://wiki.haskell.org/Monad_tutorials_timeline

Sets). A *map / homomorphism of monads* from T_1 to T_2 is a natural transformation $\alpha\colon T_1 \Rightarrow T_2$ that commutes with unit and flatten in the sense that the two diagrams

$$\begin{tikzcd} & X \arrow[dl, bend right, "unit_1"'] \arrow[dr, bend left, "unit_2"] & \\ T_1(X) \arrow[rr, "\alpha"] & & T_2(X) \end{tikzcd} \qquad \begin{tikzcd} T_1(T_1(X)) \arrow[r, "\alpha"] \arrow[d, "flat_1"'] & T_2(T_1(X)) \arrow[r, "T_2(\alpha)"] & T_2(T_2(X)) \arrow[d, "flat_2"] \\ T_1(X) \arrow[rr, "\alpha"] & & T_2(X) \end{tikzcd}$$

commute, for each set X.

We have already seen an example of a map of monads, namely the support map $supp\colon \mathcal{L} \Rightarrow \mathcal{P}_{fin}$ in Lemma 1.5.3 (3). Other simple examples of maps of monads arise from the writer monads from Lemma 1.12.2: if $f\colon M_1 \to M_2$ is a map of monoids, then the maps $\alpha_X := f \times id\colon M_1 \times X \to M_2 \times X$ form a natural transformation and a map of monads.

For a historical account of monads and their applications we refer to Hyland and Power (2007).

Exercises

1.12.1 We have seen the functor $J\colon \mathbf{Sets} \to \mathbf{Chan}(T)$. Check that there is also a functor $\mathbf{Chan}(T) \to \mathbf{Sets}$ in the opposite direction, which is $X \mapsto T(X)$ on objects, and $c \mapsto c_*$ on morphisms. Check explicitly that composition is preserved, and find the earlier result that stated that fact implicitly.

1.12.2 Recall from (1.27) the subset $\mathcal{N}[K](X) \subseteq \mathcal{M}(X)$ of natural multisets with K elements. Prove that $\mathcal{N}[K]$ is a functor $\mathbf{Sets} \to \mathbf{Sets}$.

1.12.3 Show that Exercise 1.11.1 implicitly describes a functor $\mathbf{Sets}^{\mathrm{op}} \to \mathbf{Chan}(\mathcal{P})$, which is the identity on objects.

1.12.4 Show that the *zip* function from Exercise 1.3.7 is natural: for each pair of functions $f\colon X \to U$ and $g\colon Y \to V$ the following diagram commutes.

$$\begin{tikzcd} X^K \times Y^K \arrow[r, "zip"] \arrow[d, "f^K \times g^K"'] & (X \times Y)^K \arrow[d, "(f \times g)^K"] \\ U^K \times V^K \arrow[r, "zip"] & (U \times V)^K \end{tikzcd}$$

1.12.5 Fill in the remaining details in the proof of Lemma 1.12.2: that T is a functor, that *unit* and *flat* are natural transformation, and that the flatten equation holds.

1.12.6 For arbitrary sets X, A, write $X + A$ for the disjoint union (coproduct) of X and A, which may be described explicitly by tagging elements with numbers $1, 2$ in order to distinguish them:

$$X + A = \{(x, 1) \mid x \in X\} \cup \{(a, 2) \mid a \in A\}.$$

Write $\kappa_1\colon X \to X + A$ and $\kappa_2\colon A \to X + A$ for the two obvious functions.

1 Keep the set A fixed and show that the mapping $X \mapsto X + A$ can be extended to a functor $\mathbf{Sets} \to \mathbf{Sets}$.

2 Show that it is actually a monad; it is sometimes called the *exception* monad, where the elements of A are seen as exceptions in a computation.

1.12.7 Check that the support and accumulation functions form maps of monads in the situations:

1 $supp\colon \mathcal{M}(X) \Rightarrow \mathcal{P}_{fin}(X)$;
2 $acc\colon \mathcal{L}(X) \Rightarrow \mathcal{N}(X)$.

This property was announced in the very beginning of this chapter, in Diagram (1.1).

1.12.8 Let $T = (T, unit, flat)$ be a monad. By definition, it involves T as a functor $T\colon \mathbf{Sets} \to \mathbf{Sets}$. Show that T can be 'extended' to a functor $\overline{T}\colon \mathbf{Chan}(T) \to \mathbf{Chan}(T)$. It is defined on objects as $\overline{T}(X) := T(X)$ and on a morphism $f\colon X \twoheadrightarrow Y$ as:

$$\overline{T}(f) := \Big(T(X) \xrightarrow{T(f)} T(T(Y)) \xrightarrow{flat} T(Y) \xrightarrow{unit} T(T(Y))\Big).$$

Prove that $\overline{T}$ is a functor, *i.e.* that it preserves (channel) identities and composition.

2

Discrete probability distributions

At this stage, we are well-prepared to move into the area of probability theory. This chapter introduces the basics of discrete probability distributions and of probabilistic channels / computations. These notions will play a central role in the rest of this book. Distributions will be defined as special multisets where multiplicities add up to one. As a result, there is a simple inclusion of the set of distributions in the set of multisets (on the same space). This allows us to use the same (ket) notation for distributions that we used for multisets. Also, much of the structure that we have seen for multisets restricts to distributions, especially functoriality and the monad structure given by unit and flatten.

In the other direction, this chapter describes the 'frequentist learning' operation, which turns a (non-empty) multiset into a distribution, essentially by normalisation. It yields probabilities via counting. This basic operation will show up in many situations.

Distributions can be put in parallel, via an operation called tensor product $\otimes$. It gives a 'joint' distribution, on the product of the underlying spaces. This tensor also exists for subsets and multisets. However, joint distributions are special since they may involve dependencies / correlations which create a dynamic that is essential to the field. These dependencies arise because all probabilities in a distribution have to add up to one. This chapter introduces the basics of such joint / product distributions and illustrates how they can be used to introduce new (image) distributions, like 'multinomial', 'coupon' and 'coincidence'.

The previous chapter introduced the concept of channel, as a special function for a form of computation that is determined by the monad involved. An example is non-deterministic computation for the powerset monad $\mathcal{P}$. This chapter extends these ideas to probabilistic computation, via the distribution monad $\mathcal{D}$. The resulting probabilistic channels are in essence conditional probabilities or stochastic matrices. They will be used for transformation of states / distributions along the channel, in a forward direction. This is also known as prediction. Channels can be composed both sequentially and in parallel, giving an expressive calculus of channels. For this calculus we use a graphical representation in terms of so-called string diagrams Piedeleu and Zanasi (2025); Coecke and Kissinger (2016). They look a bit like Bayesian networks, since the conditional probability tables that are associated with nodes in a Bayesian network are instances of probabilistic channels. But there are essential graphical differences between Bayesian networks and string diagrams, for instance in the way copying is handled. We shall prefer string diagrams over Bayesian networks since they have a clear semantics in terms of channels.

Important examples of probability distributions are obtained via the intuitive model of an urn filled with multiple balls of different colours. The probability of drawing a ball of a particular colour is determined by the proportion of balls of that colour in the urn. There are multiple ways to proceed after a ball has been drawn: it can be left out, it can be returned to the urn, or it can

be returned together with an additional ball of the same colour as the drawn ball. These different modes give rise to hypergeometric, multinomial and Pólya distributions. They are introduced in this chapter as illustration, but they will be studied more systematically in the next chapter.

This chapter also introduces the convolution (sum) of two distributions. It works when the underlying space of these distributions is the same and happens to be a commutative monoid. This construction is mathematically well-behaved. It occurs regularly and is thus included in this first chapter on probability.

Apart from a first mathematical perspective on discrete probability theory, this chapter also offers an impression of how two particular probabilistic programming languages support probabilistic computations, see Section 2.8 for the language ProbLog and Section 2.9 for WebPPL. Both languages are open source and can be installed[1] easily, for one's own experiments. Later on we occasionally illustrate other concepts, notably conditioning, in ProbLog and WebPPL.

One way to compare two distributions (on the same space) is via the so-called Kullback-Leibler divergence. Its essential properties will be discussed towards the end of the chapter. Later on we shall also define a metric distance function on distributions. The chapter closes with a systematic look at exchangeability, in two forms: transposition and substitution. These two versions are closely related to the operations of accumulation and matching that we saw in the previous chapter.

2.1 Probability distributions

This section introduces discrete probability distributions, which we often simply call distributions. We may also call them states. In the literature they are sometimes called multinomial or categorical distributions, but we avoid those terms since they clash with the terminology in this book. The notation and definitions that we use for distributions are inherited from multisets. Indeed, a distribution is a special multiset, with multiplicities adding up to one.

In this section we first introduce *finite* distributions, having finite support, like for multisets. Towards the end we also describe distributions with *infinite* support. There are several important examples of such infinite distributions, but for the most part the emphasis will be on the finite ones.

A (finite discrete probability) *distribution* over a set X is a finite formal convex sum of the form:

$$r_1|x_1\rangle + \cdots + r_n|x_n\rangle \qquad \text{where} \qquad x_i \in X \text{ and } r_i \in [0,1] \text{ with } \textstyle\sum_i r_i = 1.$$

We can write such an expression as a (formal) sum $\sum_i r_i|x_i\rangle$. It is called a *convex* sum since the probabilities $r_i \in [0,1]$ add up to one. Thus, a distribution over X is a special 'probabilistic' multiset, inhabiting a subset $\mathcal{D}(X) \subseteq \mathcal{M}(X)$. In particular, we may equivalently describe a distribution as a function $\omega\colon X \to [0,1]$ with finite support and with $\sum_x \omega(x) = 1$. As for multisets, we often switch back-and-forth between the representations as formal sum and as function.

Definition 2.1.1 For an arbitrary set X we write $\mathcal{D}(X) \subseteq \mathcal{M}(X)$ for the set of distributions on X, consisting of multisets $\omega \in \mathcal{M}(X) \subseteq \mathbb{R}^X_{\geq 0}$ whose multiplicities add up to one: $\sum_{x \in X} \omega(x) = 1$.

In functional form a distribution $\omega \in \mathcal{D}(X)$ restricts to a map $\omega\colon X \to [0,1]$ with the unit

[1] See the webpages https://dtai.cs.kuleuven.be/problog/problog1/problog1.html and https://webppl.readthedocs.io/en/master/ for details.

interval $[0,1] \subseteq \mathbb{R}_{\geq 0}$ as codomain. It is sometimes called the *probability mass function* or the *density function*. However, the term 'density' is more often reserved for continuous distributions. Therefore, we shall avoid it in the discrete case.

When the set X is finite, we write $\mathcal{D}_{fs}(X) \subseteq \mathcal{D}(X)$ for the subset of distributions $\omega \in \mathcal{D}(X)$ with *full support*; this means that $supp(\omega) = X$, that is, $\omega(x) > 0$, for each $x \in X$.

Via the inclusions $\mathcal{D}(X) \subseteq \mathcal{M}(X)$, we use the same conventions for distributions, as for multisets; they were described in the three bullet points in the beginning of Section 1.6. The set X is often called the *sample space*, see *e.g.* Ross (2007), the *outcome space*, the *underlying space*, or simply the *underlying set*. Each element $x \in X$ gives rise to a distribution $1|x\rangle \in \mathcal{D}(X)$, which is 1 on x and 0 everywhere else. It is called a Dirac distribution, a point mass, a point state, or also a point distribution. The mapping $x \mapsto 1|x\rangle$ is the unit function $unit\colon X \to \mathcal{D}(X)$.

This unit map does not exist when one switches to distributions $\mathcal{D}_{fs}$ with full support — unless the underlying set is a singleton. There are other desirable properties of distributions that disappear when we require support to be full — like topological completeness, see Theorem 4.8.11. Therefore, we prefer to work with distributions in general, without requiring full support. We shall only use full support when really needed, like in stick breaking, see Theorem 2.2.6.

In a distribution, by definition, the probabilities add up to one. It makes sense to allow that the sum of probabilities is below one. One then speaks of *subdistributions*. The set of such subdistributions is written as $\mathcal{D}_{\leq 1}(X)$. We thus have inclusions $\mathcal{D}(X) \subseteq \mathcal{D}_{\leq 1}(X) \subseteq \mathcal{M}(X)$. We occasionally come accross subdistributions, but our main story is about 'proper' distributions, with probabilities adding up to one.

For a coin we can use the set $\{H, T\}$ with elements for head and tail as sample space. A fair coin is described on the left below, as a distribution over this set; the distribution on the right gives a coin with a slight bias.

$$\tfrac{1}{2}|H\rangle + \tfrac{1}{2}|T\rangle \qquad\qquad 0.51|H\rangle + 0.49|T\rangle.$$

In general, for a non-empty finite set X there is a *uniform* distribution $unif_X \in \mathcal{D}(X)$ that assigns the same probability to each element. Thus, it is given by:

$$unif_X := \sum_{x \in X} \frac{1}{|X|}\big|x\rangle \qquad \text{where } |X| \in \mathbb{N}_{>0} \text{ is the size of } X.$$

The above fair coin is a uniform distribution on the two-element set $\{H, T\}$. Similarly, a fair dice can be described as $unif_{pips} = \frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$, where $pips = \{1, 2, 3, 4, 5, 6\}$. Figure 2.1 shows bar charts of several distributions. The last one describes the letter frequencies in English for the latin alphabet. One commonly does not distinguish upper and lower cases in such frequencies, so we take the 26-element set $A = \{a, b, c, \ldots, z\}$ of lower cases as sample space. The distribution itself can be described as formal sum:

$$\begin{aligned}
&0.082|a\rangle + 0.015|b\rangle + 0.028|c\rangle + 0.043|d\rangle + 0.13|e\rangle + 0.022|f\rangle + 0.02|g\rangle \\
&\quad + 0.061|h\rangle + 0.07|i\rangle + 0.0015|j\rangle + 0.0077|k\rangle + 0.04|l\rangle + 0.024|m\rangle + \\
&\quad + 0.067|n\rangle + 0.075|o\rangle + 0.019|p\rangle + 0.00095|q\rangle + 0.06|r\rangle + 0.063|s\rangle \\
&\quad + 0.091|t\rangle + 0.028|u\rangle + 0.0098|v\rangle + 0.024|w\rangle + 0.0015|x\rangle + 0.02|y\rangle + 0.0074|z\rangle.
\end{aligned}$$

These frequencies have been copied from Wikipedia. Interestingly, they do not precisely add up to 1, but to 1.01085, probably due to rounding. Thus, strictly speaking, this is not a probability distribution but a multiset.

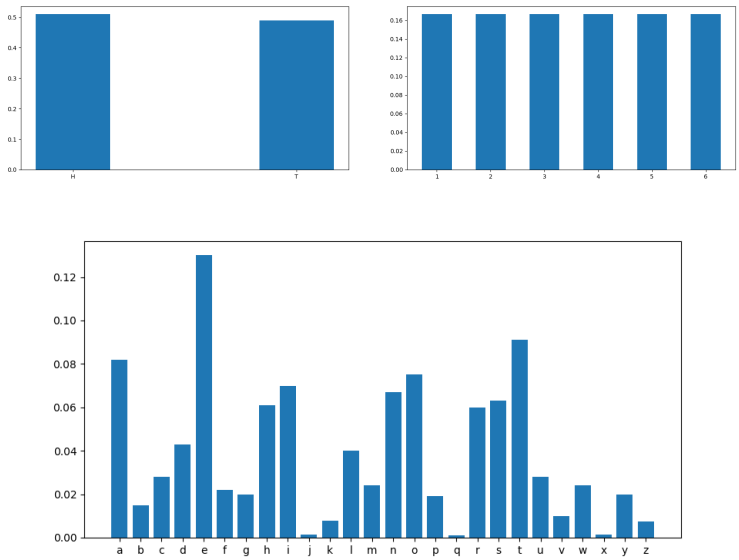


Figure 2.1 Plots, in the top row, of a slightly biased coin distribution $0.51|H\rangle + 0.49|T\rangle$ and of a fair (uniform) dice distribution on $\{1, 2, 3, 4, 5, 6\}$. The bar chart at the bottom is of the distribution of letter frequencies in English. We see that the letter e has the highest probability, as it occurs most frequently.

Data visualisation is a topic in itself, see *e.g.* Spiegelhalter (2019). In this book we shall mostly use simple bar charts, as in Figure 2.1. For continuous distributions, in Chapter **??**, they will involve continuous lines.

Below we describe several standard examples of distributions that will play an important role in the remainder of the book. We use the above formal convex sum notation with kets. Also, we use explicit names to refer to different distributions, in contrast to traditional notation where distributions are typically left implicit and are all written in the same way, as collection of probabilities $p(x)$.

Example 2.1.2 We discuss several distributions and see how they are naturally parameterised.

1 The coin that we have seen above can be parametrised via a 'bias' probability $r \in [0, 1]$. The resulting coin will be called *flip* and is defined as:

$$\mathit{flip}(r) \coloneqq r|1\rangle + (1-r)|0\rangle.$$

It uses 1 for 'head' and 0 for 'tail'. We may thus see *flip* as a function $\mathit{flip}\colon [0, 1] \to \mathcal{D}(\mathbf{2})$ from probabilities to distributions over the sample space $\mathbf{2} = \{0, 1\}$ of Booleans. This $\mathit{flip}(r)$ is often called the *Bernoulli* distribution, with parameter $r \in [0, 1]$.

2 For each number $K \in \mathbb{N}$ and probability $r \in [0, 1]$ there is the *binomial* distribution $bn[K](r) \in \mathcal{D}(\{0, 1, \ldots, K\})$. It captures probabilities for iterated coin flips, and is given by the convex

sum:

$$bn[K](r) \coloneqq \sum_{0\leq k\leq K} \binom{K}{k} \cdot r^k \cdot (1-r)^{K-k} \,|k\rangle. \tag{2.1}$$

The multiplicity probability before $|k\rangle$ in this expression is the chance of getting k heads of out K coin flips, where each flip has bias $r \in [0,1]$. The binomial coefficient $\binom{K}{k}$ is needed because we ignore the order of heads and tails. The probabilities in the above expression (2.1) add up to 1 by the Multinomial Theorem (1.42).

Here is an example, for $K = 4$ and $r = \frac{1}{3}$.

$$\begin{aligned} bn[4](\tfrac{1}{3}) &= \tbinom{4}{0} \cdot (\tfrac{1}{3})^0 \cdot (\tfrac{2}{3})^4 |0\rangle + \tbinom{4}{1} \cdot (\tfrac{1}{3})^1 \cdot (\tfrac{2}{3})^3 |1\rangle + \tbinom{4}{2} \cdot (\tfrac{1}{3})^2 \cdot (\tfrac{2}{3})^2 |2\rangle \\ &\qquad + \tbinom{4}{3} \cdot (\tfrac{1}{3})^3 \cdot (\tfrac{2}{3})^1 |3\rangle + \tbinom{4}{4} \cdot (\tfrac{1}{3})^4 \cdot (\tfrac{2}{3})^0 |4\rangle \\ &= \tfrac{16}{81}|0\rangle + \tfrac{32}{81}|1\rangle + \tfrac{8}{27}|2\rangle + \tfrac{8}{81}|3\rangle + \tfrac{1}{81}|4\rangle. \end{aligned}$$

We can organise binomial distributions as a function $bn[K]\colon [0,1] \to \mathcal{D}(\{0,1,\ldots,K\})$. The binomial probabilities are plotted as bar charts in Figure 2.2, for two binomial distributions, both for $K = 10$ on the associated sample space $\{0,1,\ldots,10\}$.

Later we make precise how the binomial distribution $bn[K](r)$ is a K-fold 'convolution' sum of the flip distribution $flip(r)$, see Proposition 2.7.6.

There is a reason for the difference between square and round brackets in writing $bn[K](r)$. In general, we write the parameters that determine the type of a function between square brackets, as in $bn[K]\colon [0,1] \to \mathcal{D}(\{0,1,\ldots,K\})$. These square brackets come before the round ones. The parameter between round brackets is the argument, to which the function is applied. We thus get $bn[K](r) \in \mathcal{D}(\{0,1,\ldots,K\})$, for an argument $r \in [0,1]$.

3 In the previous chapter we have seen several results about sums of a particular kind. We can now use them as basis for various distributions. For instance, for each number $n \in \mathbb{N}$ the binomial coefficients $\binom{n}{i}$ for $0 \leq i \leq n$ add up to 2^n, see Exercise 1.2.4. Thus we can form a binomial coefficient distribution:

$$bc[n] \coloneqq \sum_{0\leq i\leq n} \frac{\binom{n}{i}}{2^n} \,|i\rangle \in \mathcal{D}(\{0,\ldots,n\}). \tag{2.2}$$

For instance,

$$bc[5] = \tfrac{1}{32}|0\rangle + \tfrac{5}{32}|1\rangle + \tfrac{5}{16}|2\rangle + \tfrac{5}{16}|3\rangle + \tfrac{5}{32}|4\rangle + \tfrac{1}{32}|5\rangle.$$

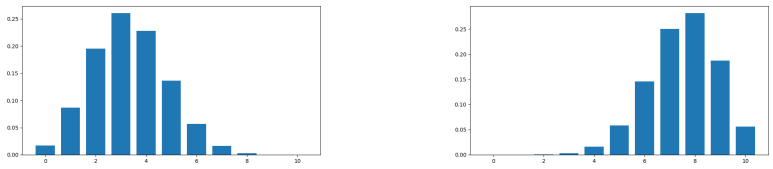
In Subsection 1.10.4 we have seen the N-nomial numbers that generalise the binomial coefficients. They can be turned into a 'nomial' distribution $nom[N,K] \in \mathcal{D}(\{0,\ldots,(N-1)\cdot K\})$ via:

$$nom[N,K] \coloneqq \sum_{0\leq i\leq (N-1)\cdot K} \frac{C_N(K,i)}{N^K} \,|i\rangle. \tag{2.3}$$

This yields a distribution by Lemma 1.10.2 (1).

4 In a similar way, for a non-empty finite set X, say with size $N \coloneqq |X| > 0$, and a number $K \in \mathbb{N}$ we can form distributions over the set of natural multisets $\mathcal{N}[K](X)$. The uniform distribution over such multisets takes the form:

$$unif_{\mathcal{N}[K](X)} \coloneqq \sum_{\varphi\in\mathcal{N}[K](X)} \frac{1}{\left(\!\binom{N}{K}\!\right)} \,|\varphi\rangle \in \mathcal{D}\big(\mathcal{N}[K](X)\big). \tag{2.4}$$



$bn[10](\frac{1}{3})$ $bn[10](\frac{3}{4})$

Figure 2.2 Plots of two binomial distributions, see Examples 2.1.2.

This is a distribution since the set $\mathcal{N}[K](X)$ contains $\left(\!\binom{N}{K}\!\right)$-many elements, see Proposition 1.8.7. When we write out such a distribution we use 'kets within kets' or 'nested kets'. For instance, for $X = \{a, b\}$ and $K = 3$ we get $\left(\!\binom{2}{3}\!\right) = \frac{4!}{1!\cdot 3!} = 4$ multisets of size. They all occur with probability $\frac{1}{4}$ in the uniform distribution of the form:

$$unif_{\mathcal{N}[3](\{a,b\})} = \tfrac{1}{4}\Big|3|a\rangle\Big\rangle + \tfrac{1}{4}\Big|2|a\rangle + 1|b\rangle\Big\rangle + \tfrac{1}{4}\Big|1|a\rangle + 2|b\rangle\Big\rangle + \tfrac{1}{4}\Big|3|b\rangle\Big\rangle.$$

The expression on the right-hand-side combines 'big' probability kets with 'small' multiset kets, inside the big kets. This makes sense.

One can also form a distribution of multiset coefficients:

$$mc[X, K] := \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\mathbf{(}\varphi\mathbf{)}}{N^K}\,|\varphi\rangle \in \mathcal{D}\big(\mathcal{N}[K](X)\big). \tag{2.5}$$

This is a distribution since the sum of all coefficients $\mathbf{(}\varphi\mathbf{)}$, for $\varphi \in \mathcal{N}[K](X)$, is N^K, see Exercise 1.7.9. Later on we shall see many more such distributions on multisets, where the multisets arise as draws of multiple coloured balls from an urn, see Section 2.6 and Chapter 3. The multiset coefficient distribution (2.5) is in fact an instance of such a 'multinomial' draw distribution, where the colour distribution is uniform, see Exercise 2.6.2.

We have written $\mathcal{D}(X) \subseteq \mathcal{M}(X)$ for the subset of distributions on a set X, with multiplicities adding up to one. In Definition 1.6.2 we have seen that $\mathcal{M}$ is functorial: it acts not only on sets, but also on functions. This works for $\mathcal{D}$ as well, via suitable restriction.

Lemma 2.1.3 *The mapping $X \mapsto \mathcal{D}(X)$ is functorial: for a function $f\colon X \to Y$ we have $\mathcal{D}(f)\colon \mathcal{D}(X) \to \mathcal{D}(Y)$ defined, either in ket form or in functional form, as:*

$$\mathcal{D}(f)\big(\textstyle\sum_i r_i|x_i\rangle\big) := \textstyle\sum_i r_i|f(x_i)\rangle \qquad \textit{or as:} \qquad \mathcal{D}(f)(\omega)(y) := \displaystyle\sum_{x \in f^{-1}(y)} \omega(x).$$

A distribution of the form $\mathcal{D}(f)(\omega) \in \mathcal{D}(Y)$, for $\omega \in \mathcal{D}(X)$, is sometimes called an image distribution*. One also says that ω is pushed forward along the function f.*

Proof. One has to check that $\mathcal{D}(f)(\omega)$ is a distribution again, that is, that its multiplicities add up to one. This works as follows.

$$\sum_{y \in Y} \mathcal{D}(f)(\omega)(y) = \sum_{y \in Y} \sum_{x \in f^{-1}(y)} \omega(x) = \sum_{x \in X} \omega(x) = 1. \qquad \square$$

We present several examples where this functoriality of $\mathcal{D}$ is used. It provides a powerful technique to produce new distributions from old, as images.

Examples 2.1.4 1 Suppose we have a distribution $\omega \in \mathcal{D}(\mathbb{R})$ on the real numbers. One may wish to shift this distribution forward or backward on the real line. There is such a shift function $shift_{\mathbb{R}} \colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ on the reals, namely $shift_{\mathbb{R}}(s,x) = s+x$, where s is the shift factor. We can adapt this function from reals to distributions on reals via functoriality of $\mathcal{D}$. This happens via a shift function $shift_{\mathcal{D}} \colon \mathbb{R} \times \mathcal{D}(\mathbb{R}) \to \mathcal{D}(\mathbb{R})$ defined via functoriality of $\mathcal{D}$, as:

$$shift_{\mathcal{D}}(s,\omega) \coloneqq \mathcal{D}(shift_{\mathbb{R}}(s,-))(\omega) = \mathcal{D}(x \mapsto s+x)(\omega). \tag{2.6}$$

Explicitly, for a distribution $\omega = \sum_i r_i|x_i\rangle \in \mathcal{D}(\mathbb{R})$ one gets, by applying the formulation of Definition 2.1.3,

$$\begin{aligned} shift_{\mathcal{D}}\Big(s, \textstyle\sum_i r_i|x_i\rangle\Big) &= \mathcal{D}(shift_{\mathbb{R}}(s,-))(\omega)\Big(\textstyle\sum_i r_i|x_i\rangle\Big) \\ &= \textstyle\sum_i r_i\Big|shift_{\mathbb{R}}(s,x_i)\rangle = \sum_i r_i\Big|s+x_i\rangle. \end{aligned}$$

This shows how all elements x_i in the support of the distribution are shifted with a factor s. An example of a distribution that is shifted 5 positions forward can be found in Figure 2.3.

Shifting involves addition of a fixed number. One may also wish to multiply with a fixed number, in what is called scaling or rescaling. It starts from the function $scale_{\mathbb{R}} \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ on the reals given by $scale_{\mathbb{R}}(s,x) = s \cdot x$, where s is now the scale factor. This scale operation on reals can also be extended to an operation on distribution, of the form $scale_{\mathcal{D}} \colon \mathbb{R}\times\mathcal{D}(\mathbb{R}) \to \mathcal{D}(\mathbb{R})$, via:

$$scale_{\mathcal{D}}(s,\omega) \coloneqq \mathcal{D}(scale_{\mathbb{R}}(s,-))(\omega) = \sum_{x\in X} \omega(x)\big|s\cdot x\rangle. \tag{2.7}$$

In Figure 2.3 the same (original) distribution is also rescaled with a factor 3. One sees that the distribution is now more spread out.

The shift and scale operations are monoid actions, both on reals and on distributions. This fact will be described more systematically in Proposition 2.1.5 below.

2 Computing marginals of 'joint' distributions involves functoriality of $\mathcal{D}$. In general, one speaks of a *joint* distribution if its sample space is a product set, of the form $X_1\times X_2$, or more generally, $X_1 \times \cdots \times X_n$, for $n \geq 2$. The i-th marginal of a joint distribution $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n)$ is defined as the image distribution $\mathcal{D}(\pi_i)(\omega) \in \mathcal{D}(X_i)$, obtained via the i-th projection function $\pi_i \colon X_1 \times \cdots \times X_n \to X_i$.

For instance, the first marginal of the joint distribution,

$$\omega = \tfrac{1}{12}\big|H,0\rangle + \tfrac{1}{6}\big|H,1\rangle + \tfrac{1}{3}\big|H,2\rangle + \tfrac{1}{6}\big|T,0\rangle + \tfrac{1}{12}\big|T,1\rangle + \tfrac{1}{6}\big|T,2\rangle$$

on the product space $\{H,T\} \times \{0,1,2\}$ is the distribution on $\{H,T\}$ that is computed explicitly as:

$$\begin{aligned} \mathcal{D}(\pi_1)(\omega) &= \tfrac{1}{12}\big|\pi_1(H,0)\rangle + \tfrac{1}{6}\big|\pi_1(H,1)\rangle + \tfrac{1}{3}\big|\pi_1(H,2)\rangle \\ &\qquad + \tfrac{1}{6}\big|\pi_1(T,0)\rangle + \tfrac{1}{12}\big|\pi_1(T,1)\rangle + \tfrac{1}{6}\big|\pi_1(T,2)\rangle \\ &= \tfrac{1}{12}\big|H\rangle + \tfrac{1}{6}\big|H\rangle + \tfrac{1}{3}\big|H\rangle + \tfrac{1}{6}\big|T\rangle + \tfrac{1}{12}\big|T\rangle + \tfrac{1}{6}\big|T\rangle = \tfrac{7}{12}\big|H\rangle + \tfrac{5}{12}\big|T\rangle. \end{aligned}$$

In the same way one obtains as second marginal $\mathcal{D}(\pi_2)(\omega) = \frac{1}{4}\big|0\rangle + \frac{1}{4}\big|1\rangle + \frac{1}{2}\big|3\rangle$.

3 Suppose we throw two (fair) dices and we look at the maximum of the pips that come up. What distribution do we get? We can describe it as an image distribution.

Recall that we write $pips = \{1,2,3,4,5,6\}$ for the sample space of a dice. Let $\max \colon pips \times$

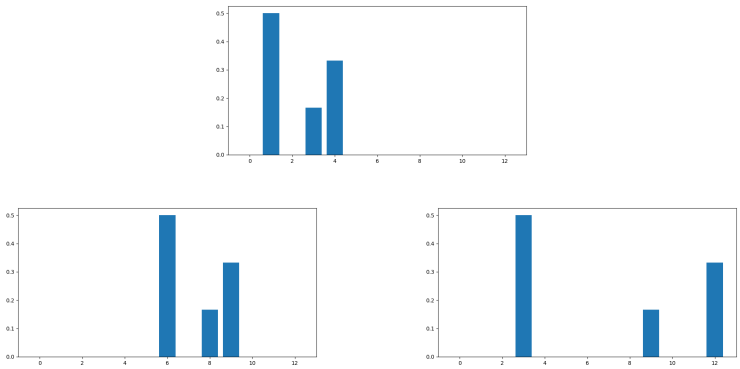


Figure 2.3 The distribution on top is shifted 5 positions forward on the lower left and rescaled with a factor 3 on the lower right, see Example 2.1.4 (1).

$pips \to pips$ be the function that take the maximum of two numbers. We use the uniform distribution $unif \in \mathcal{D}(pips \times pips)$ given by $unif = \sum_{i,j \in pips} \frac{1}{36}|i, j\rangle$. Then, in functional form:

$$\begin{aligned} \mathcal{D}(\max)(unif)(k) &= \sum_{i,j \text{ with } \max(i,j)=k} unif(i, j) \\ &= \sum_{i \leq k} unif(i, k) + \sum_{j<k} unif(k, j) = \frac{2k-1}{36}. \end{aligned}$$

We can write this image distribution in ket notation as:

$$\mathcal{D}(\max)(unif) = \tfrac{1}{36}|1\rangle + \tfrac{3}{36}|2\rangle + \tfrac{5}{36}|3\rangle + \tfrac{7}{36}|4\rangle + \tfrac{9}{36}|5\rangle + \tfrac{11}{36}|6\rangle.$$

In this illustration we use the uniform distribution $unif$ on the product space $pips \times pips$. We will discuss products $\otimes$ of distributions in Section 2.3. Then we can equivalently describe this distribution $unif$ on the product as a product $\otimes$ of two dices — which are themselves uniform distributions. This is the same, since products $\otimes$ of uniform distributions are uniform, see Exercise 2.3.4.

In the first of these examples we have seen shifting and scaling of distributions. There is some more abstract structure behind it that we like to make explicit in the proposition below. This result is a first illustration of what the approach of this book adds to the field of probability theory. It is also an illustration of the role of category theory to organise relevant structures. The result itself is not very deep, but it does capture basic properties in a compact manner. We shall see similar such results, for instance in Proposition 2.7.2 on convolutions.

We recall the notion of monoid actions $M \times X \to X$ from Definition 1.4.4, where M is the monoid. We also recall from Exercise 1.4.8 that each monoid M acts on itself via its operation $M \times M \to M$. This happens twice in Example 2.1.4 (1) where $M = \mathbb{R}$, once with the additive monoid structure $(\mathbb{R}, +, 0)$ and once with the multiplicative monoid structure $(\mathbb{R}, \cdot, 1)$.

The category $\mathbf{Act}_M$ of monoid actions for a monoid M was defined in Example 1.12.1 (4). We elaborate the proof in detail.

Proposition 2.1.5 *The distribution functor $\mathcal{D}\colon \mathbf{Sets} \to \mathbf{Sets}$ can be lifted to a functor, $\mathbf{Act}_M \to \mathbf{Act}_M$, also written $\mathcal{D}$, in a commuting diagram:*

$$\begin{tikzcd} \mathbf{Act}_M \arrow[r, "\mathcal{D}"] \arrow[d] & \mathbf{Act}_M \arrow[d] \\ \mathbf{Sets} \arrow[r, "\mathcal{D}"] & \mathbf{Sets} \end{tikzcd} \tag{2.8}$$

The vertical arrow, on both sides, is the obvious forgetful functor that maps an action $M \times X \to X$ to the set X.

This 'lifting' also works for $\mathcal{D}$ as a monad, see Exercise 2.4.12.

Proof. We use the monoid $M = (M, +, 0)$ with additive operations. For a monoid action $\alpha\colon M \times X \to X$ and a monoid element $s \in M$ we write $\alpha_s\colon X \to X$ for the function $\alpha_s(x) := \alpha(s, x)$. Then, by the fact that α is a monoid action we get:

$$\alpha_0 = id \qquad \text{and} \qquad \alpha_{s+t} = \alpha_s \circ \alpha_t \tag{$*$}$$

We extend the action $\alpha\colon M \times X \to X$ on X to an action $\alpha_{\mathcal{D}}\colon M \times \mathcal{D}(X) \to \mathcal{D}(X)$ via functoriality of $\mathcal{D}$, namely as:

$$\alpha_{\mathcal{D}}(s, \omega) := \mathcal{D}(\alpha_s)(\omega) = \sum_{x \in supp(\omega)} \omega(x) \big| \alpha_s(x) \rangle = \sum_{x \in supp(\omega)} \omega(x) \big| \alpha(s, x) \rangle.$$

We have to prove that this $\alpha_{\mathcal{D}}$ is a monoid action, for which we use the above equations $(*)$, in combination with the fact that $\mathcal{D}$ is a functor and thus preserves identity maps and composition. First,

$$\alpha_{\mathcal{D}}(0, \omega) = \mathcal{D}(\alpha_0)(\omega) \overset{(*)}{=} \mathcal{D}(id)(\omega) = id(\omega) = \omega.$$

Next, along the same lines,

$$\begin{aligned} \alpha_{\mathcal{D}}(s + t, \omega) &= \mathcal{D}(\alpha_{s+t})(\omega) \overset{(*)}{=} \mathcal{D}(\alpha_s \circ \alpha_t)(\omega) = \mathcal{D}(\alpha_s)\big(\mathcal{D}(\alpha_t)(\omega)\big) \\ &= \mathcal{D}(\alpha_s)\big(\alpha_{\mathcal{D}}(t, \omega)\big) = \alpha_{\mathcal{D}}\big(s, \alpha_{\mathcal{D}}(t, \omega)\big). \end{aligned}$$

At this stage we can define a functor $\mathbf{Act}_M \to \mathbf{Act}_M$ on objects by:

$$\big(M \times X \xrightarrow{\alpha} X\big) \longmapsto \big(M \times \mathcal{D}(X) \xrightarrow{\alpha_{\mathcal{D}}} \mathcal{D}(X)\big).$$

On morphisms we will define it as $f \mapsto \mathcal{D}(f)$. What we then have to check is that if we have a map of actions f as on the left below, then $\mathcal{D}(f)$ is a map of actions on the right.

$$\begin{tikzcd} M \times X \arrow[r, "id \times f"] \arrow[d, "\alpha"'] & M \times Y \arrow[d, "\beta"] \\ X \arrow[r, "f"] & Y \end{tikzcd} \qquad \begin{tikzcd} M \times \mathcal{D}(X) \arrow[r, "id \times \mathcal{D}(f)"] \arrow[d, "\alpha_{\mathcal{D}}"'] & M \times \mathcal{D}(Y) \arrow[d, "\beta_{\mathcal{D}}"] \\ \mathcal{D}(X) \arrow[r, "\mathcal{D}(f)"] & \mathcal{D}(Y) \end{tikzcd}$$

Using the above notation α_s, and accordingly β_s, it is not hard to see that commutation of the above diagram on the left means, for each $s \in M$,

$$f \circ \alpha_s = \beta_s \circ f. \tag{$**$}$$

This can be used as follows, to get commutation of the rectange on the right, above.

$$\begin{aligned}(\beta_{\mathcal{D}} \circ (id \times \mathcal{D}(f)))(s,\omega) \;=\; \beta_{\mathcal{D}}(s, \mathcal{D}(f)(\omega)) \;&=\; \mathcal{D}(\beta_s)\big(\mathcal{D}(f)(\omega)\big) \\ &=\; \mathcal{D}(\beta_s \circ f)(\omega) \\ &\overset{(**)}{=}\; \mathcal{D}(f \circ \alpha_s)(\omega) \\ &=\; \mathcal{D}(f)\big(\mathcal{D}(\alpha_s)(\omega)\big) \\ &=\; \mathcal{D}(f)\big(\alpha_{\mathcal{D}}(s,\omega)\big) \;=\; (\mathcal{D}(f) \circ \alpha_{\mathcal{D}})(s,\omega).\end{aligned}$$

By construction, the diagram of functors (2.8) commutes. ❑

In the previous chapter we have seen that the sets $\mathcal{L}(X)$, $\mathcal{P}(X)$ and $\mathcal{M}(X)$ of lists, subsets and multisets all carry a monoid structure — corresponding, respectively, to $\mathcal{L}$-algebras, $\mathcal{P}_{fin}$-algebras, and $\mathcal{N}$-algebras. One may expect a similar result saying that $\mathcal{D}(X)$ forms a monoid too, via an elementwise sum, like for multisets. But that does *not* work. Instead of arbitrary sums, one can take *convex sums* of distributions. This works as follows. Suppose we have two distributions $\omega, \rho \in \mathcal{D}(X)$ and a number $s \in [0,1]$. Then we can form a new distribution $\sigma \in \mathcal{D}(X)$, as convex combination of ω and ρ, namely:

$$\sigma \;\coloneqq\; s \cdot \omega + (1-s) \cdot \rho \qquad \text{that is} \qquad \sigma(x) \;=\; r \cdot \omega(x) + (1-s) \cdot \rho(x). \tag{2.9}$$

This obviously generalises to an n-ary convex sum. Such a convex sum of distributions is often called a 'mixture', see Example 2.3.6.

In Exercise 2.1.18 we axiomatise structures with such convex sums; they are called 'barycentric algebras' or just 'convex sets', see Stone (1949) or Jacobs (2010) for details. They correspond to $\mathcal{D}$-algebras and allow us to characterise sets of distributions $\mathcal{D}(X)$ as free barycentric algebras on X. A brief historical account occurs in (Keimel and Plotkin, 2017, Remark 2.9).

2.1.1 Discrete distributions with infinite support

So far we have been using multisets and distributions with finite support only. It makes sense, for certain applications, to drop this finiteness requirement, especially for distributions.

Definition 2.1.6 For an arbitrary set X we can form a set $\mathcal{D}_\infty(X)$ of distributions with (possibly) infinite support as:

$$\mathcal{D}_\infty(X) \;\coloneqq\; \{\omega \colon X \to [0,1] \mid \textstyle\sum_{x \in X} \omega(x) = 1\}.$$

This operation $\mathcal{D}_\infty$ is functorial too, like $\mathcal{D}$: for a function $f \colon X \to Y$ there is a function $\mathcal{D}_\infty(f) \colon \mathcal{D}_\infty(X) \to \mathcal{D}_\infty(Y)$ given by $\mathcal{D}_\infty(f)(\sum_i r_i|x_i\rangle) = \sum_i r_i|f(x_i)\rangle$.

The equation $\sum_n r^n = \frac{1}{1-r}$ from Theorem 1.7.4 (2), for $r \in [0,1)$, can be used as source of examples, of the form:

$$\sum_{n \in \mathbb{N}} (1-r) \cdot r^n \,\big|n\big\rangle \;\in\; \mathcal{D}_\infty(\mathbb{N}).$$

For instance, $r = \frac{1}{4}$ gives an infinite distribution $\sum_n \frac{3}{4^{n+1}}|n\rangle = \frac{3}{4}|0\rangle + \frac{3}{16}|1\rangle + \frac{3}{64}|2\rangle + \cdots$.

The sum $\sum$ in the above formulation in Definition 2.1.6 is assumed to exist, as limit of finite sums. We show that it involves at most countably many non-zero probabilities.

Lemma 2.1.7 *The support of a distribution $\omega \in \mathcal{D}_\infty(X)$ is necessarily countable, so either finite or of the same cardinality as the natural numbers.*

Proof. For $\omega \in \mathcal{D}_\infty(X)$ and $n \in \mathbb{N}_{>0}$ we write:

$$supp_n(\omega) := \{x \in X \mid \omega(x) > \tfrac{1}{n}\}.$$

Since $\sum_{x \in X} \omega(x) = 1$ this set $supp_n(\omega)$ can have at most n elements. In particular, it is finite. We can write the support of the distribution $\omega \in \mathcal{D}_\infty(X)$ as countable union:

$$supp(\omega) = \{x \in X \mid \omega(x) > 0\} = \bigcup_{n \in \mathbb{N}} supp_n(\omega).$$

We see that $supp(\omega)$ is a countable union of finite sets; hence it is either finite or countably infinite. ❑

Below we describe several (standard) examples of infinite distributions, like Poisson and negative binomials. Exercise 2.1.15 contains another example, namely the geometric distribution. There are more distributions with infinite support, such as the zeta (or zipf) distribution, see *e.g.* Ross (2018).

Examples 2.1.8 1 A famous infinite distribution on $\mathbb{N}$ is the *Poisson* distribution $pois[\lambda]$ with 'mean', 'rate' or 'intensity' parameter $\lambda \in \mathbb{R}_{\geq 0}$. It can be described as infinite formal convex sum:

$$pois[\lambda] := \sum_{k \in \mathbb{N}} e^{-\lambda} \cdot \frac{\lambda^k}{k!} \left|k\right\rangle \in \mathcal{D}_\infty(\mathbb{N}). \tag{2.10}$$

The multiplicities add up to one in the Poisson distribution because of the well known formula describing an exponential via an infinite sum:

$$e^\lambda = \sum_{k \in \mathbb{N}} \frac{\lambda^k}{k!}. \tag{2.11}$$

The Poisson distribution is typically used for counts of rare events. The *rate* or *intensity* parameter λ is the average number of events per time period. The Poisson distribution then gives for each $k \in \mathbb{N}$ the probability of having k events per time period. This works when events occur independently. As a border case, we do allow a rate $\lambda = 0$. Then $pois[0] = 1|0\rangle$ since all the terms 0^n vanish, except for $n = 0$.

2 We can modify $pois[\lambda]$ from a distribution on $\mathbb{N}$ to a distribution $mpois[\lambda]$ on natural multisets $\mathcal{N}(X)$ over a finite set X with size $N := |X|$. This happens as follows. Let X have N elements.

$$mpois[\lambda] := \sum_{\varphi \in \mathcal{N}(X)} e^{-\lambda} \cdot \frac{\left(\frac{\lambda}{N}\right)^{\|\varphi\|}}{\varphi!} \left|\varphi\right\rangle \in \mathcal{D}(\mathcal{N}(X)). \tag{2.12}$$

Notice that this is the ordinary Poisson distribution when X is a singleton. This multiset Poisson distribution $mpois[\lambda]$ is a special case of a Poisson point process $Pmn[\lambda](unif_X)$, see Definition 3.10.1, for the uniform distribution on X.

The probabilities in $mpois[\lambda]$ add up to one since:

$$\begin{aligned}
\sum_{\varphi\in\mathcal{N}(X)} e^{-\lambda}\cdot\frac{\left(\frac{\lambda}{N}\right)^{\|\varphi\|}}{\varphi!} &= e^{-\lambda}\cdot\sum_{K\in\mathbb{N}}\sum_{\varphi\in\mathcal{N}[K](X)}\frac{\left(\frac{\lambda}{N}\right)^{\|\varphi\|}}{\varphi!} \\
&= e^{-\lambda}\cdot\sum_{K\in\mathbb{N}}\frac{\left(\frac{\lambda}{N}\right)^{K}}{K!}\sum_{\varphi\in\mathcal{N}[K](X)}\frac{K!}{\varphi!} \\
&= e^{-\lambda}\cdot\sum_{K\in\mathbb{N}}\frac{\lambda^K}{N^K\cdot K!}\cdot N^K \qquad \text{by Exercise 1.7.9} \\
&\overset{(2.11)}{=} 1.
\end{aligned}$$

This is a countable sum since $\mathcal{N}(X)\cong\mathbb{N}^N$ for $N=|X|$, see Exercise 1.6.5. The multiset Poisson and ordinary Poisson distributions are connected via the size function, see Exercise 2.1.16, and also in a so-called sufficient statistic situation, see Exercise 4.5.4.

3 Our next example of an infinite distribution is the *negative binomial* distribution, of the form $nbn[K](s)\in\mathcal{D}_\infty(\mathbb{N})$, for $K\geq 1$ and $s\in(0,1]$. It captures the probability of reaching K successes, with probability s, in $n+K$ trials. This can be formulated in several ways:

$$\begin{aligned}
nbn[K](s) &:= \sum_{n\in\mathbb{N}}\left(\!\!\binom{K}{n}\!\!\right)\cdot s^K\cdot(1-s)^n\,\big|K+n\rangle \\
&= \sum_{n\in\mathbb{N}}\binom{K+n-1}{K-1}\cdot s^K\cdot(1-s)^n\,\big|K+n\rangle \\
&= \sum_{m\geq K}\binom{m-1}{K-1}\cdot s^K\cdot(1-s)^{m-K}\,\big|m\rangle.
\end{aligned}$$

One can use the second equation in Theorem 1.7.4 (1) to show that this forms a distribution.

$$\sum_{n\in\mathbb{N}}\left(\!\!\binom{K}{n}\!\!\right)\cdot s^K\cdot(1-s)^n = s^K\cdot\frac{1}{(1-(1-s))^K} = 1.$$

A multivariate generalisation occurs in Exercise 2.6.13.

Exercises

2.1.1 Define for $n\in\mathbb{N}_{>0}$ the pyramid distribution $\Lambda_n\in\mathcal{D}(\mathbb{Z})$ as:

$$\begin{aligned}
\Lambda_n &:= \tfrac{1}{n}|0\rangle + \sum_{1\leq i<n}\tfrac{n-i}{n^2}|-i\rangle + \tfrac{n-i}{n^2}|i\rangle \\
&= \tfrac{1}{n^2}|-n+1\rangle + \tfrac{2}{n^2}|-n+2\rangle + \cdots + \tfrac{n-1}{n^2}|-1\rangle \\
&\qquad + \tfrac{1}{n}|0\rangle + \tfrac{n-1}{n^2}|1\rangle + \cdots + \tfrac{2}{n^2}|n-2\rangle + \tfrac{1}{n^2}|n-1\rangle.
\end{aligned}$$

1 Describe Λ_n for $n=1,2,3,4$ explicitly.

2 Check that Λ_n is indeed a distribution: its probabilities add up to one.
Hint: Use Proposition 1.2.6 (1).

3 Describe $\mathcal{D}(|-|)(\Lambda_n)\in\mathcal{D}(\mathbb{N})$ where $|-|\colon\mathbb{Z}\to\mathbb{N}$ is the absolute value function.

2.1.2 Describe the multiset coefficient distribution $mc[X,K]$ from (2.5) for $X=\{a,b\}$ and $K=4$.

2.1.3 Let X be a finite set with $n := |X|$ elements.

1 Show that the binomial coefficient distribution $bc[n] \in \mathcal{D}(\{0,\dots,n\})$ in Example 2.1.2 (3) can be obtained as image distribution via the size-of-subsets function $size\colon \mathcal{P}(X) \to \{0,\dots,n\}$, as in:

$$bc[n] \;=\; \mathcal{D}(size)\big(unif_{\mathcal{P}(X)}\big) \qquad \text{where} \qquad unif_{\mathcal{P}(X)} \;=\; \sum_{U \subseteq X} \tfrac{1}{2^n}\big|U\rangle.$$

2 For a number $K \in \mathbb{N}$ consider the accumulation function $acc\colon X^K \to \mathcal{N}[K](X)$ and show that the multiset coefficient distribution $mc[X,K] \in \mathcal{D}(\mathcal{N}[K](X))$ from (2.5) can be described as image distribution:

$$mc[X,K] \;=\; \mathcal{D}(acc)\big(unif_{X^K}\big) \qquad \text{where} \qquad unif_{X^K} \;=\; \sum_{\vec{x} \in X^K} \tfrac{1}{n^K}\big|\vec{x}\rangle.$$

2.1.4 Consider for numbers N, K the sum of multisets map $sum\colon \mathcal{N}[K](N) \to \{0,\dots,(N-1)\cdot K\}$ and prove the following relationship between the nomial distribution from (2.3) and the multiset coefficient distribution from (2.5).

$$\mathcal{D}(sum)\big(mc[N,K]\big) \;=\; nom[N,K].$$

2.1.5 Use Exercise 1.8.13 (2) to define yourself a distribution $mcc[n] \in \mathcal{D}(\{0,1,\dots,n\})$ of multichoose coefficients, in analogy with the distribution of binomial coefficients $bc[n]$ in (2.2). Describe $mcc[3]$ concretely.

2.1.6 Check that the distributions in Example 2.1.2 (3) and (4) are instance of the following construction. Suppose X is a non-empty finite set with a function $f\colon X \to \mathbb{R}_{\geq 0}$. Let $S := \sum_{x \in X} f(x)$ be the sum. Then one can form a distribution of the form.

$$\sum_{x \in X} \frac{f(x)}{S}\big|x\rangle \;\in\; \mathcal{D}(X).$$

Check for yourself what X and f are in the distributions in Example 2.1.2 (3) and (4).

2.1.7 1 Prove that $flip\colon [0,1] \to \mathcal{D}(\mathbf{2})$ is an isomorphism.

2 Check that $flip(r)$ is the same as $bn[1](r)$.

3 Describe the distribution $bn[3](\frac{1}{4})$ concretely and interpret this distribution in terms of coin flips.

2.1.8 Recall that $\boldsymbol{n} := \{0,\dots,n-1\}$ and check that:

$$\mathcal{D}(\mathbf{0}) \;\cong\; \mathbf{0} \qquad \mathcal{D}(\mathbf{1}) \;\cong\; \mathbf{1} \qquad \mathcal{D}(\mathbf{2}) \;\cong\; [0,1].$$

The set $\mathcal{D}(\boldsymbol{n+1})$ is often called the *n-simplex*, or the *probability simplex*. Describe it as a subset of $\mathbb{R}^{n+1}$, and also as a subset of $\mathbb{R}^n$.

2.1.9 Consider the binomial distribution $bn[K]$ from Example 2.1.2 (2). Show that for $r \in [0,1]$ and $k \in \{0,1,\dots,K\}$,

$$bn[K](r)(K-k) \;=\; bn[K](1-r)(k).$$

Check that this means that the rectangle

$$\begin{tikzcd}
{[0,1]} \arrow[r, "bn[K]"] \arrow[d, "minus(1,-)"', "\cong"] & \mathcal{D}(\{0,1,\dots,K\}) \arrow[d, "\mathcal{D}(minus(K,-))"] \\
{[0,1]} \arrow[r, "bn[K]"] & \mathcal{D}(\{0,1,\dots,K\})
\end{tikzcd}$$

commutes, where $minus(x, y) := x - y$.

2.1.10 Assume out of $K > 0$ trials we see $k \in \{0, 1, \ldots, K\}$ successes. Which success rate probability $r \in [0, 1]$ gives maximal binomial probability $bn[K](r)(k)$, that is, what is $\mathrm{argmax}_{r\in[0,1]}\, bn[K](r)(k)$?

1 What do you expect this probability to be?

2 Write $f(r) = \ln\big(bn[K](r)(k)\big)$, where ln is the natural logarithm. Check for yourself that it suffices to determine the maximum of f.

3 Check that the derivative of f is:

$$f'(r) = \frac{k}{r} - \frac{K-k}{1-r}.$$

4 Show that solving $f'(r) = 0$ gives $r = \frac{k}{K}$.

Example 3.4.3 (1) will generalise this result to multinomial distributions.

2.1.11 Let a number $r \in [0, 1]$ and a finite set X be given. Show that:

$$\sum_{U\in\mathcal{P}(X)} r^{|U|} \cdot (1-r)^{|X\setminus U|} = 1.$$

Hint: Recall the binomial distribution from Example 2.1.2 (2) and recall also Exercise 1.5.6.

2.1.12 Let $\omega \in \mathcal{D}(X)$ be a distribution, considered as a function $\omega\colon X \to [0, 1]$ with finite support. Use functoriality of $\mathcal{D}$ to show that:

$$\mathcal{D}(\omega)(\omega) = \sum_{x\in supp(\omega)} \omega(x)\big|\omega(x)\big\rangle \in \mathcal{D}([0, 1]).$$

2.1.13 Check that the powerbag operation from Exercise 1.8.3 can be turned into a probabilistic powerbag PPB via:

$$\mathsf{PPB}(\psi) := \sum_{\varphi\leq\psi} \frac{\binom{\psi}{\varphi}}{2^{\|\psi\|}} \big|\varphi\big\rangle.$$

2.1.14 1 Recall Theorem 1.7.4 (1) and conclude that $\lim_{n\to\infty} \binom{n+K}{K} \cdot r^n = 0$, for $r \in [0, 1)$. (This is general result: if partial sums of a series converge, the limit of the series itself is zero.)

2 Conclude that for $r \in (0, 1]$ one has:

$$\lim_{n\to\infty} bn[n+m](r)(m) = 0.$$

Explain yourself what this means.

2.1.15 For a non-zero probability $r \in (0, 1]$ one defines the *geometric distribution* $geo[r] \in \mathcal{D}_\infty(\mathbb{N}_{>0})$ as:

$$geo[r] := \sum_{k\in\mathbb{N}_{>0}} r \cdot (1-r)^{k-1}\big|k\big\rangle.$$

It captures the probability of being successful for the first time after $k - 1$ unsuccesful tries, with success probability r. Prove that this is a distribution indeed: its multiplicities add up to one. Notice that this is essentially the negative binomial distribution from Example 2.1.8 (3) for success number $K = 1$.

2.1.16 Recall the ordinary and multiset Poisson distributions (channels) *pois* and *mpois* from Example 2.1.8 (1) and (2). Let X be a finite set. Show that there is a size connection, as in the following diagram.

$$\begin{tikzcd} \mathcal{D}_\infty(\mathbb{N}) & & \mathcal{D}_\infty(\mathcal{N}(X)) \arrow[ll, "\mathcal{D}_\infty(\|-\|)"'] \\ & \mathbb{R}_{\geq 0} \arrow[ul, bend left, "pois"] \arrow[ur, bend right, "mpois"'] & \end{tikzcd}$$

There is another pushforward connection, see Exercise 2.4.23.

2.1.17 Let $\omega \in \mathcal{D}(X)$ be an arbitrary distribution on a set X. We extend it to a distribution $\omega^\star$ on the set $\mathcal{L}(X)$ of lists of elements from X. We define the function $\omega^\star \colon \mathcal{L}(X) \to [0,1]$ by:

$$\omega^\star([x_1, \ldots, x_n]) \coloneqq \frac{\omega(x_1) \cdot \ldots \cdot \omega(x_n)}{2^{n+1}}.$$

1 Prove that $\omega^\star \in \mathcal{D}_\infty(\mathcal{L}(X))$.

2 Consider the function $f \colon \{a,b,c\} \to \{1,2\}$ with $f(a) = 1$, $f(b) = 1$, $f(c) = 2$. Take $\omega = \frac{1}{3}|a\rangle + \frac{1}{4}|b\rangle + \frac{5}{12}|c\rangle \in \mathcal{D}(\{a,b,c\})$ and $\ell = [1,2,1] \in \mathcal{L}(\{1,2\})$. Verify both equations on the next line.

$$\mathcal{D}(f)(\omega)^\star(\ell) = \tfrac{245}{27648} = \mathcal{D}_\infty(\mathcal{L}(f))(\omega^\star)(\ell).$$

Using the $\star$-operation as a function:

$$\mathcal{D}(X) \xrightarrow{(-)^\star} \mathcal{D}_\infty(\mathcal{L}(X))$$

we can describe the above equation as:

$$\begin{tikzcd} \mathcal{D}(\{a,b,c\}) \ni \omega \arrow[r, mapsto] \arrow[dd, mapsto] & \omega^\star \in \mathcal{D}_\infty(\mathcal{L}(\{a,b,c\})) \arrow[d, mapsto] \\ & \mathcal{D}_\infty(\mathcal{L}(f))(\omega^\star) \arrow[d, equal] \\ \mathcal{D}(\{1,2\}) \ni \mathcal{D}(f)(\omega) \arrow[r, mapsto] & \mathcal{D}(f)(\omega)^\star \in \mathcal{D}_\infty(\mathcal{L}(\{1,2\})) \end{tikzcd}$$

3 Prove in general that $(-)^\star$ is a natural transformation from $\mathcal{D}$ to $\mathcal{D}_\infty \circ \mathcal{L}$.

2.1.18 According to Stone Stone (1949), a *barycentric* algebra is a set A together with an operation $+ \colon A \times [0,1] \times A \to A$, written as $a +_r b$, for $a, b \in A$ and $r \in [0,1]$, such that:

- $a +_r a = a$
- $a +_0 b = a$
- $a +_r b = b +_{1-r} a$
- $a +_r (b +_s c) = (a +_{r/t} b) +_t c$, for $t = r + (1-r)s$, when not both $r = 0$ and $s = 0$.

One thinks of $a +_r b$ as the convex sum $r \cdot a + (1-r) \cdot b$. Therefore, we often call a barycentric algebra simply a *convex set*.

A function $f \colon A \to B$ between two bary-centric algebras is called *affine* if $f(a +_r a') = f(a) +_r f(a')$ holds for all $a, a' \in A$ and $r' \in [0,1]$.

Show, in analogy with Propositions 1.4.6, 1.5.5 and 1.6.6 the following three points (see (Jacobs, 2010, Thm. 4) for details).

1 A set A carries an operation $+\colon A \times [0,1] \times A \to A$ making it a bary-centric algebra if and only if A carries a $\mathcal{D}$-algebra $\alpha\colon \mathcal{D}(A) \to A$, satisfying $\alpha \circ \mathit{unit} = \mathit{id}$ and $\alpha \circ \mathit{flat} = \alpha \circ \mathcal{D}(\alpha)$.
2 Given two barycentric algebras A, B, with corresponding algebras $\alpha\colon \mathcal{D}(A) \to A$ and $\beta\colon \mathcal{D}(B) \to B$, a function $f\colon A \to B$ is affine if and only if it is a homomorphism of algebras: $f \circ \alpha = \beta \circ \mathcal{D}(f)$.
We shall write **Conv** for the resulting category of convex sets and affine maps.
3 For an arbitrary set X, the flatten map makes the set $\mathcal{D}(X)$ the *free* barycentric algebra on X: for each barycentric algebra A with a function $f\colon X \to A$ there is a unique affine map $\overline{f}\colon \mathcal{D}(X) \to A$ with $\overline{f} \circ \mathit{unit} = f$.

2.2 Frequentist learning and stick breaking

We have introduced distributions as special multisets, namely as multisets in which the multiplicities add up to one, so that $\mathcal{D}(X) \subseteq \mathcal{M}(X)$. We will encounter various relations and interactions between distributions and multisets. In Section 2.6 we shall further exploit that an urn containing coloured balls is aptly described as a multiset and that the probability distribution associated with drawing balls from the urn can then be derived from the multiset, using the proportions of the different colours. In this section we first describe this situation in more abstract terms, via a mapping from multisets to distributions, called 'frequentist learning'. In essence, it involves counting and normalisation. Later on in this book we shall see how more general forms of learning from data can be described in terms of such passages from multisets to distributions. This forms a central topic.

In the other direction, going from distributions to multisets, may be done via sampling. This is the second topic of the current section. When a probability distribution is sampled, multiple items are 'drawn' from its support, in accordance with the probabilities of the elements in the distribution. In the limit, frequentist learning from the samples must reproduce the original distribution, as explained in Subsection 2.2.1.

The third topic of this section is stick breaking. This is an alternative method to produce distributions, not from a multisets, but from a finite list of numbers between 0 and 1 (not necessarily adding up to one). Stick breaking also involves normalisation, not of all numbers at the same time, but in a successive manner. Stick breaking is a useful technique that pops up occasionally. Later on we show how it can be used to express multinomial probabilities in terms of successive binomial ones, see Proposition 2.6.9.

In the previous chapter we have seen several (natural) mappings between collection types, in the form of support and accumulation maps, see the summary in Diagram (1.37). We now add the mapping $\mathit{Flrn}\colon \mathcal{M}_*(X) \to \mathcal{D}(X)$, from (non-empty) multisets to distributions. The name Flrn stands for 'frequentist learning', and may be pronounced as 'eff-learn'. The frequentist interpretation of probability theory views probabilities as long term frequencies of occurrences. Here, these occurrences are given via multisets, which form the inputs of the Flrn function. Later on, in Theorem 4.8.11 we show that these outcomes of frequentist learning lie dense in the set of all distributions (over a fixed finite set). It means that we can approximate each distribution with arbitrary precision via frequentist learning of (natural) multisets.

Recall that $\mathcal{M}_*(X)$ is the collection of *non-empty* multisets $\sum_i r_i|x_i\rangle$, with $r_i \neq 0$ for at least one index i. Equivalently one can require that the size $s := \sum_i r_i = \|\sum_i r_i|x_i\rangle\|$ is non-zero.

The *Flrn* maps turns a (non-empty) multiset into a distribution, essentially by normalisation. It is defined as follows.

$$Flrn\big(r_1|x_1\rangle + \cdots + r_k|x_k\rangle\big) := \frac{r_1}{s}|x_1\rangle + \cdots + \frac{r_k}{s}|x_k\rangle, \qquad \text{where } s := \textstyle\sum_i r_i. \tag{2.13}$$

The normalisation step forces the formal sum on the right-hand side to be a convex sum, with factors adding up to one. Clearly, we can learn distributions only from non-empty multisets, since for an empty multiset the above size s is zero so that we cannot divide by s.

Using scalar multiplication from Lemma 1.6.3 (2) we can define the *Flrn* function more succintly via its size as:

$$Flrn(\varphi) := \frac{1}{\|\varphi\|} \cdot \varphi \qquad \text{where} \qquad \|\varphi\| := \textstyle\sum_x \varphi(x). \tag{2.14}$$

We use frequentist learning for arbitrary multisets, and not just for the natural ones, with natural numbers as multiplicities.

Remark 2.2.1 As noted before, (natural) multisets are mathematical representations of urns filled with finitely many coloured balls. Frequentist learning *Flrn* forms a further, associated formalisation step: for a non-empty urn $\upsilon \in \mathcal{N}_*(X)$ over a set of colours X, the probability of drawing a ball of colour $x \in X$ from υ is given by the fraction of balls of colour X and the total number of balls in the urn:

$$Flrn(\upsilon)(x) = \frac{\upsilon(x)}{\|\upsilon\|} \in [0, 1].$$

This formalises the idea that was presented pictorially in (0.1) in the Preface.

Example 2.2.2 We present two illustrations of frequentist learning.

1 Suppose we have some coin of which the bias is unkown. Experimental data show that out of 50 tossings, head (H) comes up 20 times and tail (T) 30 times. We can represent this data as a *multiset* $\varphi = 20|H\rangle + 30|T\rangle \in \mathcal{M}_*(\{H, T\})$. When we wish to learn the resulting probabilities, we apply the frequentist learning map *Flrn* and get a *distribution* in $\mathcal{D}(\{H, T\})$, namely:

$$Flrn(\varphi) = \tfrac{20}{20+30}|H\rangle + \tfrac{30}{20+30}|T\rangle = \tfrac{2}{5}|H\rangle + \tfrac{3}{5}|T\rangle.$$

Thus, the bias (towards head) is $\frac{2}{5}$. In this simple case we could have obtained this bias immediately from the data, but the *Flrn* map captures the general mechanism.

Notice that with frequentist learning, more (or less) consistent data gives the same outcome. For instance if we knew that 40 out of 100 tosses were head, or 2 out of 5, we would still get the same bias. Intuitively, more (or less) data should give more (or less) confidence about the distribution. However, these aspects are not covered by frequentist learning, see Equation (2.15) below. A more sophisticated form of 'Bayesian' learning will be used for this later. It has the additional advantage that it can handle prior knowledge, if any, about the bias.

2 Recall the medical table (1.31) captured by the multiset $\tau \in \mathcal{N}(B \times M)$. Learning from τ yields the following joint distribution:

$$Flrn(\tau) = \tfrac{10}{100}|H, 0\rangle + \tfrac{35}{100}|H, 1\rangle + \tfrac{25}{100}|H, 2\rangle + \tfrac{5}{100}|L, 0\rangle + \tfrac{10}{100}|L, 1\rangle + \tfrac{15}{100}|L, 2\rangle.$$

Such a distribution, directly derived from a table, is sometimes called an *empirical* distribution Darwiche (2009).

In the above coin example we saw a property that is typical of frequentist learning, namely that learning from more of the same does not have any effect. We can make this precise via the equation:

$$Flrn(s \cdot \varphi) \;=\; Flrn(\varphi), \qquad \text{for } s \in \mathbb{R}_{>0}. \tag{2.15}$$

In Gigerenzer and Hoffrage (1995) it is argued that in general, people are not very good at probabilistic (esp. Bayesian) reasoning, but that they are much better at reasoning with "frequency formats". Simply put: the information (from $\mathcal{D}$) that there is a 0.04 probability of getting a disease is more difficult to process than the information (from $\mathcal{M}$) that 4 out of 100 people get the disease. In the current setting these frequency formats would correspond to natural multisets; they can be turned into distributions via the frequentist learning map $Flrn$.

Example 2.2.3 The following is part of a formulation of Simpson's paradox in (Rice, 2006, §1.4).

A black urn contains 5 red and 6 green balls, and a white urn contains 3 red and 4 green balls. You are allowed to choose an urn and then choose a ball at random from the urn. If you choose a red ball, you get a prize. Which urn should you choose to draw from? ... Now consider another game in which a second black urn has 6 red and 3 green balls, and a second white urn has 9 red and 5 green balls. ... In the final game, the contents of the second black urn are added to the first black urn, and the contents of the second white urn are added to the first white urn. Again, you can choose which urn to draw from.

The situation is captured by two channels $c_1, c_2 \colon \{B, W\} \to \mathcal{N}(\{R, G\})$ for (natural) multisets $\mathcal{N}$, namely:

$$\begin{cases} c_1(B) \;=\; 5|R\rangle + 6|G\rangle \\ c_1(W) \;=\; 3|R\rangle + 4|G\rangle \end{cases} \qquad\qquad \begin{cases} c_2(B) \;=\; 6|R\rangle + 3|G\rangle \\ c_2(W) \;=\; 9|R\rangle + 5|G\rangle \end{cases}$$

The probabilities of drawing a red or green ball are determined by the two probabilistic channels $Flrn \circ c_i \colon \{B, W\} \to \mathcal{D}(\{R, G\})$ obtained via frequentist learning. In both games one chooses the black urn since it has the highest probalities for red, namely $(Flrn \circ c_1)(B)(R) = \frac{5}{11} > \frac{3}{7} = (Flrn \circ c_1)(W)(R)$ and $(Flrn \circ c_2)(B)(R) = \frac{6}{9} = \frac{2}{3} > \frac{9}{14} = (Flrn \circ c_2)(W)(R)$.

The adding of the contents of the urns is represented mathematically via the sum of multisets. We can do this pointwise for channels, giving $c \coloneqq c_1 + c_2$, so that $c(x) = c_1(x) + c_2(x)$. Then:

$$\begin{cases} (Flrn \circ c)(B) \;=\; Flrn\big(11|R\rangle + 9|G\rangle\big) \;=\; \frac{11}{20}|R\rangle + \frac{9}{20}|G\rangle \\ (Flrn \circ c)(W) \;=\; Flrn\big(12|R\rangle + 9|G\rangle\big) \;=\; \frac{4}{7}|R\rangle + \frac{3}{7}|G\rangle. \end{cases}$$

We now have $\frac{11}{20} < \frac{4}{7}$ so that after combining the urns, choosing the white one is wiser, since it has a higher probability of choosing red.

Is this a paradox? The change of choices may be a bit unexpected at first, but not on closer inspection. Frequentist learning $Flrn$ maps sums of multisets to *convex* sums of distributions, as in (2.9). We can make this precise in the following commuting diagram, for $K, L \in \mathbb{N}_{>0}$.

$$\begin{tikzcd} \mathcal{N}[K](X) \times \mathcal{N}[L](X) \arrow[r, "+"] \arrow[d, "Flrn \times Flrn"'] & \mathcal{N}[K+L](X) \arrow[d, "Flrn"] \\ \mathcal{D}(X) \times \mathcal{D}(X) \arrow[r, "\frac{K}{K+L}(-) + \frac{L}{K+L}(-)"] & \mathcal{D}(X) \end{tikzcd} \tag{2.16}$$

In the example, the black and white urns have different sizes in the two games. The resulting different fractions lead to differently rescaled probabilities. The lesson is that in frequentist learning

of multiple multisets, probabilities do not simply add up but are rescaled — giving possibly unexpected outcomes.

We review this example in Exercise 9.1.2 in terms of 'adjustment'.

The following elementary observation captures the essence of frequentist learning.

Lemma 2.2.4 *For each set X the tuple of size and frequentist learning forms an isomorphism in:*

$$\mathcal{M}_*(X) \xrightarrow[\cong]{\langle size, Flrn\rangle} \mathbb{R}_{>0} \times \mathcal{D}(X)$$

Proof. The inverse is scaling: it sends a positive number $s \in \mathbb{R}_{>0}$ and a distribution $\omega \in \mathcal{D}(X)$ to the multiset $s \cdot \omega = \sum_x s \cdot \omega(x)\big|x\rangle$. ❑

It turns out that the learning map *Flrn* is 'natural', in the sense that it works uniformly for each set.

Lemma 2.2.5 *The frequentist learning maps $\mathrm{Flrn}\colon \mathcal{M}_*(X) \to \mathcal{D}(X)$ from (2.13) are natural in X. This means that for each function $f\colon X \to Y$ the following diagram commutes.*

$$\begin{tikzcd} \mathcal{M}_*(X) \arrow[r, "Flrn"] \arrow[d, "\mathcal{M}(f)"'] & \mathcal{D}(X) \arrow[d, "\mathcal{D}(f)"] \\ \mathcal{M}_*(Y) \arrow[r, "Flrn"'] & \mathcal{D}(Y) \end{tikzcd}$$

As a special case, frequentist learning commutes with marginalisation, via projection functions.

Proof. Pick an arbitrary non-empty multiset $\varphi = \sum_i r_i|x_i\rangle$ in $\mathcal{M}_*(X)$ and write $s := \|\varphi\| = \sum_i r_i$. By non-emptyness of φ we have $s \neq 0$. Then:

$$\begin{aligned}(Flrn \circ \mathcal{M}_*(f))(\varphi) &= Flrn(\textstyle\sum_i r_i|f(x_i)\rangle) = \textstyle\sum_i \frac{r_i}{s}|f(x_i)\rangle) \\ &= \mathcal{D}(f)(\textstyle\sum_i \frac{r_i}{s}|x_i\rangle) = (\mathcal{D}(f) \circ Flrn)(\varphi).\end{aligned}$$ ❑

We can apply this basic result to the medical data in Table (1.31), via the multiset $\tau \in \mathcal{N}(B \times M)$. We have already seen in Section 1.6 that the multiset-marginals $\mathcal{N}(\pi_i)(\tau)$ produce the marginal columns and rows, with their totals. We can learn the distributions from the colums as:

$$Flrn\Big(\mathcal{M}(\pi_1)(\tau)\Big) = Flrn\Big(70|H\rangle + 30|L\rangle\Big) = \tfrac{7}{10}|H\rangle + \tfrac{3}{10}|L\rangle.$$

We can also take the distribution-marginal of the 'learned' distribution from the table, as described in Example 2.2.2 (2):

$$\mathcal{M}(\pi_1)\Big(Flrn(\tau)\Big) = \Big(\tfrac{10}{100} + \tfrac{35}{100} + \tfrac{25}{100}\Big)|H\rangle + \Big(\tfrac{5}{100} + \tfrac{10}{100} + \tfrac{15}{100}\Big)|L\rangle = \tfrac{7}{10}|H\rangle + \tfrac{3}{10}|L\rangle.$$

Thus, frequentist learning and marginalisation commute. This is a simple result, which many practitioners in probability are surely aware of, at an intuitive level, but maybe not in the mathematically precise form of Lemma 2.2.5.

2.2.1 Sampling

Sampling is a technique for choosing individual elements from the support $supp(\omega) \subseteq X$ of a distribution $\omega \in \mathcal{D}(X)$, in accordance with the probabilities in ω. Sampling is important in computing with probability distributions, especially when these distributions become too large to handle. Sampling exists in many programming languages with some level of support for probability. For instance, in Python, the `numpy.random` package allows to write a command (after the prompt `>>>` below) for sampling from a multinomial (*i.e.* discrete) distribution:

```
>>> multinomial(10, [1/2, 1/3, 1/6])
array([5, 4, 1])
```

The distribution involved has three probabilities $1/2$, $1/3$, $1/6$, corresponding to a distribution in ket form $\frac{1}{2}|0\rangle + \frac{1}{3}|1\rangle + \frac{1}{6}|2\rangle$, say on the three-element set $\mathbf{3} = \{0, 1, 2\}$. The above number 10 specifies that we wish to get 10 individual samples from this distribution. These 10 samples are collected in the array $[5, 4, 1]$, given as output on the above second line. In the notation of this book this array forms a multiset $5|0\rangle + 4|1\rangle + 1|2\rangle$. These numbers correspond roughly to the original distribution from which we sample, after frequentist learning: $\mathit{Flrn}(5|0\rangle + 4|1\rangle + 1|2\rangle) = \frac{1}{2}|0\rangle + \frac{2}{5}|1\rangle + \frac{1}{5}|2\rangle$.

Running the same command again may give a different outcome, as in:

```
>>> multinomial(10, [1/2, 1/3, 1/6])
array([4, 3, 3])
```

Thus, sampling is not a deterministic operation. Hence it is mathematically awkward.

If we increase the number of samples to say 1000 we may get:

```
>>> multinomial(10000, [1/2, 1/3, 1/6])
array([509, 331, 160])
```

We now obtain a better approximation of the original distribution: $\mathit{Flrn}(509|0\rangle + 331|1\rangle + 160|2\rangle) = \frac{509}{1000}|0\rangle + \frac{331}{1000}|1\rangle + \frac{160}{1000}|2\rangle \approx \frac{1}{2}|0\rangle + \frac{1}{3}|1\rangle + \frac{1}{6}|2\rangle$.

The key idea is that, as the number N of samples taken from a distribution ω increases, the resulting multiset φ_N of samples, of size N, approaches ω via frequentist learning: $\mathit{Flrn}(\varphi_N) \to \omega$ as $N \to \infty$.

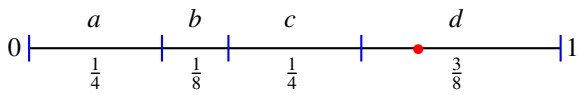
Thus, informally, sampling may be seen as an approximate inverse to frequentist learning. It exists in many probabilistic programming languages, see *e.g.* Gordon et al. (2014) for an overview. Sampling a single element x from a distribution $\omega \in \mathcal{D}(X)$ may be written as:

$$x \leftarrow \omega \qquad \text{or as} \qquad x \sim \omega. \tag{2.17}$$

Whenever we use such sampling, we shall write it in this first form $x \leftarrow \omega$.

How does this sampling from a discrete distribution work? We assume that there is some way to obtain an arbitrary number $r \in [0, 1]$ from the unit interval. Any programming language offers such functionality — in a pseudo-random manner. We sketch how to use it to obtain sampling from discrete distributions. Let's take as concrete example $\omega = \frac{1}{4}|a\rangle + \frac{1}{8}|b\rangle + \frac{1}{4}|c\rangle + \frac{3}{8}|d\rangle$. We order the elements in the support, here in the obvious way, as a, b, c, d and we break up the unit

interval in line fragments with sizes given by the respective probabilities, in order. In this case we get:



Now one asks for a random number $r \in [0, 1]$, say $r = 0.72$. One then looks up in which interval this number lands, see the above red dot. The corresponding element from the support of the distribution, in this case d, is returned as sample result. This approach obviously generalises to an arbitrary finite distribution $\sum_i r_i|x_i\rangle$, where $\sum_i r_i = 1$.

Computation via sampling is a very powerful technique, especially in situations where the distribution and/or the problem at hand is complicated and does not lend itself easily to an analytical solution. Consider for instance the following challenge. If we pick three random numbers i, j, k from the set of numbers $\{1, 2, \ldots, 100\}$, what is the probability that $i \leq j \leq k$? It is easy to write a probabilistic program for this. Let $unif_{100}$ be the uniform distribution on $\{1, 2, \ldots, 100\}$.

```
i ← unif100
j ← unif100
k ← unif100
if i <= i and j <= k:
   return yes
else:
   return no
```
(2.18)

This program yields as distribution, approximately, $0.17|yes\rangle + 0.83|no\rangle$. The precise outcome appears in Exercise 2.2.6 below.

This book is not about such probabilistic programming: it has a more mathematical ('exact') focus. From this mathematical perspective, the multinomial distribution $mn[K](\omega) \in \mathcal{D}(\mathcal{N}[K](X))$, that we introduce later on, can be understood as a 'correct' distribution over multiset samples, see Theorem 3.3.4 and also Theorem 5.6.1, and the subsequent discussions.

2.2.2 Stick breaking

Frequentist learning is a method for obtaining a distribution from a sequence of multiplicities (of a multiset). Stick breaking is an alternative method for producing a distribution, from a sequence of probabilities. It does not work via a single normalisation, but via iterated normalisations. It is a basic technique (see *e.g.* Sethuraman (1994) or (Joyce, 1998, Defn. 1)) that will be introduced below and will be used occasionally in the sequel. We follow the description of Jacobs (2022d) and start with some basic observations.

We write $\mathcal{D}_{fs}(\{R, G, B, Y\})$ for the set of all distributions with full support on the set $\{R, G, B, Y\}$ of four colours (Red, Green, Blue, Yellow). Explicitly:

$$\begin{aligned}&\mathcal{D}_{fs}(\{R, G, B, Y\})\\ &= \Big\{ r_0|R\rangle + r_1|G\rangle + r_2|B\rangle + r_3|Y\rangle \;\Big|\; r_0, r_1, r_2, r_3 \in (0, 1) \text{ with } r_0 + r_1 + r_2 + r_3 = 1 \Big\}.\end{aligned}$$

Because fullness of support is required, none of the r_i may be zero or one. It is needed below to prevent division by zero.

The above equation describes the set of distributions (on these four colours) as a *simplex*, of dimension *three*. Indeed, it is easy to see that one of the r_i is superfluous, since it is determined by the others. Explicitly, there is an isomorphism:

$$\mathcal{D}_{\mathrm{fs}}(\{R,G,B,Y\}) \cong \Big\{(r_0,r_1,r_2) \in (0,1)^3 \,\Big|\, r_0 + r_1 + r_2 < 1\Big\}.$$

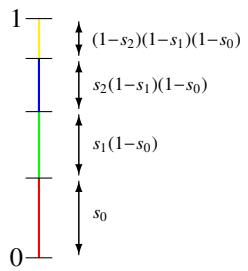
The above set on the right-hand-side is clearly a proper subset of the cube $(0,1)^3$. In essence, the stick breaking construction that we are about to describe yields an isomorphism:

$$\mathcal{D}_{\mathrm{fs}}(\{R,G,B,Y\}) \cong (0,1)^3. \tag{2.19}$$

This may not be immediate at first sight. One has to do (appropriate) rescaling.

There is an intuitive explanation of stick breaking in terms of successively breaking up a stick. We adapt this account to the above set of four colours. We start from three numbers $s_0, s_1, s_2 \in (0,1)$ and intend to turn them into a distribution on the set of colour $\{R,G,B,Y\}$.

Imagine a stick of length one, as described vertically on the right. We take our first number $s_0 \in (0,1)$ and decide to paint the lower part / proportion s_0 red. We now have an unpainted part of length $1-s_0$. We paint the s_1 proportion of it green. The newly painted part then has length $s_1(1-s_0)$. The unpainted part is now $(1-s_2)(1-s_0)$. We paint the s_2-proportion of this remainder blue. The final remainder is then of length $(1-s_2)(1-s_1)(1-s_0)$. We paint it yellow. Note that the resulting distribution has full support.



This construction can also be described in terms of breaking a stick, at each position where we have a change of colour in the above picture. The effect is a map $(0,1)^3 \to \mathcal{D}_{\mathrm{fs}}(\{R,G,B,Y\})$.

We turn to the general description.

Theorem 2.2.6 *For a number $r \in [0,1]$ we write $r^\perp = 1-r$.*

1 *For each number $N > 1$, with associated N-element set $\boldsymbol{N} = \{0,1,\ldots,N-1\}$, there is a "stick breaking" isomorphism:*

$$\begin{aligned} (0,1)^{N-1} &\xrightarrow[\cong]{\mathit{stbr}} \mathcal{D}_{\mathrm{fs}}(\boldsymbol{N}) \\ (r_0,\ldots,r_{N-2}) &\longmapsto r_0\big|0\rangle + r_0^\perp r_1\big|1\rangle + r_0^\perp r_1^\perp r_2\big|2\rangle + \cdots \\ &\qquad + r_0^\perp \cdots r_{N-3}^\perp r_{N-2}\big|N-2\rangle + r_0^\perp \cdots r_{N-2}^\perp\big|N-1\rangle. \end{aligned} \tag{2.20}$$

2 *This isomorphism extends to countably infinite sequences, as:*

$$\begin{aligned} (0,1)^{\mathbb{N}} &\xrightarrow[\cong]{\mathit{stbr}} \mathcal{D}_{\infty,\mathrm{fs}}(\mathbb{N}) \\ \vec{r} &\longmapsto \sum_{n\in\mathbb{N}} \big(\textstyle\prod_{i<n} r_i^\perp\big) r_n \,\big|n\rangle, \end{aligned} \tag{2.21}$$

where $\mathcal{D}_{\infty,\mathrm{fs}}$ is used for distributions with full support, without finiteness restriction, see Definition 2.1.6.

After the proof, Example 2.2.7 elaborates these descriptions in concrete form. Notice that the numbers 0, 1 are excluded in the domain type of the stick breaking function. The number $r = 1$ cannot be used because of division by $r^\perp = 1-r$. The number $r = 0$ is subsequently exclused in order to get a symmetric formulation, corresponding to having distributions with full support.

Proof. 1 It is easy to see that the finite stick break definition in (2.20) satisfies for $1 \leq i < N-1$,

$$1 - \sum_{j\leq i} stbr(\vec{r})(j) = \prod_{j\leq i} r_j^\perp \tag{2.22}$$

This gives $\sum_{i<N} stbr(\vec{r})(i) = 1$, so that stickbreaking produces a probability distribution, on $\boldsymbol{N}$.

In the reverse direction we start from a distribution $\omega \in \mathcal{D}_{fs}(\boldsymbol{N})$ and define a sequence of $N-1$ numbers in $(0,1)$ by:

$$stbr^{-1}(\omega)_0 := \omega(0) \quad \text{and} \quad stbr^{-1}(\omega)_{i+1} := \frac{\omega(i+1)}{1 - \sum_{j\leq i}\omega(j)}$$

This inverse operation satisfies, for $0 \leq i < N-2$,

$$\prod_{j\leq i} (stbr^{-1}(\omega)_j)^\perp = 1 - \sum_{j\leq i} \omega(j). \tag{2.23}$$

These operations are each other's inverses, since:

$$\begin{aligned} stbr^{-1}\Big(stbr(\vec{r})\Big)_0 &= stbr(\vec{r})(0) = r_0 \\ stbr^{-1}\Big(stbr(\vec{r})\Big)_{i+1} &= \frac{stbr(\vec{r})(i+1)}{1 - \prod_{j\leq i} stbr(\vec{r})(j)} \overset{(2.22)}{=} \frac{(\prod_{j\leq i} r_j^\perp) r_{i+1}}{\prod_{j\leq i} r_j^\perp} = r_{i+1}. \end{aligned}$$

And, the other way around, the case $i = 0$ is simple again. And for $0 < i < N-1$,

$$\begin{aligned} stbr\Big(stbr^{-1}(\omega)\Big)(i+1) &= \left(\prod_{j\leq i} (stbr^{-1}(\omega)_j)^\perp\right) stbr^{-1}(\omega)_{i+1} \\ &\overset{(2.23)}{=} \left(1 - \sum_{j\leq i}\omega(j)\right) \frac{\omega(i+1)}{1 - \sum_{j\leq i}\omega(j)} = \omega(i+1). \end{aligned}$$

2 For an infinite sequence $\vec{r} \in (0,1)^{\mathbb{N}}$ definition (2.21) yields a probability distribution since:

$$\begin{aligned} 1 - \sum_{n\in\mathbb{N}} stbr(\vec{r})(n) = 1 - \lim_{N\to\infty} \sum_{n\leq N} (\textstyle\prod_{i<n} r_i^\perp) r_n &= \lim_{N\to\infty} 1 - \sum_{n\leq N} stbr(r_0, \ldots, r_n)(n) \\ &\overset{(2.22)}{=} \lim_{N\to\infty} \prod_{n\leq N} r_n^\perp = 0. \end{aligned}$$

The latter holds since $r_n^\perp = 1 - r_n < 1$ for each n.

In the other direction we can define, for $\omega \in \mathcal{D}_{\infty,fs}(X)$,

$$stbr^{-1}(\omega)_i := \frac{\omega(i)}{1 - \sum_{j<i}\omega(j)}.$$

These two definitions are pointwise each other's inverses, as we have seen in the previous item. ❑

Example 2.2.7 1 Here is a simple illustrations of the stick break operation, in the finite case, for $N = 4$.

$$\begin{aligned} stbr\big(\tfrac{3}{4}, \tfrac{2}{3}, \tfrac{1}{5}\big) &= \tfrac{3}{4}\big|0\rangle + \tfrac{1}{4}\cdot\tfrac{2}{3}\big|1\rangle + \tfrac{1}{4}\cdot\tfrac{1}{3}\cdot\tfrac{1}{5}\big|2\rangle + \tfrac{1}{4}\cdot\tfrac{1}{3}\cdot\tfrac{4}{5}\big|3\rangle \\ &= \tfrac{3}{4}\big|0\rangle + \tfrac{1}{6}\big|1\rangle + \tfrac{1}{60}\big|2\rangle + \tfrac{1}{15}\big|3\rangle. \end{aligned}$$

In the other direction we get, as expected, the original sequence:

$$\begin{aligned} stbr^{-1}\big(\tfrac{3}{4}\big|0\rangle + \tfrac{1}{6}\big|1\rangle + \tfrac{1}{60}\big|2\rangle + \tfrac{1}{15}\big|3\rangle\big) &= \left(\tfrac{3}{4}, \frac{1/6}{1 - 3/4}, \frac{1/60}{1 - 3/4 - 1/6}\right) \\ &= \left(\tfrac{3}{4}, \frac{1/6}{1/4}, \frac{1/60}{1/12}\right) = \big(\tfrac{3}{4}, \tfrac{2}{3}, \tfrac{1}{5}\big). \end{aligned}$$

2 Consider the infinite distribution:

$$\omega = \sum_{n\in\mathbb{N}} \tfrac{2}{5}\cdot(\tfrac{3}{5})^n\,\big|n\rangle = \tfrac{2}{5}\big|0\rangle + \tfrac{6}{25}\big|1\rangle + \tfrac{18}{125}\big|2\rangle + \tfrac{54}{625}\big|3\rangle + \cdots$$

We can see that it is a distribution via Theorem 1.7.4 (2):

$$\sum_{n\in\mathbb{N}} \omega(n) = \tfrac{2}{5}\cdot\sum_{n\in\mathbb{N}}(\tfrac{3}{5})^n = \tfrac{2}{5}\cdot\frac{1}{1 - 3/5} = \frac{2}{5-3} = 1.$$

The sequence of numbers in $(0, 1)$ corresponding to ω is constant:

$$stbr^{-1}(\omega) = (\tfrac{2}{5}, \tfrac{2}{5}, \tfrac{2}{5}, \ldots).$$

In general, for $r \in (0, 1)$, we have $stbr(r, r, r, \ldots) = \sum_{n\in\mathbb{N}} r(1-r)^n\big|n\rangle$.

Exercises

2.2.1 Recall the data / multisets about child ages and blood types in the beginning of Subsection 1.6.1. Compute the associated (empirical) distributions.

Plot these distributions as a graph. How do they compare to the plots (1.29) and (1.30)?

2.2.2 Check that frequentist learning from a constant multiset yields a uniform distribution. And also that frequentist learning is invariant under (non-zero) scalar multiplication, as described in (2.15).

2.2.3 Prove explicitly, as described in Diagram (2.16), that for natural multisets $\varphi, \psi \in \mathcal{M}_*(X)$ one has:

$$Flrn(\varphi + \psi) = \frac{\|\varphi\|}{\|\varphi\| + \|\psi\|}\cdot Flrn(\varphi) + \frac{\|\psi\|}{\|\varphi\| + \|\psi\|}\cdot Flrn(\psi).$$

This means that when one has already learned $Flrn(\varphi)$ and new data ψ arrives, all probabilities have to be adjusted, as in the above convex sum of distributions.

2.2.4 Show that Diagram (1.37) can be refined to:

$$\begin{tikzcd} \mathcal{L}_*(X) \arrow[rr, "supp"] \arrow[dr, bend right, "acc"'] & & \mathcal{P}_{fin}(X) \\ & \mathcal{M}_*(X) \arrow[r, "Flrn"'] & \mathcal{D}(X) \arrow[u, bend right, "supp"'] \end{tikzcd} \tag{2.24}$$

where $\mathcal{L}_*(X) \subseteq \mathcal{L}(X)$ is the subset of non-empty lists.

2.2.5 Let $\upsilon \in \mathcal{N}[L](X)$ be a natural multiset. For the first point below there is a size requirement $L \geq 2$, but for the second point we only need $L \geq 1$.

1 Define the distribution $\tau \in \mathcal{D}(X \times X)$ for two consecutive draws without replacement from υ, as urn:

$$\tau \;=\; \sum_{x \in supp(\upsilon)} \; \sum_{y \in supp(\upsilon - 1|x\rangle)} Flrn(\upsilon)(x) \cdot Flrn(\upsilon - 1|x\rangle)(y) \,\big|x, y\rangle.$$

Show that both marginals of τ are equal to $Flrn(\upsilon)$, that is:

$$\mathcal{D}(\pi_1)(\tau) \;=\; Flrn(\upsilon) \qquad \text{and} \qquad \mathcal{D}(\pi_2)(\tau) \;=\; Flrn(\upsilon).$$

2 Now define $\sigma \in \mathcal{D}(X \times X)$ for two consecutive draws with addition:

$$\sigma \;=\; \sum_{x \in supp(\upsilon)} \; \sum_{y \in supp(\upsilon + 1|x\rangle)} Flrn(\upsilon)(x) \cdot Flrn(\upsilon + 1|x\rangle)(y) \,\big|x, y\rangle.$$

Show that both marginals of σ are also equal to $Flrn(\upsilon)$.

2.2.6 Consider the program in (2.18) and show analytically that the outcome distribution is $\frac{1717}{10000}\big|yes\rangle + \frac{8283}{10000}\big|no\rangle$.

2.2.7 Consider the distribution $\omega = \frac{1}{6}|0\rangle + \frac{1}{5}|1\rangle + \frac{1}{4}|2\rangle + \frac{1}{3}|3\rangle + \frac{1}{20}|4\rangle \in \mathcal{D}_{fs}(\mathbf{5})$.

1 Calculate $\vec{x} := stbr^{-1}(\omega) \in (0, 1)^4$.

2 Check what $stbr(\vec{x})$ is.

2.2.8 Fix $N > 1$.

1 Let $\upsilon = unif_N = \sum_{i \in N} \frac{1}{N}|i\rangle$. Show that:

$$stbr^{-1}(\upsilon) \;=\; \big(\tfrac{1}{N}, \tfrac{1}{N-1}, \tfrac{1}{N-2}, \ldots, \tfrac{1}{2}\big) \;\in\; (0, 1)^{N-1}.$$

2 Show that for an arbitrary $r \in (0, 1)$ one has:

$$stbr(\underbrace{r, \ldots, r}_{N-1 \text{ times}}) \;=\; \sum_{0 \leq i < N-1} r(1-r)^i \big|i\rangle \;+\; (1-r)^{N-1}\big|N-1\rangle.$$

2.2.9 Prove Equations (2.22) and (2.23).

2.3 Parallel products

Early on, in Section 1.3, we have seen Cartesian products $X \times Y$ of sets X, Y. Here we shall look at products for various collection types: subsets, multisets, and distributions. These new products will be written as tensors $\otimes$. They form parallel combinations. These tensors exist for subsets, multisets and distributions, but not for lists because they are not commutative, see Remark 2.3.3 (2).

In this section we start with a brief uniform description of parallel products, for multiple collection types — in the style of the first chapter — but we shall quickly zoom in on the probabilistic case. Products $\otimes$ of distributions have their own dynamics, due to the requirement that probabilities, also over a product, must add up to one. This means that the two components of a 'joint' distribution, over a product space, can be correlated. Indeed, a joint distribution is typically not equal to the product of its marginals: the whole is more than the product of its parts, giving crossover effects: as we shall see in Chapter 6, updating in one component has effect in other components: the components of a joint distribution 'listen' to each other.

Once parallel product (tensors) of distributions are defined, we can produce other, new distributions via images, such as 'coincidence' and 'coupon' distributions, see Subsection 2.3.1. Another construction of interest using products are convolutions of distributions, see Section 2.7.

Definition 2.3.1 Tensors $\otimes$, also called parallel products, will be defined separately for for subsets, multisets and distributions. For arbitrary sets X, Y, we describe functions $\otimes\colon \mathcal{P}(X) \times \mathcal{P}(Y) \to \mathcal{P}(X \times Y)$ and $\otimes\colon \mathcal{M}(X) \times \mathcal{M}(Y) \to \mathcal{M}(X \times Y)$ and $\otimes\colon \mathcal{D}(X) \times \mathcal{D}(Y) \to \mathcal{D}(X \times Y)$.

1 For subsets $U \in \mathcal{P}(X)$ and $V \in \mathcal{P}(Y)$ the product subset $U \otimes V \in \mathcal{P}(X \times Y)$ is defined as:

$$U \otimes V \coloneqq \{(x, y) \in X \times Y \mid x \in U \text{ and } y \in V\}.$$

The product $U \otimes V$ is often written simply as a product of sets $U \times V$, but for reasons of uniformity we prefer to have separate notation for this product of subsets.

2 For multisets $\varphi \in \mathcal{M}(X)$ and $\psi \in \mathcal{M}(Y)$ there is a product multiset $\varphi \otimes \psi \in \mathcal{M}(X \times Y)$, namely:

$$\varphi \otimes \psi \coloneqq \sum_{x \in X,\, y \in Y} \varphi(x) \cdot \psi(y) \,|x, y\rangle \quad \text{that is} \quad (\varphi \otimes \psi)(x, y) = \varphi(x) \cdot \psi(y).$$

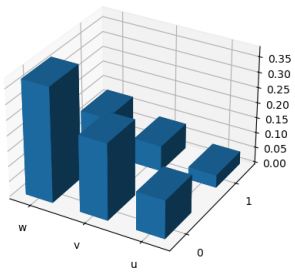
3 For distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ we use $\omega \otimes \rho$ as in the previous item (for multisets). This is well-defined, with outcome in $\mathcal{D}(X \times Y)$, since the relevant multiplicities add up to one. This tensor of distributions also works for infinite support, *i.e.* for $\mathcal{D}_\infty$.

We shall use tensors not only in binary form $\otimes$, but also in n-ary form $\otimes \cdots \otimes$, for $n \geq 2$.

We see that all of these tensors $\otimes$ involve the product $\times$ of the underlying sets. A simple illustration of a (probabilistic) tensor product is:

$$\begin{aligned} &\left(\tfrac{1}{6}|u\rangle + \tfrac{1}{3}|v\rangle + \tfrac{1}{2}|w\rangle\right) \otimes \left(\tfrac{3}{4}|0\rangle + \tfrac{1}{4}|1\rangle\right) \\ &= \tfrac{1}{8}|u, 0\rangle + \tfrac{1}{24}|u, 1\rangle + \tfrac{1}{4}|v, 0\rangle + \tfrac{1}{12}|v, 1\rangle + \tfrac{3}{8}|w, 0\rangle + \tfrac{1}{8}|w, 1\rangle. \end{aligned}$$

Such a (two-dimensional) joint distribution can be represented as a (three-dimensional) bar chart, of the form:



(2.25)

In general, these tensor products tend to grow quickly in size, since the number of entries of the two parts have to be multiplied.

We collect some basic properties.

Lemma 2.3.2 *1 The size of a tensor product of subsets or multisets is the multiplication of the sizes of the components:*

$$|U \otimes V| = |U| \cdot |V| \qquad \|\varphi \otimes \psi\| = \|\varphi\| \cdot \|\psi\|.$$

2 *Marginalisation separates tensors into the original components — assuming that the subsets U, V are non-empty.*

$$\begin{aligned} \mathcal{P}(\pi_1)(U \otimes V) &= U \qquad & \mathcal{M}(\pi_1)(\varphi \otimes \psi) &= \|\psi\| \cdot \varphi \qquad & \mathcal{D}(\pi_1)(\omega \otimes \rho) &= \omega \\ \mathcal{P}(\pi_2)(U \otimes V) &= V \qquad & \mathcal{M}(\pi_2)(\varphi \otimes \psi) &= \|\varphi\| \cdot \psi \qquad & \mathcal{D}(\pi_2)(\omega \otimes \rho) &= \rho. \end{aligned}$$

3 *Tensors are natural: for functions $f\colon X \to A$ and $g\colon Y \to B$,*

$$\begin{aligned} \mathcal{P}(f \times g)(U \otimes V) &= \mathcal{P}(f)(U) \otimes \mathcal{P}(g)(V) \\ \mathcal{M}(f \times g)(\varphi \otimes \psi) &= \mathcal{M}(f)(\varphi) \otimes \mathcal{M}(g)(\psi) \\ \mathcal{D}(f \times g)(\omega \otimes \rho) &= \mathcal{D}(f)(\omega) \otimes \mathcal{D}(g)(\rho). \end{aligned}$$

In diagrammatic form, the last equation, for instance, is written as:

$$\begin{tikzcd} \mathcal{D}(X) \times \mathcal{D}(Y) \arrow[r, "\mathcal{D}(f) \times \mathcal{D}(g)"] \arrow[d, "\otimes"'] & \mathcal{D}(A) \times \mathcal{D}(B) \arrow[d, "\otimes"] \\ \mathcal{D}(X \times Y) \arrow[r, "\mathcal{D}(f \times g)"] & \mathcal{D}(A \times B) \end{tikzcd}$$

4 *The support maps $\mathrm{supp}\colon \mathcal{M} \Rightarrow \mathcal{P}$ and $\mathrm{supp}\colon \mathcal{D} \Rightarrow \mathcal{P}$ commute with tensors, in the sense that:*

$$\mathrm{supp}(\varphi \otimes \psi) = \mathrm{supp}(\varphi) \otimes \mathrm{supp}(\psi) \qquad \mathrm{supp}(\omega \otimes \rho) = \mathrm{supp}(\omega) \otimes \mathrm{supp}(\rho).$$

5 *The frequentist learning map $\mathrm{Flrn}\colon \mathcal{M}_* \Rightarrow \mathcal{D}$ also commutes with tensors:*

$$\mathrm{Flrn}(\varphi \otimes \psi) = \mathrm{Flrn}(\varphi) \otimes \mathrm{Flrn}(\psi).$$

The last two items express that support and frequentist learning are 'monoidal' natural transformations.

Proof. 1 The case of subsets is trivial. For multisets:

$$\big\| \varphi \otimes \psi \big\| = \sum_{x \in X,\, y \in Y} \varphi(x) \cdot \psi(y) = \left(\sum_{x \in X} \varphi(x) \right) \cdot \left(\sum_{y \in Y} \psi(y) \right) = \big\|\varphi\big\| \cdot \big\|\psi\big\|.$$

2 For subsets we have

$$\mathcal{P}(\pi_1)(U \otimes V) = \{\pi_1(x, y) \mid x \in U, y \in V\} = U, \qquad \text{using that } V \neq \emptyset.$$

For multisets the size shows up:

$$\begin{aligned} \mathcal{M}(\pi_1)(\varphi \otimes \psi) &= \sum_{x \in X,\, y \in Y} \varphi(x) \cdot \psi(y) \,\big|\pi_1(x, y)\big\rangle \\ &= \sum_{x \in X} \varphi(x) \cdot \left(\sum_{y \in Y} \psi(y) \right) \big|x\big\rangle = \sum_{x \in X} \varphi(x) \cdot \|\psi\| \,\big|x\big\rangle = \|\psi\| \cdot \varphi. \end{aligned}$$

For distributions ω, ρ, the computation $\mathcal{D}(\pi_1)(\omega \otimes \rho)$ works similarly and yields $\|\rho\| \cdot \omega = \omega$ since $\|\rho\| = 1$.

3 We do the distribution case and leave the other two cases to the reader.

$$\begin{aligned}
\mathcal{D}(f \times g)(\omega \otimes \rho) &= \sum_{x \in X,\, y \in Y} \omega(x) \cdot \rho(y) \,\big|\, (f \times g)(x, y) \rangle \\
&= \sum_{x \in X,\, y \in Y} \omega(x) \cdot \rho(y) \,\big|\, f(x), g(y) \rangle \\
&= \left(\sum_{x \in X} \omega(x) \,\big|\, f(x) \rangle \right) \otimes \left(\sum_{y \in y} \rho(y) \,\big|\, g(y) \rangle \right) \\
&= \mathcal{D}(f)(\omega) \otimes \mathcal{D}(g)(\rho).
\end{aligned}$$

4 We prove the required result for multisets; the proof for distributions works in the same way.

$$\begin{aligned}
supp(\varphi \otimes \psi) &= \{(x, y) \mid (\varphi \otimes \psi)(x, y) \neq 0\} \\
&= \{(x, y) \mid \varphi(x) \cdot \psi(y) \neq 0\} \\
&= \{(x, y) \mid \varphi(x) \neq 0 \text{ and } \psi(y) \neq 0\} \\
&= \{(x, y) \mid x \in supp(\varphi) \text{ and } y \in supp(\psi)\} \;=\; supp(\varphi) \otimes supp(\psi).
\end{aligned}$$

5 We use commutation of size with tensor from item (1) in:

$$\begin{aligned}
Flrn(\varphi \otimes \psi)(x, y) \;=\; \frac{(\varphi \otimes \psi)(x, y)}{\|\varphi \otimes \psi\|} \;=\; \frac{\varphi(x)}{\|\varphi\|} \cdot \frac{\psi(y)}{\|\psi\|} \;&=\; Flrn(\varphi)(x) \cdot Flrn(\psi)(y) \\
&=\; (Flrn(\varphi) \otimes Flrn(\psi))(x, y).
\end{aligned}$$

❑

As promised, we look into why parallel products do *not* work for lists.

Remark 2.3.3 1 In traditional notation a joint distribution is often written via an expression $p(x, y)$ with two variables x, y. The corresponding first and second marginals are written as $p(x) = \sum_y p(x, y)$ and $p(y) = \sum_x p(x, y)$. The tensor product of two arbitrary distributions $p(x)$ and $p(y)$ is $p(x) \cdot p(y)$. This notation leaves the distributions involved implicit and makes it difficult to express basic properties, like $\mathcal{D}(\pi_i)(\omega_1 \otimes \omega_2) = \omega_i$, saying that marginalisations of a product of two distributions yield the original distributions (for $i = 1, 2$). In this book we avoid the (strange) tradition of writing every distribution as 'p', since it limits what can be expressed and leaves so much implicit — easily leading to unclarity and confusion. Here, distributions are mathematical objects, like lists, subsets and multisets, inhabiting their own sets $\mathcal{D}(X)$, with their own operations, like parallel product $\otimes \colon \mathcal{D}(X_1) \times \mathcal{D}(X_2) \to \mathcal{D}(X_1 \times X_2)$ and marginalisation $\mathcal{D}(\pi_i) \colon \mathcal{D}(X_1 \times X_2) \to \mathcal{D}(X_i)$, satisfying their own properties, like $\mathcal{D}(\pi_i) \circ \otimes = \pi_i$.

2 Suppose we have two list $[a, b, c]$ and $[u, v]$ and we wish to form their parallel product. Then there are many ways to do so. For instance, two obvious choices are:

$$\begin{aligned}
&[\langle a, u\rangle, \langle a, v\rangle, \langle b, u\rangle, \langle b, v\rangle, \langle c, u\rangle, \langle c, v\rangle] \\
&[\langle a, u\rangle, \langle b, u\rangle, \langle c, u\rangle, \langle a, v\rangle, \langle b, v\rangle, \langle c, v\rangle].
\end{aligned}$$

There are many other possibilities. The problem is that there is no canonical choice. Since the order of elements in a list matters, there is no commutativity property which makes all options equivalent, like for multisets. Technically, the tensor for $\mathcal{L}$ does not exist because $\mathcal{L}$ is not a *commutative* (*i.e. monoidal*) monad; this is an early result in category theory going back to Kock (1971b).

Back to tensors of distributions. We can form a K-ary product X^K for a set X. But also for a distribution $\omega \in \mathcal{D}(X)$ we can form $\omega^K = \omega \otimes \cdots \otimes \omega \in \mathcal{D}(X^K)$. This lies at the heart of the notion of 'independent and identical distributions'. We define separate functions for such constructions.

Definition 2.3.4 Fix a set X and a number K.

1 Independent and identical K-ary distributions are obtained via the function:

$$\mathcal{D}(X) \xrightarrow{iid[K]} \mathcal{D}(X^K) \qquad \text{with} \qquad iid[K](\omega) := \underbrace{\omega \otimes \cdots \otimes \omega}_{K \text{ times}}. \tag{2.26}$$

2 The K-ary tensors form a function:

$$\mathcal{D}(X)^K \xrightarrow{\bigotimes[K]} \mathcal{D}(X^K) \qquad \text{with} \qquad \bigotimes[K](\vec{\omega}) := \omega_1 \otimes \cdots \otimes \omega_K. \tag{2.27}$$

The parameter K in $iid[K]$ and $\bigotimes[K]$ may be omitted. For $K = 0$ we use that X^K is a singleton set, so that $\mathcal{D}(X^K)$ is also a singleton.

We can describe $iid[K]$ diagrammatically as composite, making the copying of distributions explicit:

$$\begin{tikzcd} \mathcal{D}(X) \arrow[rr, "iid[K]"] \arrow[dr, bend right, "\Delta"'] & & \mathcal{D}(X^K) \\ & \mathcal{D}(X)^K \arrow[ur, bend right, "\bigotimes[K]"'] & \end{tikzcd} \tag{2.28}$$

We collect some basic properties, involving the zip isomorphism $zip\colon X^K \times Y^K \xrightarrow{\cong} (X \times Y)^K$ from Exercises 1.3.7 and 1.12.4.

Lemma 2.3.5 *Fix a number $K \in \mathbb{N}$.*

1 The iid and $\bigotimes$ maps are natural: for each function $f\colon X \to Y$ the following two diagrams commute.

$$\begin{tikzcd} \mathcal{D}(X) \arrow[r, "iid"] \arrow[d, "\mathcal{D}(f)"'] & \mathcal{D}(X^K) \arrow[d, "\mathcal{D}(f^K)"] \\ \mathcal{D}(Y) \arrow[r, "iid"] & \mathcal{D}(Y^K) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{D}(X)^K \arrow[r, "\bigotimes"] \arrow[d, "\mathcal{D}(f)^K"'] & \mathcal{D}(X^K) \arrow[d, "\mathcal{D}(f^K)"] \\ \mathcal{D}(Y^K) \arrow[r, "\bigotimes"] & \mathcal{D}(Y^K) \end{tikzcd}$$

2 The zip function commutes with multiple tensors in the following way.

$$\begin{tikzcd} \mathcal{D}(X)^K \times \mathcal{D}(Y)^K \arrow[r, "\bigotimes \times \bigotimes"] \arrow[d, "zip"'] & \mathcal{D}(X^K) \times \mathcal{D}(Y^K) \arrow[r, "\otimes"] & \mathcal{D}(X^K \times Y^K) \arrow[d, "\mathcal{D}(zip)"] \\ (\mathcal{D}(X) \times \mathcal{D}(Y))^K \arrow[r, "\otimes^K"] & \mathcal{D}(X \times Y)^K \arrow[r, "\bigotimes"] & \mathcal{D}((X \times Y)^K) \end{tikzcd}$$

3 Zip also commutes with copied tensors:

$$\begin{tikzcd} \mathcal{D}(X) \times \mathcal{D}(Y) \arrow[r, "iid \times iid"] \arrow[d, "\otimes"'] & \mathcal{D}(X^K) \times \mathcal{D}(Y^K) \arrow[r, "\otimes"] & \mathcal{D}(X^K \times Y^K) \arrow[d, "\mathcal{D}(zip)"] \\ \mathcal{D}(X \times Y) \arrow[rr, "iid"] & & \mathcal{D}((X \times Y)^K) \end{tikzcd}$$

Proof. 1 This is a direct consequence of Lemma 2.3.2 (3).

2 For distributions $\omega_i \in \mathcal{D}(X)$, $\rho_i \in \mathcal{D}(Y)$ we elaborate:

$$\begin{aligned}
&\Big(\mathcal{D}(zip) \circ \otimes \circ (\bigotimes \times \bigotimes)\Big)(\vec{\omega}, \vec{\rho}) \\
&= \sum_{\vec{x} \in X^K,\, \vec{y} \in Y^K} \Big(\bigotimes(\vec{\omega}) \otimes \bigotimes(\vec{\rho})\Big)(\vec{x}, \vec{y}) \,\big|\, zip(\vec{x}, \vec{y})\big\rangle \\
&= \sum_{\vec{x} \in X^K,\, \vec{y} \in Y^K} \bigotimes(\vec{\omega})(\vec{x}) \cdot \bigotimes(\vec{\rho})(\vec{y}) \,\big|\, zip(\vec{x}, \vec{y})\big\rangle \\
&= \sum_{\vec{x} \in X^K,\, \vec{y} \in Y^K} \omega_1(x_1) \cdot \ldots \cdot \omega_K(x_K) \cdot \rho_1(y_1) \cdot \ldots \cdot \rho_K(y_K) \,\big|\, zip(\vec{x}, \vec{y})\big\rangle \\
&= \sum_{\vec{x} \in X^K,\, \vec{y} \in Y^K} \omega_1(x_1) \cdot \rho_1(y_1) \cdot \ldots \cdot \omega_K(x_K) \cdot \rho_K(y_K) \,\big|\, zip(\vec{x}, \vec{y})\big\rangle \\
&= \sum_{\vec{x} \in X^K,\, \vec{y} \in Y^K} (\omega_1 \otimes \rho_1)(x_1, y_1) \cdot \ldots \cdot (\omega_K \otimes \rho_K)(x_K, y_K) \,\big|\, zip(\vec{x}, \vec{y})\big\rangle \\
&= \sum_{\vec{x} \in X^K,\, \vec{y} \in Y^K} \bigotimes\Big(\omega_1 \otimes \rho_1, \ldots, \omega_K \otimes \rho_K\Big)((x_1, y_1), \ldots, (x_K, y_K)) \,\big|\, zip(\vec{x}, \vec{y})\big\rangle \\
&= \sum_{\vec{z} \in (X \times Y)^K} \bigotimes\Big(\omega_1 \otimes \rho_1, \ldots, \omega_K \otimes \rho_K\Big)(\vec{z}) \,\big|\, \vec{z}\big\rangle \\
&= (\bigotimes \circ \otimes^K)((\omega_1, \rho_1), \ldots, (\omega_K, \rho_K)) \;=\; (\bigotimes \circ \otimes^K \circ zip)(\vec{\omega}, \vec{\rho}).
\end{aligned}$$

3 This follows from the previous item and (2.28), using the equation $zip \circ (\Delta \times \Delta) = \Delta$, see Exercise 1.3.7. ❑

Example 2.3.6 Suppose we have a distribution $\omega \in \mathcal{D}(\boldsymbol{n})$ for $n \geq 2$, where, recall, $\boldsymbol{n} = \{0, \ldots, n-1\}$. We like to use ω to form a mixture of n distributions $\rho_0, \ldots, \rho_{n-1} \in \mathcal{D}(X)$, for some set X. This takes the form of a convex sum of distributions (2.9):

$$mix(\omega, \vec{\rho}) \;\coloneqq\; \omega(0) \cdot \rho_0 + \cdots + \omega(n-1) \cdot \rho_{n-1} \;=\; \sum_{i \in \boldsymbol{n}} \omega(i) \cdot \rho_i. \tag{2.29}$$

Using tensors we can describe this function *mix* as a composite of the following form:

$$\begin{tikzcd}
\mathcal{D}(\boldsymbol{n}) \times \mathcal{D}(X)^n \arrow[r, "mix"] \arrow[d, "id \times \bigotimes"'] & \mathcal{D}(X) \\
\mathcal{D}(\boldsymbol{n}) \times \mathcal{D}(X^n) \arrow[r, "\otimes"] & \mathcal{D}(\boldsymbol{n} \times X^n) \arrow[u, "\mathcal{D}(proj)"']
\end{tikzcd}$$

where $proj(i, \vec{x}) = x_i$. Indeed:

$$\begin{aligned}
\Big(\mathcal{D}(proj) \circ \otimes \circ (id \times \bigotimes)\Big)(\omega, \vec{\rho}) &= \sum_{i \in \boldsymbol{n},\, \vec{x} \in X^n} (\omega \otimes (\bigotimes \vec{\rho}))(i, \vec{x}) \,\big|\, proj(i, \vec{x})\big\rangle \\
&= \sum_{i \in \boldsymbol{n},\, \vec{x} \in X^n} \omega(i) \cdot (\bigotimes \vec{\rho})(\vec{x}) \,\big|\, x_i\big\rangle \\
&= \sum_{i \in \boldsymbol{n},\, \vec{x} \in X^n} \omega(i) \cdot \rho_0(x_0) \cdot \ldots \cdot \rho_{n-1}(x_{n-1}) \,\big|\, x_i\big\rangle \\
&= \sum_{i \in \boldsymbol{n},\, y \in X} \omega(i) \cdot \rho_i(y) \,\big|\, y\big\rangle \;=\; \sum_{i \in \boldsymbol{n}} \omega(i) \cdot \rho_i.
\end{aligned}$$

Later on, Exercise 2.4.7 will describe this mixture in terms of state transformation.

2.3.1 Image distributions from products

In this section we describe two classes of distributions via images of products. More specifically, for a distribution $\omega \in \mathcal{D}(X)$ we first form the K-ary product $iid[K](\omega) = \omega^K = \omega \otimes \cdots \otimes \omega \in \mathcal{D}(X^K)$ and then take image / pushforward distributions of the form $\mathcal{D}(f)(\omega^K)$, for two different functions $f\colon X^K \to Y$. The two functions f that we use come from the triangular prism (1.52), namely matching *mat*, and support *supp*. Later, in Theorem 2.6.7, we shall also use the accumulation function *acc* in this manner (for the multinomial distribution). The two distributions that we introduce below will be called *coincidence cd* and *coupon cpn*. The descriptions are copied from Jacobs (2022a) and use the match function $mat\colon X^K \to SP(K)$ from lists to set partitions (1.23) of $\{1, \ldots, K\}$ and the support fuction $supp\colon X^K \to \mathcal{P}_{\leq}[K](X)$ from lists to subsets (1.15), where, recall, $\mathcal{P}_{\leq}[K](X) = \{U \in \mathcal{P}_{fin}(X) \mid 1 \leq |U| \leq K\}$. Explanations and illustrations will be provided below.

Definition 2.3.7 Let $\omega \in \mathcal{D}(X)$ be distribution and let $K \in \mathbb{N}_{>0}$.

1 The *coincidence* distribution *cd* is defined via pushforward along the match function as:

$$cd[K](\omega) \coloneqq \mathcal{D}(mat)\Big(iid[K](\omega)\Big) \in \mathcal{D}\Big(SP(K)\Big). \tag{2.30}$$

2 The *coupon* distribution *cpn* arises via pushforward along the support function:

$$cpn[K](\omega) \coloneqq \mathcal{D}(supp)\Big(iid[K](\omega)\Big) \in \mathcal{D}\Big(\mathcal{P}_{\leq}[K](X)\Big). \tag{2.31}$$

We explain the latter coupon distribution first. We now think of the elements of the set X as different 'coupons' that may occur in cereal boxes. The number $\omega(x) \in [0, 1]$ gives the probability of finding coupon x in an arbitrary box. Thus, we are assuming that each box contains a coupon, but it is unclear from the outside which one. The idea is that people are interested in collecting a complete set of coupons. Hence duplicates are irrelevant — except possibly for exchange, which we ignore here.

To calculate the probability of such subsets of coupons we apply pushforward along the support function $supp\colon X^K \to \mathcal{P}_{\leq}[K](X)$ to a product distribution ω^K, see (2.31). The recipe (2.31) tells that for a subset U we have to sum the probabilities of all list ℓ with $supp(\ell) = U$.

For an illustration we use the distribution $\rho = \frac{1}{2}|a\rangle + \frac{1}{3}|b\rangle + \frac{1}{6}|c\rangle$. The probabilities for the various non-empty subsets (of coupons) of $X = \{a, b, c\}$, when we buy four boxes of cereal, are given by the following distribution.

$$\begin{aligned} cpn[4](\rho) &= \tfrac{1}{16}\Big|\{a\}\Big\rangle + \tfrac{1}{81}\Big|\{b\}\Big\rangle + \tfrac{11}{27}\Big|\{a,b\}\Big\rangle + \tfrac{1}{1296}\Big|\{c\}\Big\rangle \\ &\quad + \tfrac{29}{216}\Big|\{a,c\}\Big\rangle + \tfrac{4}{81}\Big|\{b,c\}\Big\rangle + \tfrac{1}{3}\Big|\{a,b,c\}\Big\rangle. \end{aligned}$$

We elaborate the probability $\frac{11}{27}$ of the subset $\{a, b\}$. Lemma 1.4.9 tells that the number of lists ℓ of length 4 with support $\{a, b\}$ is $2 \cdot \left\{ {4 \atop 2} \right\} = 2 \cdot 7 = 14$. Below we describe the probabilities $\omega^4(\ell)$ for all these lists ℓ with $supp(\ell) = \{a, b\}$.

$$\begin{array}{lll} \omega^4(a,a,a,b) = \frac{1}{24} & \omega^4(a,a,b,a) = \frac{1}{24} & \omega^4(a,a,b,b) = \frac{1}{36} \\ \omega^4(a,b,a,a) = \frac{1}{24} & \omega^4(a,b,a,b) = \frac{1}{36} & \omega^4(a,b,b,a) = \frac{1}{36} \\ \omega^4(a,b,b,b) = \frac{1}{54} & \omega^4(b,a,a,a) = \frac{1}{24} & \omega^4(b,a,a,b) = \frac{1}{36} \\ \omega^4(b,a,b,a) = \frac{1}{36} & \omega^4(b,a,b,b) = \frac{1}{54} & \omega^4(b,b,a,a) = \frac{1}{36} \\ \omega^4(b,b,a,b) = \frac{1}{54} & \omega^4(b,b,b,a) = \frac{1}{54} & \end{array}$$

One may check that these 14 probabilities add up to $\frac{11}{27}$. In a similar way one obtains the other probabilities of the above distribution $cpn[4](\rho)$. It is quite a bit of work to do this manually, but a computer program does it easily.

This coupon distribution is most familiar for the special case when the distribution is uniform. Let's assume that the set X has $N \geq 1$ elements, with uniform distribution $\upsilon = \sum_{x \in X} \frac{1}{N}|x\rangle$. The resulting coupon probabilities are determined by the size of the subsets:

$$cpn[K](\upsilon) = \sum_{U \in \mathcal{P}_{\leq}[K](X)} \frac{|U|!}{N^K} \cdot \left\{ \begin{matrix} K \\ |U| \end{matrix} \right\} |U\rangle.$$

This follows from Equation (1.16).

We next illustrate the coincidence distribution (2.30), obtained via a pushforward along the match function $mat\colon X^K \to SP(K)$ from (1.23). It assigns to each set partition $P \in SP(K)$ a probability, given as sum of probabilities of lists in X^K that match to P.

For instance, for the same distribution $\rho = \frac{1}{2}|a\rangle + \frac{1}{3}|b\rangle + \frac{1}{6}|c\rangle$ that we used above, and for parameter $K = 3$, there is the following coincidence distribution over the set $SP(3)$ of set partitions over $\{1, 2, 3\}$.

$$\begin{aligned} cd[3](\rho) &= \tfrac{1}{6}\Big|\{\{1,2,3\}\}\Big\rangle + \tfrac{2}{9}\Big|\{\{1,2\},\{3\}\}\Big\rangle + \tfrac{2}{9}\Big|\{\{1,3\},\{2\}\}\Big\rangle \\ &\quad + \tfrac{2}{9}\Big|\{\{1\},\{2,3\}\}\Big\rangle + \tfrac{1}{6}\Big|\{\{1\},\{2\},\{3\}\}\Big\rangle. \end{aligned}$$

We elaborate the probability $\frac{2}{9}$ of the set partition $P = \{\{1,2\},\{3\}\} \in SP(3)$. According to Lemma 1.5.9 (2) there are $\frac{3!}{(3-2)!} = 6$ lists ℓ of length 3 with $mat(\ell) = P$. The probabilities $\omega^3(\ell)$ are:

$$\begin{array}{lll} \omega^3(a,a,b) = \frac{1}{12} & \omega^3(a,a,c) = \frac{1}{24} & \omega^3(b,b,a) = \frac{1}{18} \\ \omega^3(b,b,c) = \frac{1}{54} & \omega^3(c,c,a) = \frac{1}{72} & \omega^3(c,c,b) = \frac{1}{108} \end{array}$$

These 6 probabilities add up to $\frac{2}{9}$.

When a set X has $N \geq K$ elements and υ is the uniform distribution on X, then via Equation (1.25) we get as coincidence distribution:

$$cd[K](\upsilon) = \sum_{P \in SP(K)} \frac{N!}{(N - |P|)! \cdot N^K} |P\rangle \overset{(1.25)}{=} \sum_{P \in SP(K)} \frac{|mat^{-1}(P)|}{N^K} |P\rangle.$$

This is a proper probability distribution by Exercise 1.5.7. For instance, for $N = 4$ and $K = 3$,

$$\begin{aligned} cd[3](\upsilon) &= \tfrac{1}{16}\Big|\{\{1,2,3\}\}\Big\rangle + \tfrac{3}{16}\Big|\{\{1,2\},\{3\}\}\Big\rangle + \tfrac{3}{16}\Big|\{\{1,3\},\{2\}\}\Big\rangle \\ &\quad + \tfrac{3}{16}\Big|\{\{1\},\{2,3\}\}\Big\rangle + \tfrac{3}{8}\Big|\{\{1\},\{2\},\{3\}\}\Big\rangle. \end{aligned}$$

2.3.2 Marginalisation and copying

In traditional notation in probability theory one simply omits variables to express marginalisation. For instance, for a distribution $\omega \in \mathcal{D}(X_1 \times X_2 \times X_3 \times X_4)$, considered as function $\omega(x_1, x_2, x_3, x_4)$ in four variables x_i, one writes:

$$\omega(x_2, x_3) \qquad \text{for the marginal} \qquad \sum_{x_1, x_4} \omega(x_1, x_2, x_3, x_4).$$

This notation can be ambiguous and confusing. Since marginalisation plays an important role, we like to deal with it via an explicit operation, and not just via a reduction of the number of variables without leaving a trace of the variables that have been marginalised out. Therefore we introduce special notation for marginalisation, in post-fix form, via 'masks'. Such a mask lists not only the variables that remain (via a 1) but also the variable that are marginalised out and disappear (via a 0).

Definition 2.3.8 Let T be one of $\mathcal{P}$, $\mathcal{M}$ or $\mathcal{D}$.

1 A *mask* M is a finite list of 0's and 1's, that is, an element $M \in \mathcal{L}(\{0,1\})$. For a joint state $\omega \in T(X_1 \times \cdots \times X_n)$ of type T and a mask M of length n we write:

$$\omega M$$

for the marginal with mask M. Informally, it keeps all the parts from ω at a position where there is 1 in M and it projects away parts where there is 0. This is best illustrated via an example:

$$\omega[1,0,1,0,1] \;=\; T(\langle \pi_1, \pi_3, \pi_5 \rangle)(\omega) \;\in\; T(X_1 \times X_3 \times X_5).$$

2 A joint state $\omega \in T(X \times Y)$ is called *non-entwined* if it is the parallel product of its marginals:

$$\omega \;=\; \omega[1,0] \otimes \omega[0,1].$$

Otherwise it is called *entwined*. This notion of entwinedness may be formulated with respect to n-ary states, via a mask, see for example Exercise 2.3.2, but it may then need some rearrangement of components.

3 Let $\sigma \in T(X)$ and $\tau \in T(Y)$ be given. A *coupling* of σ, τ is a joint state $\omega \in T(X \times Y)$ that marginalises to both σ and τ, as in:

$$\omega[1,0] \;=\; \sigma \qquad\qquad \omega[0,1] \;=\; \tau.$$

In the probabilistic case there typically are infinitely many couplings, see Exercise 2.3.8.

A product distribution of the form $\omega_1 \otimes \omega_2 \in \mathcal{D}(X_1 \times X_2)$ is always non-entwined, see Lemma 2.3.2 (2). But in general, joint distributions are entwined, as will be illustrated below.

Entwinedness of a joint distribution means that the different parts are correlated and can influence each other. This is a mechanism that will play an important role in the sequel, via updating, see for instance in Example 6.1.2 (2). A joint distribution is thus more than the product of its parts.

Non-entwined joint distributions are called separable in Coecke and Kissinger (2016). Sometimes they are called independent, although independence is also used for random variables. We like to have separate terminology for states only, so we use the phrase (non-)entwinedness, which is a new expression. Independence for random variables is described in Section 5.4.

Example 2.3.9 Take sets $X = \{u, v\}$ and $A = \{a, b\}$.

1 Consider the joint subset:

$$W \;=\; \{\, (u,a),\ (u,b),\ (v,a) \,\} \;\subseteq\; X \times A.$$

Its marginals are the subsets:

$$W[1,0] \;=\; \mathcal{P}(\pi_1)(W) \;=\; \{u, v\} \;=\; X \qquad\qquad W[0,1] \;=\; \mathcal{P}(\pi_2)(W) \;=\; \{a, b\} \;=\; A.$$

But clearly, $W \neq X \otimes A = W[1,0] \otimes W[0,1]$. Hence, W is an entwined joint subset.

2 Now consider the joint distribution $\omega \in \mathcal{D}(X \times A)$ given by:

$$\omega \;=\; \tfrac{1}{8}|u,a\rangle + \tfrac{1}{4}|u,b\rangle + \tfrac{3}{8}|v,a\rangle + \tfrac{1}{4}|v,b\rangle.$$

We claim that ω is entwined. The first and second marginals $\omega[1,0] = \mathcal{D}(\pi_1)(\omega) \in \mathcal{D}(X)$ and $\omega[0,1] = \mathcal{D}(\pi_2)(\omega) \in \mathcal{D}(A)$ are:

$$\omega[1,0] \;=\; \tfrac{3}{8}|u\rangle + \tfrac{5}{8}|v\rangle \qquad \text{and} \qquad \omega[0,1] \;=\; \tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle.$$

The original state ω differs from the product of its marginals, which is:

$$\omega[1,0] \otimes \omega[0,1] \;=\; \tfrac{3}{16}|u,a\rangle + \tfrac{3}{16}|u,b\rangle + \tfrac{5}{16}|v,a\rangle + \tfrac{5}{16}|v,b\rangle.$$

This entwinedness in the 2×2 case follows from a general characterisation, see Exercise 2.3.10 below.

We move from discarding (via marginalisation) to copying. Recall the copy function $\Delta\colon X \to X^K$ from Subsection 1.3.2. It is used in the following observation.

Fact 2.3.10 A pushforward along a copy differs from an iterated tensor. This will be illustrated for subsets and distributions. We consider the binary case, with copier $\Delta\colon X \to X \times X$, given by $\Delta(x) = (x,x)$.

1 For a subset $U \subseteq X$,

$$\mathcal{P}(\Delta)(U) \;\neq\; U \otimes U.$$

For instance, for $X = \{a,b,c\}$ and $U = \{a,b\}$ one has:

$$\begin{aligned} \mathcal{P}(\Delta)(U) \;&=\; \{\Delta(x) \mid x \in U\} \;=\; \{\,(a,a),\ (b,b)\,\}. \\ U \otimes U \;&=\; \{(x,y) \mid x,y \in U\} \;=\; \{\,(a,a),\ (a,b), (b,a),\ (b,b)\,\}. \end{aligned}$$

2 Similarly, for a distribution $\omega \in \mathcal{D}(X)$,

$$\mathcal{D}(\Delta)(\omega) \;\neq\; \omega \otimes \omega$$

Indeed, the outcome of flipping two coins is not the same as flipping a coin once and copying it. In detail, for $\omega = \frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle$,

$$\mathcal{D}(\Delta)(\omega) \;=\; \tfrac{1}{2}\big|\Delta(0)\big\rangle + \tfrac{1}{2}\big|\Delta(1)\big\rangle \;=\; \tfrac{1}{2}\big|0,0\big\rangle + \tfrac{1}{2}\big|1,1\big\rangle. \tag{2.32}$$

In contrast:

$$\omega \otimes \omega \;=\; \tfrac{1}{4}\big|0,0\big\rangle + \tfrac{1}{4}\big|0,1\big\rangle + \tfrac{1}{4}\big|1,0\big\rangle + \tfrac{1}{4}\big|1,1\big\rangle. \tag{2.33}$$

In Exercise 2.3.13 we see that only point distributions are 'copyable'.

At the end of this section on products (tensors $\otimes$) of distributions we re-emphasise two special properties of these products.

- A joint distribution is typically *not* the product of its marginals: for $\tau \in \mathcal{D}(X \times Y)$ there is in general a difference between τ and the product $\tau[1,0] \otimes \tau[0,1]$ of its two marginals.
- The product of a distribution with itself is *not* a copy: for $\omega \in \mathcal{D}(X)$ there is typically a difference between $\omega \otimes \omega$ and $\mathcal{D}(\Delta)(\omega)$.

In Section 2.10 we shall use so-called KL-divergence to measure differences between distributions. The divergence between a joint distribution and the product of its marginals is called its *mutual information*; the divergence between a product of a distribution with itself and a copy is called its *entropy*, see Subsection 2.10.1.

Exercises

2.3.1 1 Prove that if a joint state has full support, then each of its marginals has full support.

2 Consider the joint distribution $\omega = \frac{1}{2}|a,b\rangle + \frac{1}{2}|a^\perp,b^\perp\rangle \in \mathcal{D}(A \times B)$ for $A = \{a,a^\perp\}$ and $B = \{b,b^\perp\}$. Check that both marginals $\omega[1,0] \in \mathcal{D}(A)$ and $\omega[0,1] \in \mathcal{D}(B)$ have full support, but ω itself does not. Conclude that the converse of the previous point does not hold.

2.3.2 Check that the distribution $\omega \in \mathcal{D}(\{a,b\} \times \{a,b\} \times \{a,b\})$ given by:

$$\omega = \tfrac{1}{24}|aaa\rangle + \tfrac{1}{12}|aab\rangle + \tfrac{1}{12}|aba\rangle + \tfrac{1}{6}|abb\rangle + \tfrac{1}{6}|baa\rangle + \tfrac{1}{3}|bab\rangle + \tfrac{1}{24}|bba\rangle + \tfrac{1}{12}|bbb\rangle$$

satisfies:

$$\omega = \omega[1,1,0] \otimes \omega[0,0,1].$$

2.3.3 Consider the following table-as-multiset $\psi \in \mathcal{N}[800](F \times S \times D)$ from (Jewell, 2009, Table 8.4),

$$\begin{aligned}\psi = {} & 160\big|f,s,d\rangle + 40\big|f,s,d^\perp\rangle + 120\big|f,s^\perp,d\rangle + 80\big|f,s^\perp,d^\perp\rangle \\ & + 80\big|f^\perp,s,d\rangle + 120\big|f^\perp,s,d^\perp\rangle + 40\big|f^\perp,s^\perp,d\rangle + 160\big|f^\perp,s^\perp,d^\perp\rangle,\end{aligned}$$

with sets F for fluoridation, S for high sugar-diet and D for tooth decay, where:

$$F = \{f,f^\perp\} \qquad S = \{s,s^\perp\} \qquad D = \{d,d^\perp\}.$$

1 Compute the emperical distribution $\tau := \mathit{Flrn}(\psi)$.

2 Show that the fluoridation and sugar-diet marginal is non-entwined, that is,

$$\tau[1,1,0] = \tau[1,0,0] \otimes \tau[0,1,0].$$

The 'shape' of the distribution τ will appear in Exercise 7.8.5, using this non-entwinedness.

2.3.4 Let X, Y be two non-empty finite sets. Show that:

$$\mathit{unif}_X \otimes \mathit{unif}_Y = \mathit{unif}_{X \times Y}.$$

1 Conclude that the uniform distribution on a product is non-entwined and that its marginals are again uniform.

2 Apply the above equation to the standard 52-card deck, given as product of a 4-element set 'Suit' and a 13-element set 'Rank'.

2.3.5 Find different joint states σ, τ with equal marginals:

$$\sigma[1,0] = \tau[1,0] \quad \text{and} \quad \sigma[0,1] = \tau[0,1] \quad \text{but} \quad \sigma \neq \tau.$$

Hint: Use Example 2.3.9 (2).

2.3.6 Consider $\omega = \frac{1}{4}\big|0,0,0\rangle + \frac{1}{4}\big|0,1,1\rangle + \frac{1}{4}\big|1,0,1\rangle + \frac{1}{4}\big|1,1,0\rangle \in \mathcal{D}(\mathbf{2} \times \mathbf{2} \times \mathbf{2})$.

1 Check that the three distributions $\omega[1,1,0], \omega[1,0,1], \omega[0,1,1] \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$ are all non-entwined.

2 Show that ω itself is entwined: it differs from the product $\omega[1,0,0]\otimes\omega[0,1,0]\otimes\omega[0,0,1]$ of its marginals.

2.3.7 Let $\tau \in \mathcal{D}(X \times X)$ satisfy $\tau[1,0] = 1|x\rangle$ and $\tau[0,1] = \omega$. Show that τ must be of the form $1|x\rangle \otimes \omega$.

2.3.8 1 For numbers $0 \leq s \leq r \leq 1$ form the joint distribution:

$$\omega \;:=\; s\big|1,1\rangle + (r-s)\big|1,0\rangle + (1-r)\big|0,0\rangle \;\in\; \mathcal{D}(\mathbf{2}\times\mathbf{2}).$$

Show that ω is a coupling of $\mathit{flip}(r)$ and $\mathit{flip}(s)$, *i.e.* that: $\omega[1,0] = \mathit{flip}(r)$ and $\omega[0,1] = \mathit{flip}(s)$, see Definition 2.3.8 (3).

2 Let $s, t \in (0,1)$ be given. Show that there infinitely many couplings of $\mathit{flip}(s)$ and $\mathit{flip}(t)$.

3 Use the previous points to construct a non-uniform coupling of the two uniform distributions $\mathit{flip}(\frac{1}{2})$ and $\mathit{flip}(\frac{1}{2})$.

2.3.9 Check that:

$$\omega[0,1,1,0,1,1][0,1,1,0] \;=\; \omega[0,0,1,0,1,0].$$

What is the general result behind this?

2.3.10 Let $X = \{u,v\}$ and $A = \{a,b\}$ as in Example 2.3.9. Prove that a distribution $\omega = r_1|u,a\rangle + r_2|u,b\rangle + r_3|v,a\rangle + r_4|v,b\rangle \in \mathcal{D}(X\times A)$, where $r_1+r_2+r_3+r_4 = 1$, is non-entwined if and only if $r_1 \cdot r_4 = r_2 \cdot r_3$.

2.3.11 Show that tensoring of multisets is linear, in the sense that for $\varphi \in \mathcal{M}(X)$ the 'tensor with φ' operation $\varphi \otimes (-)\colon \mathcal{M}(Y) \to \mathcal{M}(X\times Y)$ is linear w.r.t. the cone structure of Lemma 1.6.3 (2): for $\psi_1,\ldots,\psi_n \in \mathcal{M}(Y)$ and $r_1,\ldots,r_n \in \mathbb{R}_{\geq 0}$ one has:

$$\varphi \otimes \Big(r_1\cdot\psi_1 + \cdots + r_n\cdot\psi_n\Big) \;=\; r_1\cdot(\varphi\otimes\psi_1) + \cdots + r_n\cdot(\varphi\otimes\psi_n).$$

The same hold in the other coordinate, for $(-)\otimes\psi$. As a special case we obtain that when φ is a probability distribution, then $\varphi\otimes(-)\colon \mathcal{D}(Y) \to \mathcal{D}(Y)$ preserves convex sums of distributions. One says that the tensor of multisets is *bi-linear*, and the tensor of distributions is *bi-affine* (see Exercise 2.1.18).

2.3.12 (See Kock (1971a) for a broader perspective) Show that the tensor $\otimes\colon \mathcal{D}(X)\times\mathcal{D}(Y) \to \mathcal{D}(X\times Y)$ is universally bi-affine, in the following sense: for each convex set A — *i.e.* barycentric algebra, see Exercise 2.1.18 — and function $f\colon \mathcal{D}(X)\times\mathcal{D}(Y) \to A$ that is bi-affine — *i.e.* affine in each coordinate — there is a unique affine map $\overline{f}\colon \mathcal{D}(X\times Y) \to A$ in:

$$\begin{tikzcd}
\mathcal{D}(X)\times\mathcal{D}(Y) \arrow[r, "\otimes"] \arrow[dr, bend right, "\text{bi-affine } f"'] & \mathcal{D}(X\times Y) \arrow[d, dashed, "\overline{f},\ \text{affine}"] \\
& A
\end{tikzcd}$$

Hint: Define $\overline{f}(\tau) = \sum_{x,y} \tau(x,y)\cdot f(1|x\rangle, 1|y\rangle)$, so that, $\overline{f}$ is the free extension, as in Exercise 2.1.18 (3), of the map $f \circ (\mathit{unit}\times\mathit{unit})\colon X\times Y \to A$. Explicitly: $\overline{f} = \alpha \circ \mathcal{D}(f \circ (\mathit{unit}\times\mathit{unit}))$, where $\alpha\colon \mathcal{D}(A) \to A$ is the algebra corresponding to the barycentric structure on A.

2.3.13 For a state $\omega \in \mathcal{D}(X)$, show that the following statements are equivalent.

- ω is *copyable*, that is, $\omega\otimes\omega = \mathcal{D}(\Delta)(\omega)$;
- ω is a point distribution;

- ω is $\{0,1\}$-valued, *i.e.* restricts to a function $X \to \{0,1\}$.

2.3.14 Show that the following diagram commutes, in which *flat* is the flatten operation for lists, that removes inner brackets.

$$\begin{tikzcd} \mathcal{D}(X) \arrow[r, "iid[K]"] \arrow[drr, bend right, "iid[K\cdot L]"'] & \mathcal{D}(X^K) \arrow[r, "iid[L]"] & \mathcal{D}((X^K)^L) \arrow[d, "\mathcal{D}(flat)"] \\ & & \mathcal{D}(X^{K\cdot L}) \end{tikzcd}$$

2.4 Probabilistic channels

In the previous chapter we have seen channels of the form $X \to \mathcal{L}(Y)$, or $X \to \mathcal{P}(Y)$, or $X \to \mathcal{M}(Y)$. We will now introduce probabilistic channels of the form $X \to \mathcal{D}(Y)$. Recall that we often use a special arrow $\rightsquigarrow$ for such channels and then simply write them as $X \rightsquigarrow Y$.

We have seen that the operations $\mathcal{L}$, $\mathcal{P}$, $\mathcal{M}$ used for these channels are all monads, via special 'unit' and 'flatten' functions. Moreover, in terms of these unit and flatten maps we have defined identity and composition for these channels, leading to categories of channels $\mathbf{Chan}(\mathcal{L})$, $\mathbf{Chan}(\mathcal{P})$ and $\mathbf{Chan}(\mathcal{M})$, for associated forms of computation. In this section we will first show that the same monad structure exists for the distribution functor $\mathcal{D}$ — and for the infinite version $\mathcal{D}_\infty$ too — and that we can thus also organise probabilistic channels in the form of a category $\mathbf{Chan}(\mathcal{D})$, with sets as objects and probabilistic computations $X \to \mathcal{D}(Y)$ as morphisms $X \rightsquigarrow Y$. In the remainder of this book the emphasis will be almost exclusively on this probabilistic case, so that 'channel' will standardly mean 'probabilistic channel' and that the notation **Chan** will be used for $\mathbf{Chan}(\mathcal{D})$.

Lemma 2.4.1 *The unit and flatten maps for multisets from Subsection 1.6.2 restrict to distributions. The unit function* unit$\colon X \to \mathcal{D}(X)$ *is simply* unit$(x) := 1|x\rangle$. *Flattening is the function* flat$\colon \mathcal{D}(\mathcal{D}(X)) \to \mathcal{D}(X)$ *with:*

$$\mathit{flat}\Big(\textstyle\sum_i r_i|\omega_i\rangle\Big) := \textstyle\sum_i r_i \cdot \omega_i = \displaystyle\sum_{x\in X} \big(\textstyle\sum_i r_i \cdot \omega_i(x)\big)\big|x\rangle.$$

The formulation in the middle uses a convex sum of distributions (2.9).

The familiar properties of unit and flatten hold for distributions too: the analogue of Lemma 1.6.4 holds, with 'multiset' replaced by 'distribution'.

Proof. The only thing that needs to be checked is that flattening yields a convex sum, *i.e.* that its probabilities add up to one. This is easy:

$$\textstyle\sum_x \big(\sum_i r_i \cdot \omega_i(x)\big) = \sum_i r_i \cdot \sum_x \omega_i(x) = \sum_i r_i \cdot 1 = 1. \qquad \square$$

We conclude that $\mathcal{D}$, with these *unit* and *flat* maps, is a monad. The same holds for the 'infinite' variation $\mathcal{D}_\infty$ from Subsection 2.1.1, and also for the 'subdistribution' version $\mathcal{D}_{\leq 1}$ that was mentioned briefly after Definition 2.1.1.

In Lemma 2.1.3 we have seen pushforward of a distribution ω along a *function* f, producing an image distribution $\mathcal{D}(f)(\omega)$. There is also pushforward along a *channel*, which we describe next. It may be formulated in terms of multiplication of stochastic matrices, see Exercise 2.4.3.

Definition 2.4.2 Let X, Y be arbitrary sets.

1 A probabilistic channel $c\colon X \rightarrowtriangle Y$ is a function $c\colon X \to \mathcal{D}(Y)$. One can think of it as an X-indexed collection $(c(x))_{x\in X}$ of distributions $c(x) \in \mathcal{D}(Y)$ on Y. As such it is often written as a conditional probability $P(y \mid x)$.

2 For a distribution / state $\omega \in \mathcal{D}(X)$ on the domain X of a channel $c\colon X \rightarrowtriangle Y$ we can do a *pushforward*, also known as *state transformation* or *prediction*, and produce a new distribution / state $c_*(\omega) \in \mathcal{D}(Y)$ on the channel's codomain Y. This happens via the standard definition for pushforward from Section 1.11, see especially (1.58) and (1.59):

$$c_*(\omega) \;=\; \mathit{flat}\Big(\mathcal{D}(c)(\omega)\Big) \;=\; \sum_{x\in X} \omega(x)\cdot c(x) \;=\; \sum_{y\in Y}\Bigg(\sum_{x\in X} c(x)(y)\cdot\omega(x)\Bigg)\big|y\rangle. \tag{2.34}$$

The first sum $\sum_x \omega(x)\cdot c(x)$ involves a convex sum (mixture) over distributions $c(x) \in \mathcal{D}(Y)$. This formulation is easiest in actual computations.

The pushforward $c_*(\omega)$ of distribution ω along channel c is a very useful construction. However, there is no notation for it in traditional probability theory. Here is an example, copied from the statement of (Murphy, 2023, Thm. 5.1.3)[2] where pushforward does occur, in a rather indirect manner: "Consider two different distributions $p(x)$ and $q(x)$ combined with a probabilistic channel $t(y|x)$. If $p(y)$ is the distribution that results from sending samples from $p(x)$ through the channel $t(y|x)$, and similarly for $q(y)$, ...". This may be understood as $p(y) = t(y|x)_*(p(x))$. We do not distuinguish different distributions via different variable. Instead, we use different names for different distributions and leave the variables implicity. The above formulation of $c_*(\omega)$ in (2.34) makes precise what the "sending samples" from the quote means. It is also described in (2.35) below.

A distribution $\omega \in \mathcal{D}(X)$ on a set X is sometimes called a 'random element' of X. We can describe such a distribution as a channel $\omega\colon \mathbf{1} \rightarrowtriangle X$, where $\mathbf{1}$ is a singleton set. This fits with ordinary elements $x \in X$, since they correspond to ordinary functions $x\colon \mathbf{1} \to X$. More generally, one can view a channel $X \rightarrowtriangle Y$ as a 'random function' from X to Y.

We have already seen that unit and flatten restrict to distributions, so as a result, the description (2.34) produces a new distribution $c_*(\omega)$. But we can double-check that probabilities add up to one:

$$\sum_{y\in Y} c_*(\omega)(y) \;=\; \sum_{y\in Y}\Bigg(\sum_{x\in X} c(x)(y)\cdot\omega(x)\Bigg) \;=\; \sum_{x\in X}\Bigg(\sum_{y\in Y} c(x)(y)\Bigg)\cdot\omega(x) \;=\; \sum_{x\in X} 1\cdot\omega(x) \;=\; 1.$$

One may describe state tranformation $c_*(\omega)$ in operational terms via sampling (see Subsection 2.2.1). Then it looks as a probabilistic program fragment of the form:

```
x ← ω
y ← c(x)
return y
```
(2.35)

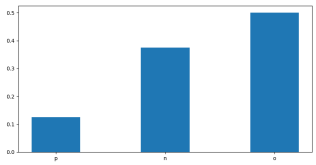
Later, in Subsection 3.8.1, we provide a mathematical justification for this notation.

Example 2.4.3 (Copied from Jacobs (2021b)) Let's assume we wish to capture the mood of a teacher, as a probabilistic mixture of three possible options namely: pessimistic (p), neutral (n)

[2] We describe this same result later on, as Proposition 2.10.6 (3), using pushforward notation.

or optimistic (o). We thus have a three-element probability space $X = \{p, n, o\}$. We assume a mood distribution:

$$\omega \;=\; \tfrac{1}{8}|p\rangle + \tfrac{3}{8}|n\rangle + \tfrac{1}{2}|o\rangle \qquad \text{with plot}$$

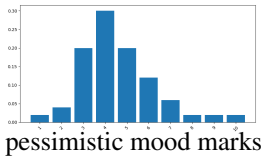


This mood thus tends towards optimism.

Associated with these different moods the teacher has different views on how pupils perform in a particular test. This performance is expressed in terms of marks, which can range from 1 to 10, where 10 is best. The probability space for these marks is written as $Y = \{1, 2, \ldots, 10\}$.

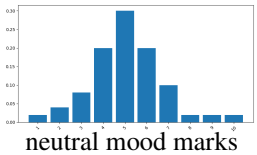
The view of the teacher is expressed via a channel $c\colon X \twoheadrightarrow Y$. It is defined via the following three mark distributions, one for each element in the set of moods $X = \{p, n, o\}$.

$$\begin{aligned} c(p) \;=\; & \tfrac{1}{50}|1\rangle + \tfrac{2}{50}|2\rangle + \tfrac{10}{50}|3\rangle + \tfrac{15}{50}|4\rangle + \tfrac{10}{50}|5\rangle \\ & + \tfrac{6}{50}|6\rangle + \tfrac{3}{50}|7\rangle + \tfrac{1}{50}|8\rangle + \tfrac{1}{50}|9\rangle + \tfrac{1}{50}|10\rangle \end{aligned}$$



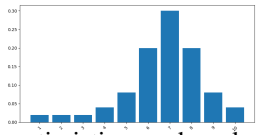
pessimistic mood marks

$$\begin{aligned} c(n) \;=\; & \tfrac{1}{50}|1\rangle + \tfrac{2}{50}|2\rangle + \tfrac{4}{50}|3\rangle + \tfrac{10}{50}|4\rangle + \tfrac{15}{50}|5\rangle \\ & + \tfrac{10}{50}|6\rangle + \tfrac{5}{50}|7\rangle + \tfrac{1}{50}|8\rangle + \tfrac{1}{50}|9\rangle + \tfrac{1}{50}|10\rangle \end{aligned}$$



neutral mood marks

$$\begin{aligned} c(o) \;=\; & \tfrac{1}{50}|1\rangle + \tfrac{1}{50}|2\rangle + \tfrac{1}{50}|3\rangle + \tfrac{2}{50}|4\rangle + \tfrac{4}{50}|5\rangle \\ & + \tfrac{10}{50}|6\rangle + \tfrac{15}{50}|7\rangle + \tfrac{10}{50}|8\rangle + \tfrac{4}{50}|9\rangle + \tfrac{2}{50}|10\rangle. \end{aligned}$$

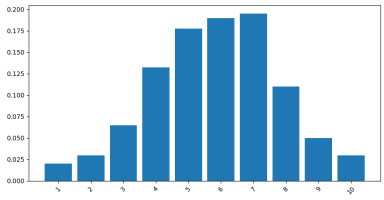


optimistic mood marks

Now that the state $\omega \in \mathcal{D}(X)$ and the channel $c\colon X \twoheadrightarrow Y$ have been described, we can form the transformed state $c_*(\omega) \in \mathcal{D}(Y)$. Following the formulation (2.34) we get for each mark $i \in Y$ the predicted probability:

$$\begin{aligned} c_*(\omega)(i) \;=\; \sum_{x \in X} \omega(x) \cdot c(x)(i) \;&=\; \omega(p) \cdot c(p)(i) + \omega(n) \cdot c(n)(i) + \omega(o) \cdot c(o)(i) \\ &=\; \tfrac{1}{8} \cdot c(p)(i) + \tfrac{3}{8} \cdot c(n)(i) + \tfrac{1}{2} \cdot c(o)(i). \end{aligned}$$

The outcome is in the plot below. It contains a convex combination of the above three distributions (and plots), for $c(p)$, $c(n)$ and $c(o)$, where the weights are determined by the mood distribution ω. This plot contains the 'predicted' marks, corresponding to the state of mind of the teacher.



We view channels $X \rightsquigarrow Y$ as probabilistic computations. Such computations can be composed, both sequentially, via $\circ$, and in parallel, via $\otimes$.

Definition 2.4.4 Let X, Y, Z be arbitrary sets.

1 For two channels $X \overset{c}{\rightsquigarrow} Y$ and $Y \overset{d}{\rightsquigarrow} Z$ we can form the composite channel $d \circ c \colon X \rightsquigarrow Z$ via the monad structure of $\mathcal{D}$.

$$d \circ c \;:=\; \Big(X \xrightarrow{c} \mathcal{D}(Y) \xrightarrow{\mathcal{D}(d)} \mathcal{D}(\mathcal{D}(Z)) \xrightarrow{flat} \mathcal{D}(Z)\Big).$$

More explicitly, there is a formula, for $x \in X$,

$$(d \circ c)(x) \;=\; d_*(c(x)) \;=\; \sum_{z \in Z} \left(\sum_{y \in Y} c(x)(y) \cdot d(y)(z) \right) \big| z \rangle.$$

This is sometimes called the Chapman-Kolmogorov formula, see *e.g.* Fritz (2020).

Channel composition $\circ$ is associative and has unit channels $unit \colon X \rightsquigarrow X$ as identity maps, see Lemma 1.11.3. It gives us a category of probabilistic channels, written as $\mathbf{Chan}(\mathcal{D})$, or simply as $\mathbf{Chan}$.

2 Now let's assume that we have channels $X \overset{c}{\rightsquigarrow} Y$ and $A \overset{d}{\rightsquigarrow} B$. Then we can form the parallel product (or tensor) channel $c \otimes d \colon X \times A \rightsquigarrow Y \times B$ via:

$$c \otimes d \;:=\; \Big(X \times A \xrightarrow{c \times d} \mathcal{D}(Y) \times \mathcal{D}(B) \xrightarrow{\otimes} \mathcal{D}(Y \times B)\Big).$$

This means for elements $x \in X$ and $a \in A$,

$$(c \otimes d)(x, a) \;=\; c(x) \otimes d(a) \;=\; \sum_{y \in Y,\, b \in B} c(x)(y) \cdot d(a)(b) \,\big| y, b \rangle.$$

3 When two channels $c \colon Z \rightsquigarrow X$ and $d \colon Z \rightsquigarrow Y$ have the same domain we can form a *tuple channel* $\langle c, d \rangle \colon Z \rightsquigarrow X \times Y$ as:

$$\langle c, d \rangle \;:=\; (c \otimes d) \circ \Delta \qquad \text{so that} \qquad \langle c, d \rangle(z) \;=\; c(z) \otimes d(z).$$

4 In particular, for a distribution ω on Z and a channel $d \colon Z \to Y$ we define the *graph* as the joint distribution on $Z \times Y$ defined by:

$$\langle id, d \rangle_*(\omega) \qquad \text{so that} \qquad \Big(\langle id, d \rangle_*(\omega)\Big)(y, z) \;=\; \omega(y) \cdot d(y)(z).$$

5 Finally, for a channel $c \colon X \rightsquigarrow Y_1 \times \cdots \times Y_n$ and a mask M of length n we write cM for the channel $x \mapsto c(x)M$. This means that masks are used for channels too, in a pointwise manner.

Thus, the category **Chan** = **Chan**($\mathcal{D}$) of probabilistic channels has sets X as objects and channels $X \multimap Y$ as morphisms / maps. These morphisms can be composed sequentially via $\circ\!\!\!\circ$ and in parallel via $\otimes$. More formally, this means that **Chan** is a symmetric monoidal category. Below, in Subsection 2.5.1, we will introduce a convenient graphical notation for this category.

An ordinary function $f\colon X \to Y$ can be turned into a (probabilistic) channel $\langle f \rangle\colon X \multimap Y$. Explicitly, $\langle f \rangle(x) = 1|f(x)\rangle$. This can be formalised in terms of a functor **Sets** $\to$ **Chan**. Often, we don't write the $\langle - \rangle$ angles when the context makes clear what is meant. For convenience we recall the most basic laws for deterministic (single-valued) channels — see Lemma 1.11.3 (4) — and add two equations for tensors and tuples.

$$\begin{array}{rclcrclcrcl} \langle g \rangle \circ\!\!\!\circ \langle f \rangle & = & \langle g \circ f \rangle & \qquad & c \circ\!\!\!\circ \langle f \rangle & = & c \circ f & \qquad & \langle f \rangle \otimes \langle g \rangle & = & \langle f \times g \rangle \\ \langle f \rangle_*(\omega) & = & \mathcal{D}(f)(\omega) & & g \circ\!\!\!\circ c & = & \mathcal{D}(g) \circ c & & \langle \langle f \rangle, \langle g \rangle \rangle & = & \langle \langle f, g \rangle \rangle. \end{array} \tag{2.36}$$

As a special case we get:

$$unit \otimes unit \;=\; \langle id \rangle \otimes \langle id \rangle \;=\; \langle id \times id \rangle \;=\; \langle id \rangle \;=\; unit. \tag{2.37}$$

We have already seen several examples of probabilistic channels. For instance, the flip and binomial distributions from Example 2.1.2 form channels:

$$[0,1] \xrightarrow{flip} \{0,1\} \qquad \text{and} \qquad [0,1] \xrightarrow{bn[K]} \{0,1,\ldots,K\}$$

Frequentist learning has been described as a function $Flrn\colon \mathcal{M}_*(X) \to \mathcal{D}(X)$ and thus forms a channel, as on the left below. A tensor also forms a channel, as in the middle. Even the identity function $\mathcal{D}(X) \to \mathcal{D}(X)$ can be described as a channel, which we call *sample*, abbreviated to *sam*.

$$\mathcal{M}_*(X) \xrightarrow{Flrn} X \qquad \mathcal{D}(X) \times \mathcal{D}(Y) \xrightarrow{\otimes} X \times Y \qquad \mathcal{D}(X) \xrightarrow{sam} X$$

The coincidence and coupon functions from Subsection 2.3.1 form channels of the form:

$$\mathcal{D}(X) \xrightarrow{cd} SP(K) \qquad\qquad \mathcal{D}(X) \xrightarrow{cpn} \mathcal{P}_{\leq}[K](X)$$

The formalism of channels, including their sequential and parallel composition $\circ\!\!\!\circ$ and $\otimes$, allows us to express basic facts of probability theory. This is an underlying theme in this book. It forms the basis for the area of categorical probability theory.

Remark 2.4.5 In Lemma 2.2.5 we have seen that $Flrn\colon \mathcal{M}_* \Rightarrow \mathcal{D}$ is a natural transformation. Since both $\mathcal{M}_*$ and $\mathcal{D}$ are monads, one can ask if $Flrn$ is also a map of monads. It would mean that $Flrn$ also commutes with the unit and flatten maps, see Definition 1.12.3. This is *not* the case.

It is easy to see that $Flrn$ does commute with the unit maps, simply because $Flrn(1|x\rangle) = 1|x\rangle$. But commutation with flatten's fails. Here is a simple counterexample. Consider the multiset of multisets $\Phi \in \mathcal{M}(\mathcal{M}(\{a,b,c\}))$ given by:

$$\Phi \;:=\; 1\Big|\, 2|a\rangle + 4|c\rangle \,\Big\rangle + 2\Big|\, 1|a\rangle + 1|b\rangle + 1|c\rangle \,\Big\rangle.$$

First flattening the multiset, and then doing frequentist learning gives:

$$Flrn\Big(flat(\Phi)\Big) \;=\; Flrn\Big(4|a\rangle + 2|b\rangle + 6|c\rangle\Big) \;=\; \tfrac{1}{3}|a\rangle + \tfrac{1}{6}|b\rangle + \tfrac{1}{2}|c\rangle.$$

However, first (outer and inner) learning and then flattening the resulting distribution of distributions yields a different outcome:

$$\begin{aligned} \mathit{flat}\Big(\mathit{Flrn}(\mathcal{M}(\mathit{Flrn})(\Phi))\Big) &= \mathit{flat}\Big(\tfrac{1}{3}\Big|\, \tfrac{1}{3}|a\rangle + \tfrac{2}{3}|c\rangle \,\rangle + \tfrac{2}{3}\Big|\, \tfrac{1}{3}|a\rangle + \tfrac{1}{3}|b\rangle + \tfrac{1}{3}|c\rangle \,\rangle\Big) \\ &= \tfrac{1}{3}|a\rangle + \tfrac{2}{9}|b\rangle + \tfrac{4}{9}|c\rangle. \end{aligned}$$

In the next example we show how products, multisets and distributions come together in an elementary combinatorial situation. It shows how the accumulator function from lists to multisets has a probabilistic inverse.

Example 2.4.6 Let X be an arbitrary set and K be a fixed positive natural number. We recall the accumulation function $\mathit{acc}\colon X^K \to \mathcal{N}[K](X)$ mapping lists to natural multisets, via $\mathit{acc}(x_1, \ldots, x_K) = 1|x_1\rangle + \cdots + 1|x_K\rangle$, see (1.39).

In Proposition 1.7.2 we have seen that for a natural multiset $\varphi \in \mathcal{N}[K](X)$ there are $\mathbf{(}\varphi\mathbf{)}$ many lists $\vec{x} \in X^K$ that accumulate to φ, that is, with $\mathit{acc}(\vec{x}) = \varphi$. We are interested in an inverse to accumulation. Which of these $\mathbf{(}\varphi\mathbf{)}$ many lists in $\mathit{acc}^{-1}(\varphi)$ do we pick? The trick is not to make a choice, but to take them all, in a uniform distribution. This gives a channel $\mathcal{N}[K](X) \to \mathcal{D}(X^K)$ in the other direction. We call it *arrangement*, abbreviated as *arr*.

$$\mathit{arr}(\varphi) \coloneqq \sum_{\vec{x} \in \mathit{acc}^{-1}(\varphi)} \frac{1}{\mathbf{(}\varphi\mathbf{)}} \big|\vec{x}\rangle = \sum_{\vec{x} \in \mathit{acc}^{-1}(\varphi)} \frac{\prod_x \varphi(x)!}{K!} \big|\vec{x}\rangle \tag{2.38}$$

For instance, for $X = \{a, b\}$ with multiset $\varphi = 3|a\rangle + 1|b\rangle$ there are $\mathbf{(}\varphi\mathbf{)} = \binom{4}{3,1} = \frac{4!}{3! \cdot 1!} = 4$ arrangements of φ, namely $[a, a, a, b]$, $[a, a, b, a]$, $[a, b, a, a]$, and $[b, a, a, a]$, so that:

$$\mathit{arr}(3|a\rangle + 1|b\rangle) = \tfrac{1}{4}\big|a, a, a, b\rangle + \tfrac{1}{4}\big|a, a, b, a\rangle + \tfrac{1}{4}\big|a, b, a, a\rangle + \tfrac{1}{4}\big|b, a, a, a\rangle.$$

Our next question is: how are accumulation *acc* and arrangement *arr* related? We like to say that arrangement is left-inverse of accumulation, that is: accumulation after arrangement is the identity. Here we hit the problem that accumulation is a function, but arrangement is a channel. What kind of composition should we use, as 'after'? The trick is to turn the accumulation function $\mathit{acc}\colon X^K \to \mathcal{N}[K](X)$ into a (deterministic) channel $\langle \mathit{acc} \rangle\colon X^K \to \mathcal{D}(\mathcal{N}[K](X))$, where $\langle \mathit{acc} \rangle(\vec{x}) = 1\big|\mathit{acc}(\vec{x})\rangle$, and then to compose as channels.

The intended left-inverse property can now be expressed via commutation of the following diagram of channels. Recall that *unit* is the identity channel, which we simply write as id.

$$\begin{tikzcd} \mathcal{N}[K](X) \arrow[r, "\mathit{arr}"] \arrow[dr, equal, "\mathit{id}"'] & X^K \arrow[d, "\langle \mathit{acc} \rangle"] \\ & \mathcal{N}[K](X) \end{tikzcd} \qquad \text{that is:} \qquad \langle \mathit{acc} \rangle \circ \mathit{arr} = \mathit{id}. \tag{2.39}$$

We elaborate the details, for an arbitrary multiset $\varphi \in \mathcal{N}[K](X)$.

$$\begin{aligned}
(\langle acc \rangle \circ arr)(\varphi) &\overset{(2.36)}{=} \mathcal{D}(acc)(arr(\varphi)) \overset{(2.38)}{=} \mathcal{D}(acc)\left(\sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big|\vec{x}\rangle\right) \\
&= \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big|acc(\vec{x})\rangle \\
&= \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big|\varphi\rangle \\
&= 1\big|\varphi\rangle \qquad \text{by Proposition 1.7.2} \\
&= unit(\varphi).
\end{aligned}$$

In the other direction, $arr(acc(\vec{x}))$ does not return $\vec{x}$. It yields a uniform distribution over all permutations of the sequence $\vec{x} \in X^K$, see Section 2.11 for more information.

What we have done in this example can be generalised to a probabilistic inverse of a function h. It will be written as $h^{\sim 1}$. The tilde $\sim$ in the notation $h^{\sim 1}$, at the place where one writes -1 for ordinary inverses, is meant to induce an association with uncertainty.

Definition 2.4.7 Let $h\colon X \twoheadrightarrow Y$ be a surjective function with finite inverse images $h^{-1}(y) = \{x \in X \mid f(x) = y\}$. These subsets are non-empty by surjectivity of h. We define the probabilistic inverse as a channel $h^{\sim 1}\colon Y \to \mathcal{D}(X)$ via uniform distributions:

$$h^{\sim 1}(y) := \sum_{x \in h^{-1}(y)} \frac{1}{|h^{-1}(y)|} \big|x\rangle. \tag{2.40}$$

Then one has: $\langle h \rangle \circ h^{\sim 1} = unit$, precisely as in the previous example.

Thus, the arrangement map is defined in (2.38) as $arr = acc^{\sim 1}$, with an explicit formula for the sizes of inverse images.

Exercises

2.4.1 Consider the sets $X = \{x, y, z\}$ and $Y = \{u, v\}$, with state $\omega = \frac{1}{6}|x\rangle + \frac{1}{2}|y\rangle + \frac{1}{3}|z\rangle \in \mathcal{D}(X)$ and channel $c\colon X \to \mathcal{D}(Y)$ given by:

$$c(x) = \tfrac{1}{2}|u\rangle + \tfrac{1}{2}|v\rangle \qquad c(y) = 1|u\rangle \qquad c(z) = \tfrac{3}{4}|u\rangle + \tfrac{1}{4}|v\rangle.$$

1 Compute the state transformation $c_*(\omega) \in \mathcal{D}(Y)$.

2 Consider a new channel $d\colon \{u, v\} \rightsquigarrow \{1, 2, 3, 4\}$ given by:

$$d(u) = \tfrac{1}{4}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{2}|3\rangle + \tfrac{1}{8}|4\rangle \qquad d(v) = \tfrac{1}{4}|1\rangle + \tfrac{1}{8}|3\rangle + \tfrac{5}{8}|4\rangle.$$

Describe the composite channel $d \circ c\colon \{a, b, c\} \rightsquigarrow \{1, 2, 3, 4\}$ concretely.

2.4.2 Let $dr, tr \in [0, 1]$ two probabilities for 'disease rate' and 'transmission rate'.

1 Suppose that dr is the fraction of the population that has a particular (infectious) disease, say $dr = \frac{1}{20}$. Let $M \in \mathbb{N}$ be the number of people you meet in a day. Argue that $bn[M](dr)$ is the distribution of the number of people with the disease that you meet in a day.

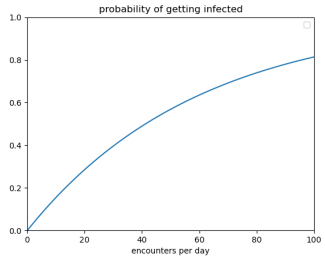
2 We understand tr to be the chance that you get infected after meeting a person with the disease. Let $I = \{i, i^\perp\}$ be the set describing infection and non-infection. The channel $ic[tr]\colon \mathbb{N} \to \mathcal{D}(I)$ gives the infection distribution $ic[tr](n)$ after meeting n people with the disease. Argue that:

$$ic[tr](n) = (1 - (1 - tr)^n)\big|i\rangle + (1 - tr)^n\big|i^\perp\rangle.$$

3 Now compute the infection distribution after meeting $M = 5$ (arbitrary) people on a day, with rates $dr = \frac{1}{20}$ and $tr = \frac{1}{3}$, via pushforward as:

$$ic[tr]_*\Big(bn[M](dr)\Big) = 0.08\big|i\rangle + 0.92\big|i^\perp\rangle.$$

As an aside, with these rates, the curve of infection probablities for $0 \leq M \leq 100$ encounters per day looks as follows.



2.4.3 Consider for the channel c and the distribution ω in the previous exercise the associated matrices:

$$M_c = \begin{pmatrix} \frac{1}{2} & 1 & \frac{3}{4} \\ \frac{1}{2} & 0 & \frac{1}{4} \end{pmatrix} \qquad\qquad M_\omega = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{2} \\ \frac{1}{3} \end{pmatrix}$$

Such matrices are called stochastic, since each of their columns has non-negative entries that add up to one.

1 Check that the matrix associated with the transformed state $c_*(\omega)$ is the matrix-column multiplication $M_c \cdot M_\omega$.

2 Check also that channel composition corresponds to matrix multiplication: $M_{d \circ c} = M_d \cdot M_c$, for the channel d described above.

(A general description appears in Remark 4.5.9.)

2.4.4 For a set X and a number $K \in \mathbb{N}_{>0}$ write:

$$\mathcal{D}[K](X) := \{Flrn(\varphi) \mid \varphi \in \mathcal{N}[K](X)\}.$$

We then write $\mathcal{FD}(X) := \bigcup_K \mathcal{D}[K](X) \subseteq \mathcal{D}(X)$ for the subset of fractional distributions on X. Check that:

1 $\sum_i r_i|x_i\rangle \in \mathcal{FD}(X)$ if and only if $r_i \in [0, 1] \cap \mathbb{Q}$ and $\sum_i r_i = 1$;

2 the mapping $X \mapsto \mathcal{D}[K](X)$ forms a functor **Sets** $\to$ **Sets**;

3 $\mathcal{FD}$ forms a monad on **Sets** in such a way that the inclusion functions $\mathcal{FD}(X) \hookrightarrow \mathcal{D}(X)$ form a map of monads $\mathcal{FD} \Rightarrow \mathcal{D}$.

In Theorem 4.8.11 we shall see that these fractional distributions $\mathcal{FD}(X)$ ly dense in the set $\mathcal{D}(X)$ of all distributions on a set X.

2.4.5 Recall from Proposition 1.8.7 that the set $\mathcal{N}[K](X)$ of K-sized natural multisets over a set X with n elements contains $\left(\!\binom{n}{K}\!\right)$ members. We write $\mathit{unif}_{\mathcal{N}[K](X)} \in \mathcal{D}(\mathcal{N}[K](X))$ for the uniform distribution over such multisets. Note that, or recall from Example 2.1.2 (3), that:

$$\mathit{unif}_{\mathcal{N}[K](X)} \;=\; \sum_{\varphi \in \mathcal{N}[K](n)} \frac{1}{\left(\!\binom{n}{K}\!\right)} \big|\varphi\rangle.$$

Use Lemma 1.8.8 to prove that $\mathit{Flrn}_*(\mathit{unif}_{\mathcal{N}[K](X)}) = \mathit{unif}_X \in \mathcal{D}(X)$.

2.4.6 Recall that the identity function $\mathcal{D}(X) \to \mathcal{D}(X)$ is used as sample channel $\mathit{sam}\colon \mathcal{D}(X) \rightsquigarrow X$. Check that for $\omega \in \mathcal{D}(X)$, $c\colon X \rightsquigarrow Y$, and $\Omega \in \mathcal{D}(\mathcal{D}(X))$ one has:

$$(c \circ \mathit{sam})(\omega) \;=\; c_*(\omega) \qquad \text{and} \qquad \mathit{sam}_*(\Omega) \;=\; \mathit{flat}(\Omega).$$

2.4.7 Recall the mixture map $\mathit{mix}\colon \mathcal{D}(\boldsymbol{n}) \times \mathcal{D}(X)^n \to \mathcal{D}(X)$ from Example 2.3.6. Show that it is essentially state transformation:

$$\mathit{mix}(\omega, c) \;=\; c_*(\omega),$$

where a channel $c\colon \boldsymbol{n} \rightsquigarrow X$ is identified with an n-tuple of distributions $c(0), \ldots, c(n-1) \in \mathcal{D}(X)$.

2.4.8 Let $f\colon X \to Y$ be a function and $c\colon Y \rightsquigarrow Z$ be a channel. Show that for a distribution $\omega \in \mathcal{D}(X)$ the following expressions are all the same.

$$c_*(\mathcal{D}(f)(\omega)) \;=\; c_*(f_*(\omega)) \;=\; (c \circ f)_*(\omega).$$

We have left the brackets $\langle - \rangle$ that promote f to a channel implicit. Fill them in yourself where appropriate.

2.4.9 For an arbitrary set X and a number $K \in \mathbb{N}_{>0}$ define a 'uniform projection' channel $\mathit{unpr}[K]\colon X^K \rightsquigarrow X$ by:

$$\mathit{unpr}[K](\vec{x}) \;:=\; \sum_{1 \leq i \leq K} \tfrac{1}{K} \big|x_i\rangle.$$

This channel is written as $\varepsilon[K]$ in Jacobs (2021c). It can be understood as frequentist learning for sequences, see the triangles below.

1 Prove that $\mathit{unpr}[K]$ is natural in X.

2 Check that the following diagram commutes, for $K, L \geq 1$.

$$\begin{tikzcd}
X^K \times X^L \arrow[r, "\mathit{unpr}[K] \times \mathit{unpr}[L]"] \arrow[d, "+\!+", "\cong"'] & \mathcal{D}(X) \times \mathcal{D}(X) \arrow[d, "\frac{K}{K+L}\cdot(-) + \frac{L}{K+L}\cdot(-)"] \\
X^{K+L} \arrow[r, "\mathit{unpr}[K+L]"] & \mathcal{D}(X)
\end{tikzcd}$$

3 Show that the following two triangles (of channels) commute.

$$\begin{tikzcd}
X^K \arrow[rr, "\mathit{unpr}[K]"] \arrow[dr, "\mathit{acc}"'] & & X \\
& \mathcal{N}[K](X) \arrow[ur, "\mathit{Flrn}"'] &
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
X^K \arrow[rr, "\mathit{unpr}[K]"] & & X \\
& \mathcal{N}[K](X) \arrow[ul, "\mathit{arr}"] \arrow[ur, "\mathit{Flrn}"'] &
\end{tikzcd}$$

4 Let $\omega_1, \ldots, \omega_K \in \mathcal{D}(X)$ be given. Show that:

$$\mathit{unpr}[K]_*(\omega_1 \otimes \cdots \otimes \omega_K) \;=\; \tfrac{1}{K} \cdot \omega_1 + \cdots + \tfrac{1}{K} \cdot \omega_K \;=\; \mathit{flat}\big(\mathit{unpr}[K](\omega_1, \ldots, \omega_K)\big).$$

This means that the rectangle below commutes, via Exercise 2.4.6:

$$\begin{tikzcd}
\mathcal{D}(X)^K \arrow[r, "unpr[K]"] \arrow[d, "\bigotimes"'] & \mathcal{D}(X) \arrow[d, "sam"] \\
X^K \arrow[r, "unpr[K]"] & X
\end{tikzcd}$$

Deduce that $unpr[K]_*(\omega^K) = \omega$.

2.4.10 1 Check that the following two diagrams commute.

$$\begin{tikzcd}
X_1 \times \cdots \times X_n \arrow[r, "\pi_i"] \arrow[d, "c_1 \otimes \cdots \otimes c_n"'] & X_i \arrow[d, "c_i"] \\
Y_1 \times \cdots \times Y_n \arrow[r, "\pi_i"] & Y_i
\end{tikzcd}
\qquad
\begin{tikzcd}
X \arrow[d, "\langle d_1, \ldots, d_n \rangle"'] \arrow[dr, bend left, "d_i"] & \\
Y_1 \times \cdots \times Y_n \arrow[r, "\pi_i"] & Y_i
\end{tikzcd}$$

2 Show that:

$$(c \otimes d)_*(\omega \otimes \rho) \;=\; c_*(\omega) \otimes d_*(\rho).$$

3 Use the previous equation to show:

$$(e \otimes f) \circ (c \otimes d) \;=\; (e \circ c) \otimes (f \circ d).$$

Check that this equation trivialises when written in diagrammatic form.

4 Prove that for a channel $c \colon X \rightsquigarrow Y$ the following diagram commutes.

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "iid"] \arrow[d, "c_*(-)"'] & X^K \arrow[d, "c^K = c \otimes \cdots \otimes c"] \\
\mathcal{D}(Y) \arrow[r, "iid"] & Y^K
\end{tikzcd}$$

5 Show that the following diagram commutes.

$$\begin{tikzcd}
\mathcal{D}(\mathcal{D}(X)) \arrow[r, "iid"] \arrow[d, "flat"'] & \mathcal{D}(X)^K \arrow[d, "\bigotimes"] \\
\mathcal{D}(X) \arrow[r, "iid"] & X^K
\end{tikzcd}
\qquad \text{that is} \qquad
\bigotimes\nolimits_*(iid(\Omega)) \;=\; iid(flat(\Omega)).$$

6 Recall the swap isomorphism $(X^K)^L \xrightarrow{\cong} (X^L)^K$ from Exercise 1.7.18 and prove that:

$$\begin{tikzcd}
 & \mathcal{D}(X^K) \arrow[r, "iid[L]"] & \mathcal{D}\big((X^K)^L\big) \arrow[d, "\cong", "\mathcal{D}(swap)"'] \\
\mathcal{D}(X)^K \arrow[ur, bend left, "\bigotimes[K]"] \arrow[dr, bend right, "iid[L]^K"'] & & \\
 & \mathcal{D}(X^L)^K \arrow[r, "\bigotimes[K]"'] & \mathcal{D}\big((X^L)^K\big)
\end{tikzcd}$$

2.4.11 Show that the big tensor $\bigotimes \colon \mathcal{D}(X)^K \to \mathcal{D}(X^K)$ from (2.28) commutes with unit and flatten, as described below.

$$\begin{tikzcd}
X^K \arrow[d, "unit^K"'] \arrow[dr, bend left, "unit"] & \\
\mathcal{D}(X)^K \arrow[r, "\bigotimes"] & \mathcal{D}(X^K)
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathcal{D}^2(X)^K \arrow[r, "\bigotimes"] \arrow[d, "flat^K"'] & \mathcal{D}(\mathcal{D}(X)^K) \arrow[r, "\mathcal{D}(\bigotimes)"] & \mathcal{D}^2(X^K) \arrow[d, "flat"] \\
\mathcal{D}(X)^K \arrow[rr, "\bigotimes"] & & \mathcal{D}(X^K)
\end{tikzcd}$$

Abstractly this shows that the K-fold tensor functor $(-)^K$ distributes over the monad $\mathcal{D}$.

2.4.12 Check that the functor $\mathcal{D}\colon \mathbf{Act}_M \to \mathbf{Act}_M$ in Diagram (2.8) is a monad, by transferring the unit and flatten map from **Sets** to $\mathbf{Act}_M$.

2.4.13 Recall that we write $sam\colon \mathcal{D}(X) \twoheadrightarrow X$ for the identity function, as channel. Prove that one can write the *iid* and $\bigotimes$ channels from Definition 2.3.4 as:

$$iid \;=\; \langle sam, \ldots, sam\rangle \qquad \text{and} \qquad \bigotimes \;=\; sam \otimes \cdots \otimes sam.$$

2.4.14 Prove that state transformation along $\mathcal{D}$-channels preserves convex combinations (2.9), that is, for $r \in [0, 1]$,

2.4.15 1 Prove that the arrangement map $arr = acc^{\smile 1}$ from Example 2.4.6 is natural: for each function $f\colon X \to Y$ the following diagram commutes.

$$\begin{tikzcd} \mathcal{N}[K](X) \arrow[r, "arr"] \arrow[d, "\mathcal{N}(f)"'] & \mathcal{D}(X^K) \arrow[d, "\mathcal{D}(f^K)"] \\ \mathcal{N}[K](Y) \arrow[r, "arr"] & \mathcal{D}(Y^K) \end{tikzcd}$$

(This is not easy; you may wish to check first what this means in a simple case and then content yourself with a 'proof by example'.)

2 Show that arrangement can be interpreted as a channel $arr\colon \mathcal{N}(X) \twoheadrightarrow \mathcal{L}(X)$ and that it is also natural in this form, without explicit size K.

2.4.16 Recall the match function $mat\colon X^K \to SP(K)$ from (1.23). We assume $K \leq |X|$. Use (1.25) to check that the probabilistic inverse channel $mat^{\smile 1}\colon SP(K) \twoheadrightarrow X^K$ can be described as:

$$mat^{\smile 1}(P) \;=\; \sum_{\vec{x} \in mat^{-1}(P)} \frac{(|X| - |P|)!}{|X|!} \,\big|\vec{x}\rangle.$$

Describe this channel for $K = 3$ and $X = \{a, b, c, d\}$.

2.4.17 Use Lemma 1.4.9 to define a probabilistic inverse for the support function $supp[K]\colon X^K \to \mathcal{P}_{\leq}[K](X)$, where $K \geq 1$.

2.4.18 This exercise is meant to show that probabilistic inverses do not compose. Consider the following commutative triangle of surjective functions with finite inverse images:

$$\begin{tikzcd} \{1,2,3,4\} \arrow[rr, two heads, "h"] \arrow[dr, two heads, "f = g \circ h"'] & & \{u,v,w\} \arrow[dl, two heads, "g"] \\ & \{a,b\} & \end{tikzcd} \qquad \text{where} \qquad \begin{aligned} h(1) &= h(2) = u \\ h(3) &= v \\ h(4) &= w \\ g(u) &= g(v) = a \\ g(v) &= b. \end{aligned}$$

Check the inequality $\neq$ in:

$$h^{\smile 1} \circledcirc g^{\smile 1} \;\neq\; f^{\smile 1} \;=\; (g \circ h)^{\smile 1}.$$

2.4.19 Recall the bijective correspondences from Exercise 1.11.4.

1 Let X, Y be finite sets and $c\colon X \to \mathcal{D}(Y)$ be a $\mathcal{D}$-channel. We can then define an $\mathcal{M}$-channel $c^{\dagger}\colon Y \to \mathcal{M}(X)$ by swapping arguments: $c^{\dagger}(y)(x) = c(x)(y)$. We call c a 'bi-channel' if $c^{\dagger}$ is also a $\mathcal{D}$-channel, *i.e.* if $\sum_x c(x)(y) = 1$ for each $y \in Y$.
Prove that the identity channel is a bi-channel and that bi-channels are closed under composition.

2 Take $A = \{a_0, a_1\}$ and $B = \{b_0, b_1\}$ and define a channel $bell\colon A \times \mathbf{2} \to \mathcal{D}(B \times \mathbf{2})$ as:

$$\begin{array}{rcl} bell(a_0, 0) & = & \frac{1}{2}|b_0, 0\rangle + \qquad\qquad \frac{3}{8}|b_1, 0\rangle + \frac{1}{8}|b_1, 1\rangle \\ bell(a_0, 1) & = & \qquad\qquad \frac{1}{2}|b_0, 1\rangle + \frac{1}{8}|b_1, 0\rangle + \frac{3}{8}|b_1, 1\rangle \\ bell(a_1, 0) & = & \frac{3}{8}|b_0, 0\rangle + \frac{1}{8}|b_0, 1\rangle + \frac{1}{8}|b_1, 0\rangle + \frac{3}{8}|b_1, 1\rangle \\ bell(a_1, 1) & = & \frac{1}{8}|b_0, 0\rangle + \frac{3}{8}|b_0, 1\rangle + \frac{3}{8}|b_1, 0\rangle + \frac{1}{8}|b_1, 1\rangle \end{array}$$

Check that *bell* is a bi-channel.

(It captures the famous Bell table from quantum theory; we have deliberately used open spaces in the above description of the channel *bell* so that non-zero entries align, giving a 'bi-stochastic' matrix, from which one can read $bell^\dagger$ vertically, that is, as transpose.)

3 Show that both first marginals $bell(x, z)[1, 0]$ and $bell^\dagger(y, z)[1, 0]$ are uniform.

2.4.20 Check that the inclusions $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ form a map of monads, as described in Definition 1.12.3.

2.4.21 Let $c\colon X \to \mathcal{D}(Y)$ be a $\mathcal{D}$-channel and $\varphi \in \mathcal{M}(X)$ be a non-empty multiset. Because $\mathcal{D}(Y) \subseteq \mathcal{M}(Y)$ we can also consider c as an $\mathcal{M}$-channel, and use $c_*(\varphi)$. Prove that:

2.4.22 Consider a state $\sigma \in \mathcal{D}(X)$ and an 'endo' channel $c\colon X \twoheadrightarrow X$.

1 Check that the following two statements are equivalent.

- $\langle id, c\rangle_*(\sigma) = \langle c, id\rangle_*(\sigma)$ in $\mathcal{D}(X \times X)$;
- $\sigma(x) \cdot c(x)(y) = \sigma(y) \cdot c(y)(x)$ for all $x, y \in X$.

2 Show that the conditions in the previous item imply that σ is a fixed point for state transformation, that is:

$$c_*(\sigma) = \sigma.$$

Such a σ is also called a stationary state.

2.4.23 Recall the ordinary and multiset Poisson distributions (channels) *pois* and *mpois* from Example 2.1.8 (1) and (2) and recall also the multinomical coefficient distribution *mc* from (2.5). Let X be a finite set and $\lambda \geq 0$ be a rate. Check that $mpois[\lambda] \in \mathcal{D}_\infty(\mathbb{N})$ can be obtained as a pushforward:

$$\Big(mc[X, -]\Big)_*(pois[\lambda]) = mpois[\lambda].$$

2.4.24 In the style of the previous exercise we would like to form the pushforward of the Poisson distribution along the negative binomial distribution $nbn[K](s)$, from Example 2.1.8 (3). A small problem is that this distribution is defined only for $K \geq 1$. We extend it to a channel $nbn[-](s)\colon \mathbb{N} \to \mathcal{D}_\infty(\mathbb{N})$ on all natural numbers by stipulating that $nbn[0](s) = 1|0\rangle$. Show that in this way:

$$\Big(nbn[-](s)\Big)_*(pois[\lambda]) = e^{-\lambda}\big|0\rangle + \sum_{n \geq 1} \sum_{1 \leq K \leq n} e^{-\lambda} \cdot \frac{\lambda^K}{K!} \cdot \binom{n-1}{K-1} \cdot s^K \cdot (1-s)^{n-K}\,\big|n\rangle.$$

This last expression is called the geometric Poisson distribution, see *e.g.* Johnson et al. (2005). The name is somewhat unfortunate, since it is pushforward of Poisson along the negative binomial channel, not along any geometric one (see Exercise 2.1.15).

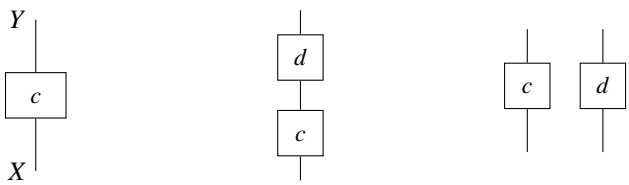
2.5 String diagrams and Bayesian networks

In this book we use commuting diagrams, like in category theory, both for functions and for channels, see for instance Diagram 2.39. For channels we will use a second graphical formalism, called string diagrams Piedeleu and Zanasi (2025); Coecke and Kissinger (2016); Selinger (2011). Such diagrams better display the flows involved. Moreover, string diagrams can be used for equational reasoning. This section first introduces these string diagrams. Then it discusses an alternative graphical formalism that is used in probabilistic reasoning: Bayesian networks, see *e.g.* Pearl (1982, 1988); Lauritzen and Spiegelhalter (1988); Spiegelhalter et al. (1993) for some early references and Larrañaga and Moral (2011) for a short history and overview of applications. We introduce these networks via an example from the literature and show that their so-called conditional probability tables correspond to channels. However, in their orginal graphical representation, Bayesian networks do not reflect sequential and parallel composition of channels. Hence we represent Bayesian networks as string diagrams, following Fong (2012).

2.5.1 A graphical language for channels

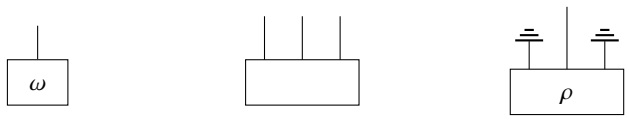
At this stage we introduce string diagrams informally. How they are used will become clear along the way. For more information we refer to the literature, see *e.g.* Selinger (2011); Piedeleu and Zanasi (2025). For usage of string diagrams in probability theory, see Fong (2012); Fritz (2020), for applications in quantum theory see Coecke and Kissinger (2016) and in electrical circuits see Boisseau and Sobociński (2022).

In the language of string diagrams a channel $c\colon X \twoheadrightarrow Y$ is written as a box with two wires, as on the left below, where the flow is from bottom to top. Wires have sets as types. Below on the left the domain and codomain X, Y are included as types of the input and output wire. These types of wires are usually omitted when they can be deduced from the context. In the middle below we see how a composite of two channels $d \circ c$ is written via a sequential connection of the wires of two boxes. On the right we see the graphical notation for the parallel composition $c \otimes d$ of two channels, as boxes.



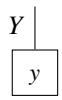
Sometimes we write a function f in such a box. We then implicitly promote it to a deterministic channel $\langle f \rangle$.

A state / distribution is a box without incoming wires, as on the left below. It may have multiple outgoing wires, when it is a joint state, for instance in $\mathcal{D}(X \times Y \times Z)$, as in the middle. On the right we see how marginalisation is described graphically, via the discard symbol $\overline{\top}$. In this case we discard the first and third wire. It corresponds to what we have written via a mask as $\rho[0, 1, 0]$, see Definition 2.3.8, where only the middle output remains.



For each non-empty finite set X there is a uniform distribution $unif_X \in \mathcal{D}(X)$, for which we use special notation, on the left below, as upside-down of discarding.

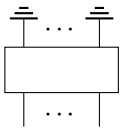
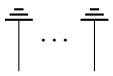
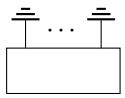
$unif_X \;=$

$\qquad 1|y\rangle \;=$ 

(2.41)

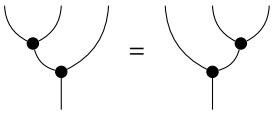
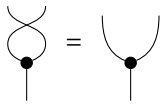
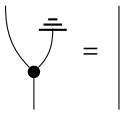
For an element $y \in Y$ of an arbitrary set we simply put y in the box, as on the right, instead of $1|y\rangle$.

There are also equations between string diagrams. When all wires of a channel are discarded, the whole box disappears, as in:

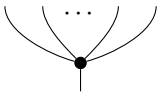

$=$ 
in particular 
$=$

The dashed box on the right indicates 'nothing'.

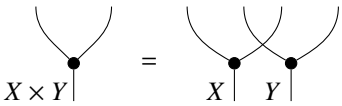
A copy is written as a splitting of wires: . It satisfies a number of diagrammatic equations, expressing associativity, commutativity and a co-unit property:

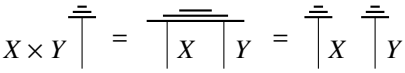

$=$ 
$=$ 
$=$
(2.42)

Because copying is associative, it does not matter which order of copying we use in successive copying. Therefore we feel free to write such n-ary copying as coming from a single node $\bullet$, as in:



In this formalism of string diagrams a single wire $|$ of type $X \times Y$ is the same as two parallel wires $|\;\;|$ of types X and Y respectively. This is notationally sometimes a bit awkward and means that we have to use equations like:





(2.43)

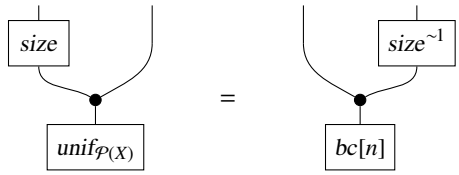
We illustrate the use of string diagrams in the next two theorems that connects several of the probabilistic operations that we have seen so far. The diagrams, with flows going from top to bottom, make clear in an intuitive manner what is connected to what and how the computation proceeds. Later on, we shall recognise that the equations between string diagrams in the two theorems below involve a sufficient statistic situation, see Section 7.7.

Theorem 2.5.1 *Fix a finite set X with $n := |X|$ elements.*

1 *Recall the size-of-subsets function $size\colon \mathcal{P}(X) \to \{0,\dots,n\}$ from Exercise 2.1.3 and consider its probabilistic inverse $size^{\sim 1}\colon \{0,\dots,n\} \to \mathcal{D}(\mathcal{P}(X))$ given by:*

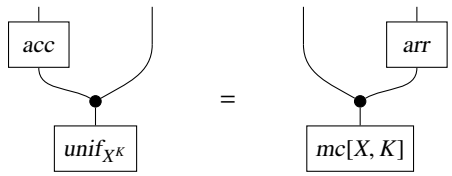
$$size^{\sim 1}(i) \;=\; \sum_{U \subseteq X, |U|=i} \frac{1}{\binom{n}{i}} \big|U\big\rangle.$$

There is then an equality of string diagrams of the form as below, amounting to an equality of joint distributions on $\{0, \ldots, n\} \times \mathcal{P}(X)$.



(2.44)

2 *Recall from Example 2.4.6 the accumulation function* $\mathrm{acc}\colon X^K \to \mathcal{N}[K](X)$ *with its probabilistic inverse 'arrange' channel* $\mathrm{arr}\colon \mathcal{N}[K](X) \rightsquigarrow X^K$. *They can be related via the following equation, between distributions on* $\mathcal{N}[K](X) \times X^K$, *involving the multiset coefficient distribution* $\mathrm{mc}[X, K] \in \mathcal{D}(\mathcal{N}[K](X))$ *from* (2.5).



(2.45)

The last equation (2.45) will be generalised later on, where it appears as special case for the uniform distribution in Theorem 3.3.1 (2). Notice that the equation in Exercise 2.1.3 can be derived from the previous theorem by taking first marginals in (2.44) and (2.45).

Proof. 1 We transform the equation between string diagrams into an equation between joint (tuple) distributions:

$$\begin{aligned}
\langle size, id\rangle_*\big(unif_{\mathcal{P}(X)}\big) &= \sum_{U \subseteq X} \frac{1}{2^n} \,\big|size(U), U\rangle \\
&= \sum_{0 \le i \le n} \; \sum_{U \subseteq X,\, size(U)=i} \frac{1}{\binom{n}{i}} \cdot \frac{\binom{n}{i}}{2^n} \,\big|i, U\rangle \\
&\overset{(2.2)}{=} \sum_{0 \le i \le n} \; \sum_{U \subseteq X,\, size(U)=i} size^{\sim 1}(i)(U) \cdot bc[n](i) \,\big|i, U\rangle \\
&= \langle id, size^{\sim 1}\rangle_*\big(bc[n]\big).
\end{aligned}$$

2 Similarly,

$$\begin{aligned}
\langle acc, id\rangle_*\big(unif_{X^K}\big) &= \sum_{\vec{x} \in X^K} \frac{1}{n^K} \,\big|acc(\vec{x}), \vec{x}\rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \; \sum_{\vec{x} \in X^K,\, acc(\vec{x})=\varphi} \frac{1}{(\varphi)} \cdot \frac{(\varphi)}{n^K} \,\big|\varphi, \vec{x}\rangle \\
&\overset{(2.5)}{=} \sum_{\varphi \in \mathcal{N}[K](X)} \; \sum_{\vec{x} \in X^K,\, acc(\vec{x})=\varphi} acc^{\sim 1}(\varphi)(\vec{x}) \cdot mc[X, K](\varphi) \,\big|\varphi, \vec{x}\rangle \\
&= \langle id, arr\rangle_*\big(mc[X, K]\big).
\end{aligned}$$

❑

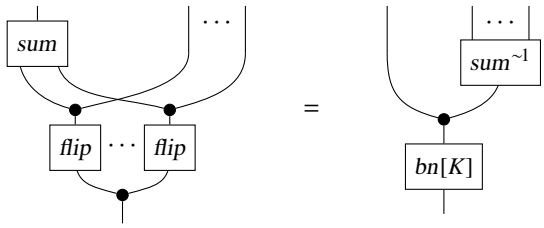
We describe a slightly more complicated situation via a similar string diagram equation.

Theorem 2.5.2 *Consider the K-fold bit-sum map* $sum\colon \mathbf{2}^K \to \{0, \ldots, K\}$ *with* $sum(b_1, \ldots, b_K) := \sum_i b_i$.

1 *The probabilistic inverse* $sum^{\sim 1} \colon \{0, \ldots, K\} \rightsquigarrow \mathbf{2}^K$ *is given by:*

$$sum^{\sim 1}(n) \;=\; \sum_{\vec{b} \in sum^{-1}(n)} \frac{1}{\binom{K}{n}} \left|\vec{b}\right\rangle.$$

2 *There is the following equality of channels* $[0,1] \rightsquigarrow \{0, \ldots, K\} \times \mathbf{2}^K$.



(2.46)

Proof. 1 In order to have K bits b_i sum to $n \in \{0, \ldots, K\}$ precisely n of them need to be 1. There are $\binom{K}{n}$ ways to choose these n bits out of K. Thus, in a uniform distribution $sum^{\sim 1}(n)$ each of these ways gets probability $1/\binom{K}{n}$.

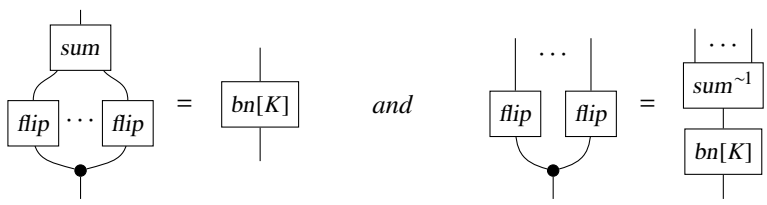
2 We have for $r \in [0,1]$,

$$\begin{aligned}
&\langle sum, id \rangle_*(\langle flip, \ldots, flip \rangle(r)) \\
&= \sum_{\vec{b} \in 2^K} \sum_{n \in \{0,\ldots,K\}} \langle sum \rangle(\vec{b})(n) \cdot flip(r)(b_1) \cdot \ldots \cdot flip(r)(b_K) \left|n, \vec{b}\right\rangle \\
&= \sum_{n \in \{0,\ldots,K\}} \sum_{\vec{b} \in sum^{-1}(n)} r^{b_1} \cdot (1-r)^{1-b_1} \cdot \ldots \cdot r^{b_K} \cdot (1-r)^{1-b_K} \left|n, \vec{b}\right\rangle \\
&= \sum_{n \in \{0,\ldots,K\}} \sum_{\vec{b} \in sum^{-1}(n)} r^{\sum_i b_i} \cdot (1-r)^{K - \sum_i b_i} \left|n, \vec{b}\right\rangle \\
&= \sum_{n \in \{0,\ldots,K\}} \sum_{\vec{b} \in sum^{-1}(n)} \frac{1}{\binom{K}{n}} \cdot \binom{K}{n} \cdot r^n \cdot (1-r)^{K-n} \left|n, \vec{b}\right\rangle \\
&= \sum_{\vec{b} \in 2^K} \sum_{n \in \{0,\ldots,K\}} sum^{\sim 1}(n)(\vec{b}) \cdot bn[K](r)(n) \left|n, \vec{b}\right\rangle \;=\; \langle id, sum^{\sim 1} \rangle_*(bn[K](r)).
\end{aligned}$$

❑

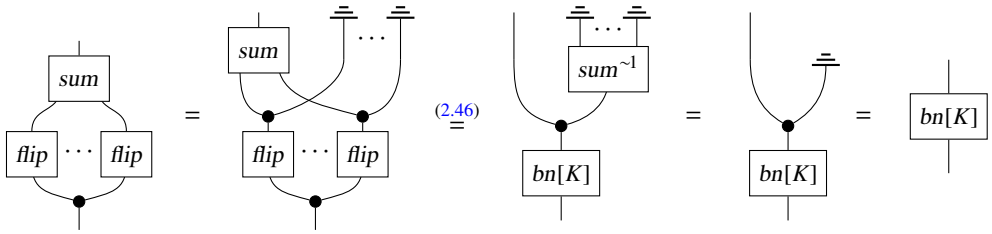
Concentrating only on the left, or only on the right, side of Diagram 2.46 gives interesting equations.

Corollary 2.5.3 *There are the following two equations of channels* $[0,1] \rightsquigarrow \{0, \ldots, K\}$ *and* $[0,1] \rightsquigarrow \mathbf{2}^K$.



Proof. We do only the equation on the left, since the one on the right can be obtained in a similar manner. We give two proofs and first reason diagrammatically by using the rule for the discarding

after a copier, in (2.42), and the rule for discarding of boxes:

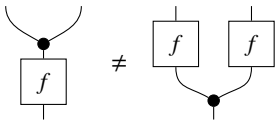


Alternatively, the corresponding equational proof uses Exercise 2.4.10:

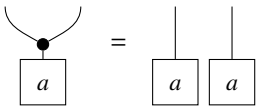
$$\begin{aligned} sum \circ \langle flip, \ldots, flip \rangle &= \pi_1 \circ \langle sum, id \rangle \circ \langle flip, \ldots, flip \rangle \\ &\overset{(2.46)}{=} \pi_1 \circ \langle id, sum^{\sim 1} \rangle \circ bn[K] \;=\; bn[K]. \end{aligned}$$

❑

Remark 2.5.4 We should be careful that boxes can in general *not* be "pulled through" copiers, as expressed below.



This equation *does* hold for deterministic channels, that is for functions f. In fact, such commutation can be used to characterise such deterministic channels, see Exercise 2.5.2. In particular, point distributions $1|a\rangle$, written as box containing a, satisfy:



(2.47)

2.5.2 A Bayesian network example

Bayesian networks are graphical representations of joint distributions and are useful for reasoning: in presence of information about certains nodes in the network, one can draw conclusions about other nodes via probabilistic updating. For instance, a Bayesian network may capture the probabilistic dependencies between phenomena in a particular medical affliction. Given certain observations, the network may give a doctor or lawyer (updated) information about the likelihoods of various scenarios, see *e.g.* et al. (2020); Taylor and Kokshoorn (2023). Indeed, Bayesian networks provide a useful framework for representing expert knowledge and experience, in their graphical layout and in their conditional probability tables.

We shall go deeper into such reasoning in Section 6.4, once probabilistic updating has been introduced. At this stage we introduce, via an example, how Bayesian networks are formulated in the literature. Then we provide a reformulation in terms of string diagrams (see Subsection 2.5.1). The latter representation will be used for prediction, that is for state transformation. The lesson of this section is that Bayesian networks can be understood as configurations of channels and are thus described appropriately via string diagrams. We will show how the conditional probability tables of a Bayesian network correspond to a channel.

The particular Bayesian network that we use as illustration is a standard one, copied from the literature, namely from Darwiche (2009). Bayesian networks were introduced in Pearl (1988),

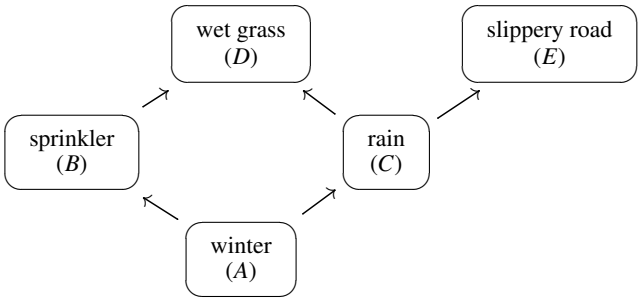


Figure 2.4 The wetness Bayesian network (from (Darwiche, 2009, Chap. 6)), with only the nodes and edges between them; the conditional probability tables associated with the nodes are given separately in the text.

see also Lauritzen (1996); Barber (2012); Bernardo and Smith (2000); Bishop (2006); Jensen and Nielsen (2007); Koller and Friedman (2009); Larrañaga and Moral (2011). They form a popular technique for displaying probabilistic connections and for efficiently presenting joint distributions, without (immediate) space explosion.

Consider the diagram / network in Figure 2.4. It is meant to capture probabilistic dependencies between several wetness phenomena in the oval boxes. For instance, in winter it is more likely to rain (than when it is not winter), and also in winter it is less likely that a sprinkler is on. Still, the grass may be wet by a combination of these occurrences. Whether a road is slippery depends on rain, not on sprinklers.

The letters A, B, C, D, E in this diagram are written exactly as in Darwiche (2009). Here they are not used as sets of Booleans, with inhabitants *true* and *false*, but instead we use these sets with elements:

$$A = \{a, a^\perp\} \qquad B = \{b, b^\perp\} \qquad C = \{c, c^\perp\} \qquad D = \{d, d^\perp\} \qquad E = \{e, e^\perp\}.$$

The notation $a^\perp$ is read as 'not a'. In this way the name of an elements suggests to which set the element belongs.

The diagram in Figure 2.4 becomes a Bayesian network when we provide it with *conditional probability tables*. For the lower three nodes they look as follows.

winter

a	$a^\perp$
$3/5$	$2/5$

sprinkler

	A	b	$b^\perp$
	a	$1/5$	$4/5$
	$a^\perp$	$3/4$	$1/4$

rain

	A	c	$c^\perp$
	a	$4/5$	$1/5$
	$a^\perp$	$1/10$	$9/10$

And for the upper two nodes we have:

wet grass

	B	C	d	$d^\perp$
	b	c	$19/20$	$1/20$
	b	$c^\perp$	$9/10$	$1/10$
	$b^\perp$	c	$4/5$	$1/5$
	$b^\perp$	$c^\perp$	0	1

slippery road

	C	e	$e^\perp$
	c	$7/10$	$3/10$
	$c^\perp$	0	1

This Bayesian network is thus given by nodes, each with a conditional probability table, describing likelihoods in terms of previous 'ancestor' nodes in the network (if any).

How to interpret all this data? How to make it mathematically precise? It is not hard to see that the first 'winter' table describes a probability distribution, abbreviated as wi, on the set $A = \{a, a^\perp\}$, which, in ket notation is:

$$wi \;=\; \tfrac{3}{5}|a\rangle + \tfrac{2}{5}|a^\perp\rangle \;\in\; \mathcal{D}(A).$$

Thus, it is assumed with probability of 60% that we are in a winter situation. This is often called the *prior* distribution, or also the *initial* state.

We move to the above 'sprinkler' table. Note that it contains two distributions on B, one for the element $a \in A$ and one for $a^\perp \in A$. Here we recognise a channel in the description of the table, in this case a channel $A \to \mathcal{D}(B)$. This is a crucial insight! We abbreviate this channel as sp, and define it explicitly as:

$$A \xrightarrow{sp} B \qquad \text{with} \qquad \begin{cases} sp(a) \;=\; \tfrac{1}{5}|b\rangle + \tfrac{4}{5}|b^\perp\rangle \\ sp(a^\perp) \;=\; \tfrac{3}{4}|b\rangle + \tfrac{1}{4}|b^\perp\rangle. \end{cases}$$

We read this channel as: if it is winter, there is a 20% chance that the sprinkler is on, but if it is not winter, there is 75% chance that the sprinkler is on.

Similarly, the 'rain' table corresponds to a channel:

$$A \xrightarrow{ra} C \qquad \text{with} \qquad \begin{cases} ra(a) \;=\; \tfrac{4}{5}|c\rangle + \tfrac{1}{5}|c^\perp\rangle \\ ra(a^\perp) \;=\; \tfrac{1}{10}|c\rangle + \tfrac{9}{10}|c^\perp\rangle. \end{cases}$$

Before continuing, we can see that the formalisation (partial, so far) of the wetness Bayesian network in Figure 2.4 in terms of states and channels, already allows us to do something meaningful, namely state transformation. Indeed, we can form distributions:

$$sp_*(wi) \;\text{ on } B \qquad \text{and} \qquad ra_*(wi) \;\text{ on } C.$$

These (transformed) distributions capture the derived, predicted probabilities that the sprinkler is on, and that it rains. Using the definition of state transformation, see (2.34), we get:

$$\begin{aligned} sp_*(wi)(b) \;=\; \textstyle\sum_x sp(x)(b) \cdot wi(x) \;&=\; sp(a)(b) \cdot wi(a) + sp(a^\perp)(b) \cdot wi(a^\perp) \\ &=\; \tfrac{1}{5} \cdot \tfrac{3}{5} + \tfrac{3}{4} \cdot \tfrac{2}{5} \;=\; \tfrac{21}{50} \\ sp_*(wi)(b^\perp) \;=\; \textstyle\sum_x sp(x)(b^\perp) \cdot wi(x) \;&=\; sp(a)(b^\perp) \cdot wi(a) + sp(a^\perp)(b^\perp) \cdot wi(a^\perp) \\ &=\; \tfrac{4}{5} \cdot \tfrac{3}{5} + \tfrac{1}{4} \cdot \tfrac{2}{5} \;=\; \tfrac{29}{50}. \end{aligned}$$

Thus the overall distribution for the sprinkler (being on or not) is:

$$sp_*(wi) \;=\; \tfrac{21}{50}|b\rangle + \tfrac{29}{50}|b^\perp\rangle.$$

In a similar way one can compute the probability distribution for rain as:

$$ra_*(wi) \;=\; \tfrac{13}{25}|c\rangle + \tfrac{12}{25}|c^\perp\rangle.$$

Such distributions for non-initial nodes of a Bayesian network are called *predictions*. They are obtained via forward state transformation, following the structure of the network.

We still have to translate the upper two nodes of the network from Figure 2.4 into channels. In the conditional probability table for the 'wet grass' node we see 4 distributions on the set D, one

for each combination of elements from the sets B and C. The table thus corresponds to a channel, with a product set $B \times C$ as domain:

$$B \times C \xrightarrow{wg} D \qquad \text{with} \qquad \begin{cases} wg(b,c) &= \frac{19}{20}|d\rangle + \frac{1}{20}|d^\perp\rangle \\ wg(b,c^\perp) &= \frac{9}{10}|d\rangle + \frac{1}{10}|d^\perp\rangle \\ wg(b^\perp,c) &= \frac{4}{5}|d\rangle + \frac{1}{5}|d^\perp\rangle \\ wg(b^\perp,c^\perp) &= 1|d^\perp\rangle. \end{cases}$$

Finally, the table for the 'slippery road' table gives:

$$C \xrightarrow{sr} E \qquad \text{with} \qquad \begin{cases} sr(c) &= \frac{7}{10}|e\rangle + \frac{3}{10}|e^\perp\rangle \\ sr(c^\perp) &= 1|e^\perp\rangle. \end{cases}$$

We illustrate how to obtain predictions for 'rain' and for 'slippery road'. We start from the latter. Looking at the network in Figure 2.4 we see that there are two arrows between the initial node 'winter' and our node of interest 'slippery road'. This means that we have to do two state successive transformations, giving:

$$(sr \circ ra)_*(wi) \;=\; sr_*(ra_*(wi)) \;=\; sr_*\Big(\tfrac{13}{25}|c\rangle + \tfrac{12}{25}|c^\perp\rangle\Big) \;=\; \tfrac{91}{250}|e\rangle + \tfrac{159}{250}|e^\perp\rangle.$$

The first equation follows from Lemma 1.11.3 (3). The second one involves elementary calculations, where we can use the distribution $ra_*(wi)$ that we calculated earlier.

Getting the predicted wet grass probability requires some care. Inspection of the network in Figure 2.4 is of some help, but leads to ambiguity — see below. One might be tempted to form the parallel product $\otimes$ of the predicted distributions for sprinkler and rain, and do state transformation on this product along the wet grass channel wg, as in:

$$wg_*(sp_*(wi) \otimes ra_*(wi)).$$

But this is wrong, since the winter probabilities are now not used consistently, see the different outcomes in the calculations (2.32) and (2.33). The correct way to obtain the wet grass prediction involves copying the winter state, via the copy channel Δ, see:

$$\begin{aligned} (wg \circ (sp \otimes ra) \circ \Delta)_*(wi) \;=\; wg_*\Big((sp \otimes ra_*(\Delta_*(wi))\Big) \;&=\; wg_*(\langle sp, ra\rangle_*(wi)) \\ &=\; \tfrac{1399}{2000}|d\rangle + \tfrac{601}{2000}|d^\perp\rangle. \end{aligned}$$

Such calcutions are laborious, but essentially straightforward. We shall do this in detail, just to see how it works. Especially, it becomes clear that all summations are automatically done at the right place. We proceed in two steps, where for each step we only elaborate the first case.

$$\begin{aligned} \langle sp, ra\rangle_*(wi)(b,c) &= \textstyle\sum_x sp(x)(b) \cdot ra(x)(c) \cdot wi(x) \\ &= sp(a)(b) \cdot ra(a)(c) \cdot \tfrac{3}{5} + sp(a^\perp)(b) \cdot ra(a^\perp)(c) \cdot \tfrac{2}{5} \\ &= \tfrac{1}{5} \cdot \tfrac{4}{5} \cdot \tfrac{3}{5} + \tfrac{3}{4} \cdot \tfrac{1}{10} \cdot \tfrac{2}{5} \;=\; \tfrac{63}{500} \\ \langle sp, ra\rangle_*(wi)(b,c^\perp) &= \cdots \;=\; \tfrac{147}{500} \\ \langle sp, ra\rangle_*(wi)(b^\perp,c) &= \tfrac{197}{500} \\ \langle sp, ra\rangle_*(wi)(b^\perp,c^\perp) &= \tfrac{93}{500}. \end{aligned}$$

We conclude that:

$$(sp \otimes ra)_*(\Delta_*(wi)) \;=\; \langle sp, ra\rangle_*(wi) \;=\; \tfrac{63}{500}|b,c\rangle + \tfrac{147}{500}|b,c^\perp\rangle + \tfrac{197}{500}|b^\perp,c\rangle + \tfrac{93}{500}|b^\perp,c^\perp\rangle.$$

This distribution is used in the next step:

$$\begin{aligned}
\Big(wg_*(\langle sp, ra\rangle_*(wi))\Big)(d) &= \textstyle\sum_{x,y} wg(x,y)(d) \cdot (sp \otimes ra)_*(\Delta_*(wi))(x,y) \\
&= wg(b,c)(d) \cdot \tfrac{63}{500} + wg(b,c^\perp)(d) \cdot \tfrac{147}{500} \\
&\qquad + wg(b^\perp,c)(d) \cdot \tfrac{197}{500} + wg(b^\perp,c^\perp)(d) \cdot \tfrac{93}{500} \\
&= \tfrac{19}{20} \cdot \tfrac{63}{500} + \tfrac{9}{10} \cdot \tfrac{147}{500} + \tfrac{4}{5} \cdot \tfrac{197}{500} + 0 \cdot \tfrac{93}{500} \\
&= \tfrac{1399}{2000} \\
\Big(wg_*(\langle sp, ra\rangle_*(wi))\Big)(d^\perp) &= \tfrac{601}{2000}.
\end{aligned}$$

We have thus shown that the predicted wet grass probability is around 70%:

$$wg_*(\langle sp, ra\rangle_*(wi)) = \tfrac{1399}{2000}|d\rangle + \tfrac{601}{2000}|d^\perp\rangle \approx 0.7|d\rangle + 0.3|d^\perp\rangle.$$

2.5.3 Redrawing Bayesian networks as string diagrams

We have illustrated how prediction computations for Bayesian networks can be done, basically by following the graph structure and translating it into suitable sequential and parallel compositions (written as $\circ$ and $\otimes$) of channels. The match between the graph and the computation is not perfect, and requires some care, especially with respect to copying. Since channels provide a solid semantics, we like to use it in order to improve the network drawing and achieve a better match between the underlying mathematical operations and the graphical representation. String diagrams neatly reflect channels with their sequential and parallel composition. Therefore we prefer to draw Bayesian networks not as in Figure 2.4, but via string diagrams. The switch involves several changes.

- Copying is written explicity in string diagrams, as in the binary case; in general one can have n-ary copying, for $n \geq 2$;
- Wires between nodes in Figure 2.4 represent channels and will thus become boxes in string diagrams.
- The relevant sets / types — like A, B, C, D, E — are not included in the nodes, but are associated with the wires of the string diagram.
- final nodes have outgoing arrows, labeled with their type.

Adapting the original Bayesian network in Figure 2.4 according to these points yields its representation as a string diagram in Figure 2.5. In this way the nodes (with their conditional probability tables) are clearly recognisable as channels, in general of type $A_1 \times \cdots \times A_n \rightsquigarrow B$, where $A_1, \ldots, A_n$ are the types on the incoming wires, and B is the type of the outgoing wire. Initial nodes have no incoming wires, which formally leads to a channel $\mathbf{1} \rightsquigarrow B$, where $\mathbf{1} = \{0\}$ is the empty product. As we have seen, such channels $\mathbf{1} \rightsquigarrow B$ correspond to distributions / states on B. In the adapted diagram one easily forms sequential and parallel compositions of channels, so that it becomes clearer that the wetness prediction is obtained via the expression $(wg \circ (sp \otimes ra) \circ \Delta)_*(wi)$. We may even replace the pushforward $(-)_*$ in the latter expression by $\circ$, when we see the winter distribution also as a channel, of type $wi \colon \mathbf{1} \rightsquigarrow A$. This tightens the match between the string diagram and the computation.

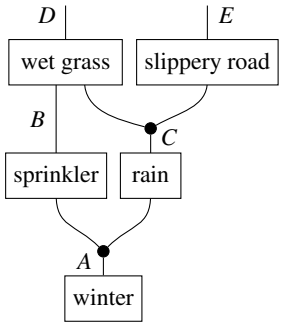


Figure 2.5 The wetness Bayesian network from Figure 2.4 redrawn as a string diagram, reflecting the underlying channel-based semantics, via explicit copiers and typed wires.

Exercises

2.5.1 Consider the probabilistic channel $c\colon X \rightsquigarrow Y$ from Exercise 2.4.1 and show that on the one hand $\Delta \circ f\colon X \rightsquigarrow Y \times Y$ is given by:

$$\begin{aligned} x &\longmapsto \tfrac{1}{2}|u,u\rangle + \tfrac{1}{2}|v,v\rangle \\ y &\longmapsto 1|u,u\rangle \\ z &\longmapsto \tfrac{3}{4}|u,u\rangle + \tfrac{1}{4}|v,v\rangle. \end{aligned}$$

On the other hand, $(c \otimes c) \circ \Delta\colon X \rightsquigarrow Y \times Y$ is described by:

$$\begin{aligned} x &\longmapsto \tfrac{1}{4}|u,u\rangle + \tfrac{1}{4}|u,v\rangle + \tfrac{1}{4}|v,u\rangle + \tfrac{1}{4}|v,v\rangle \\ y &\longmapsto 1|u,u\rangle \\ z &\longmapsto \tfrac{9}{16}|u,u\rangle + \tfrac{3}{16}|u,v\rangle + \tfrac{3}{16}|v,u\rangle + \tfrac{1}{16}|v,v\rangle. \end{aligned}$$

This demonstrates that channels do not commute with copiers, as claimed in Remark 2.5.4.

2.5.2 Deterministic channels are in fact the only ones that commute with copy (see for an abstract setting Carboni and Walters (1987)). Let $c\colon X \rightsquigarrow Y$ commute with diagonals, in the sense that $\Delta \circ c = (c \otimes c) \circ \Delta$. Use Exercise 2.3.13 to prove that c commutes with diagonals if and only if c is deterministic, *i.e.* of the form $c = \langle f \rangle = \mathit{unit} \circ f$ for a function $f\colon X \rightsquigarrow Y$.

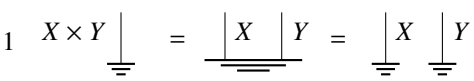
2.5.3 Show that the tuple operation $\langle -, - \rangle$ for channels, described in Definition 2.4.4 (4), satisfies both:

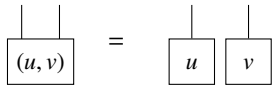
$$\pi_i \circ \langle c_1, c_2 \rangle = c_i \qquad \text{and} \qquad \langle \pi_1, \pi_2 \rangle = \mathit{unit},$$

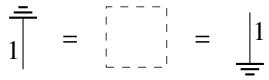
but, essentially as in Exercise 2.5.1:

$$\langle c_1, c_2 \rangle \circ d \neq \langle c_1 \circ d, c_2 \circ d \rangle.$$

2.5.4 Prove that, for finite sets X, Y, and for elements u, v,

1 

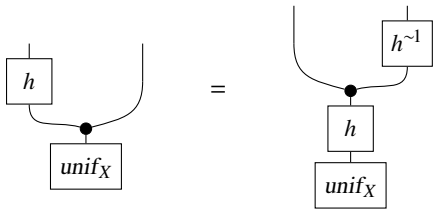
2 

3 

2.5.5 1 Describe the channel in Theorem 2.5.2 (1) in detail for $K = 3$.

2 Add types to all wires in the string diagrams in Theorem 2.5.2 (2).

2.5.6 Let $h\colon X \twoheadrightarrow Y$ be a surjective function, where the domain X is a non-empty finite set. Prove the following equality between distributions on $Y \times X$ given by string diagrams, involving the probabilistic inverse $h^{\sim 1}\colon Y \rightsquigarrow X$ from Definition 2.4.7.



2.5.7 In (Darwiche, 2009, §6.2) the (predicted) joint distribution on $D \times E$ that arises from the Bayesian network example Subsection 2.5.2 is represented as a table. It translates into a joint distribution:

$$\tfrac{30443}{100000}|d, e\rangle + \tfrac{39507}{100000}|d, e^{\perp}\rangle + \tfrac{5957}{100000}|d^{\perp}, e\rangle + \tfrac{24093}{100000}|d^{\perp}, e^{\perp}\rangle.$$

Following the structure of the diagram in Figure 2.5, it is obtained in the present setting as:

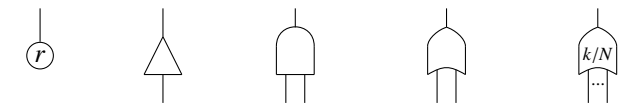
$$\begin{aligned}&\big((wg \otimes sr) \circ (id \otimes \Delta) \circ (sp \otimes ra) \circ \Delta\big)_*(wi)\\ &= (wg \otimes sr)_*((id \otimes \Delta)_*(\langle sp, ra\rangle_*(wi))).\end{aligned}$$

Perform the calculations and check that this expression equals the above distribution.

(Readers may wish to compare the different calculation methods, using sequential and parallel composition of channels — as done here — or using multiplications of tables — as in Darwiche (2009).)

2.5.8 Express the two equations in Exercise 2.4.10 in string diagrammatic form.

2.5.9 Fault trees form a graphical formalism that is used in safety and reliability engineering, see *e.g.* Jimenez et al. (2021) for an overview (or also Piedeleu et al. (2025) for a string diagrammatic approach). There are probabilistic versions of this formalism, which can be understood as instances of (Bayesian networks as) string diagrams. The wires of these trees typically have type $\mathbf{2} = \{1, 0\}$ with basic flip events as states and logical operations (and, or, voting) as boxes, written as:



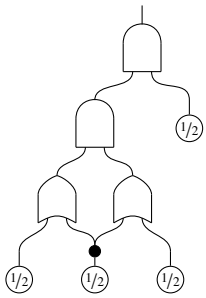
flip(r) not and or voting

As before, $flip(r) = r|1\rangle + (1-r)|0\rangle \in \mathcal{D}(\mathbf{2})$. We interprete these gates as deterministic

negation (not) / conjunction (and) / disjunction (or) channels $nc\colon \mathbf{2} \rightsquigarrow \mathbf{2}$, $ac, oc\colon \mathbf{2}\times\mathbf{2} \rightsquigarrow \mathbf{2}$ and also as an at-least-k-out-of-N voting channel $vc_{k,N}\colon \mathbf{2}^N \rightsquigarrow \mathbf{2}$. Explicitly:

$$\begin{aligned} nc(x) &:= 1\big|\neg x\rangle \\ ac(x,y) &:= 1\big|x \wedge y\rangle \\ oc(x,y) &:= 1\big|x \vee y\rangle \end{aligned} \qquad vc_{k,N}(\vec{x}) := \begin{cases} 1\big|1\rangle & \text{if } sum(\vec{x}) \geq k \\ 1\big|0\rangle & \text{otherwise.} \end{cases}$$

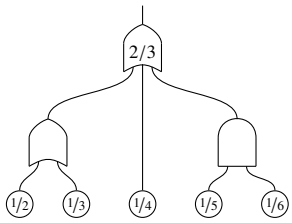
1 Translate the fault tree below, from (Lopuhaä-Zwakenberg, 2024, Fig. 2), into a push-forward expression in terms of (sequential and parallel) composition of channels and states. Then do the prediction (push forward) computation.



$$= \frac{5}{16}|1\rangle + \frac{11}{16}|0\rangle = 0.3125|1\rangle + 0.6875|0\rangle.$$

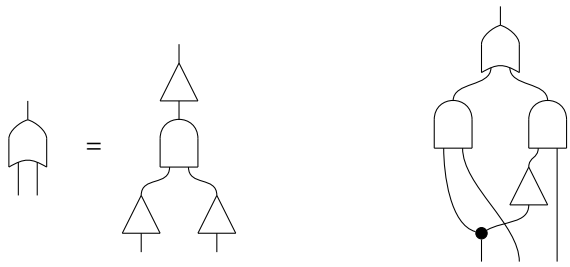
(A representation in ProbLog appears in Exercise 2.8.2.)

2 Show similarly:



$$= \frac{67}{360}|1\rangle + \frac{293}{360}|0\rangle = 0.1861|1\rangle + 0.8139|0\rangle.$$

3 Check that the disjunction channel is the 'De Morgan' dual of conjunction, as described on the left below, and that if-then-else is captured by the diagram on the right.



2.6 Draw distributions

What we call 'draw' distributions capture the probabilities associated with draws of coloured balls from an urn. The proportion of balls of each colour determines these probabilities. Urns filled with coloured balls form an intuitive often-used model in probability theory, like coins and dices. This section provides a first account of these draw distributions in the newly introduced

context of probabilistic channels. The entire next chapter is devoted to draw distributions and provides many more details.

The topic of draw distributions is about the interaction between multisets and distributions. Indeed, an urn with coloured balls forms a (natural) multiset over the set of colours. For instance, an urn with ten balls, three red, two blue and five green, will be described as a multiset $3|R\rangle + 2|B\rangle + 5|G\rangle$ of size 10 over the set $\{R, G, B\}$. Hence the urn is an element of the set $\mathcal{N}[10](\{R, G, B\})$. Similarly, a draw of five balls from this urn, say with two red, two blue and one green, is a multiset $2|R\rangle + 2|B\rangle + 1|G\rangle$. Draw distributions assign probabilities to such multisets of draws (say of size 5), and are thus elements of the set $\mathcal{D}\big(\mathcal{N}[5](\{R, G, B\})\big)$.

For successive draws of multiple balls from the same urn, the following three options are distinguished.

1 **Draw-and-delete**, where the drawn ball is removed from the urn.
2 **Draw-and-replace**, where the drawn ball is returned to the urn, so that each draw is from the same urn;
3 **Draw-and-duplicate**, where the drawn ball is returned to the urn together with an additional ball of the same colour.

These three modes of drawing are called, respectively, *hypergeometric*, *multinomial*, and *Pólya*. A short description uses the numbers -1, 0, or $+1$ to indicate the urn change. Below we devote a seperate subsection to each of these different modes of drawing.

This section contains illustrations and formalisations of these three modes of drawing. It uses a set X where the elements of X are seen as colours. An urn filled with balls whose colours are in X will be represented as a natural multiset υ — the Greek letter upsilon — from the set $\mathcal{N}(X)$. The probability of drawing a ball with colour x from an urn $\upsilon \in \mathcal{N}(X)$ is given by the fraction $\frac{\upsilon(x)}{\|\upsilon\|} = \mathit{Flrn}(\upsilon)(x) \in [0, 1]$ of x-coloured balls in the urn υ. When a ball with colour x is removed or added to an urn $\upsilon \in \mathcal{N}(X)$, we describe the resulting urn as $\upsilon - 1|x\rangle$ or as $\upsilon + 1|x\rangle$.

2.6.1 Hypergeometric draw-and-delete draws

Let us take, for instance, the multiset $\upsilon = 4|R\rangle + 6|B\rangle + 2|G\rangle$ as urn, over the set $X = \{R, B, G\}$ with colours red, green, blue. We are interested in draws of three balls, using the hypergeometric draw-and-delete mode. Such a draw can also be represented as a multiset, namely from the set $\mathcal{N}[3](X)$. We first ask: what is the probability of drawing three red balls. This draw corresponds to the multiset $3|R\rangle \in \mathcal{N}[3](X)$. We reason as follows.

1 The probability of drawing the first red ball from υ is $\mathit{Flrn}(\upsilon)(R) = \frac{\upsilon(R)}{\|\upsilon\|} = \frac{4}{12} = \frac{1}{3}$. The resulting urn is $\upsilon_1 = \upsilon - 1|R\rangle = 3|R\rangle + 6|B\rangle + 2|G\rangle$.
2 The probability of drawing the second red ball, now from $\upsilon_1 = \upsilon - 1|R\rangle$ is given by the same formula: $\mathit{Flrn}(\upsilon_1)(R) = \frac{\upsilon_1(R)}{\|\upsilon_1\|} = \frac{3}{11}$. This results in an urn $\upsilon_2 = \upsilon_1 - 1|R\rangle = 2|R\rangle + 6|B\rangle + 2|G\rangle$.
3 The probability of the third red ball is now $\mathit{Flrn}(\upsilon_2)(R) = \frac{\upsilon_2(R)}{\|\upsilon_2\|} = \frac{2}{10} = \frac{1}{5}$.

The probability that we thus assign to the draw $3|R\rangle$ from υ is $\frac{1}{3} \cdot \frac{3}{11} \cdot \frac{1}{5} = \frac{1}{55}$.

Next we look at a draw $2|R\rangle + 1|B\rangle$, consisting of two red and one blue ball. We can go through the above three steps, for each of the sequences of draws $[R, R, B]$, $[R, B, R]$ and $[B, R, R]$. Successive drawing gives the same probability for each of these three sequences. This probability

has to be multiplied by 3, to take these three orders into account. The resulting probability for the draw $2|R\rangle + 1|B\rangle$ of two red and one blue ball is thus:

$$3 \cdot \frac{4 \cdot 3 \cdot 6}{12 \cdot 11 \cdot 10} = \frac{9}{55}.$$

In this way we can assign a probability to each draw $\varphi \in \mathcal{N}[3](X)$. We then get a distribution over $\mathcal{N}[3](X)$, inhabiting $\mathcal{D}(\mathcal{N}[3](X))$. We now describe the general type and formula of these kind of 'hypergeometric' draws — where drawn balls are deleted.

Definition 2.6.1 Let X be a set (of colours) and $\upsilon \in \mathcal{N}(X)$ be an urn of size $L = \|\upsilon\|$, filled with X-coloured balls. Let $K \in \mathbb{N}$ describe the size of the draws from υ, where we assume $K \leq L = \|\upsilon\|$. This guarantees that the urn contains sufficiently many balls and that we do not 'overdraw'.

The *hypergeometric* distribution $hg[K](\upsilon)$ over $\mathcal{N}[K](X)$ assigns probabilities to K-sized draws $\varphi \leq_K \upsilon$ via the formula:

$$hg[K](\upsilon) := \sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{L}{K}} |\varphi\rangle = \sum_{\varphi \leq_K \upsilon} \frac{\prod_x \binom{\upsilon(x)}{\varphi(x)}}{\binom{L}{K}} |\varphi\rangle. \tag{2.48}$$

The formula (2.48) involves binomial coefficients for numbers and for multisets (see Definition 1.8.1). The probabilities add up to one by Vandermonde's result, see Lemma 1.8.2. The demonstration that this formula is the right one will be provided soon, in Theorem 2.6.2. Without multisets, this formula occurs in (Jaynes, 2003, (3.75)).

First we provide an illustration. The hypergeometric distribution for draws of size 3, that we used above, is:

$$\begin{aligned} & hg[3]\Big(4|R\rangle + 6|B\rangle + 2|G\rangle\Big) \\ &= \tfrac{1}{55}\Big|3|R\rangle\Big\rangle + \tfrac{9}{55}\Big|2|R\rangle + 1|B\rangle\Big\rangle + \tfrac{3}{11}\Big|1|R\rangle + 2|B\rangle\Big\rangle + \tfrac{1}{11}\Big|3|B\rangle\Big\rangle + \tfrac{3}{55}\Big|2|R\rangle + 1|G\rangle\Big\rangle \\ &\quad + \tfrac{12}{55}\Big|1|R\rangle + 1|B\rangle + 1|G\rangle\Big\rangle + \tfrac{3}{22}\Big|2|B\rangle + 1|G\rangle\Big\rangle + \tfrac{1}{55}\Big|1|R\rangle + 2|G\rangle\Big\rangle + \tfrac{3}{110}\Big|1|B\rangle + 2|G\rangle\Big\rangle. \end{aligned}$$

The hypergeometric distribution can be described as a channel of the form:

$$\mathcal{N}[L](X) \xrightarrow{hg[K]} \mathcal{N}[K](X). \tag{2.49}$$

The domain is the set of multisets / urns $\mathcal{N}[L](X)$ of size L. The codomain is the set of distributions over the set of draws $\mathcal{N}[K](X)$ of size K, where we assume $K \leq L$.

The next result demonstrates that the hypergeometric distribution captures the probabilities of successive draw-and-deletes, as illustrated in the beginning of this subsection. We see that the frequentist learning operation *Flrn* plays a role in the formalisation of drawing, see Remark 2.2.1.

Theorem 2.6.2 *Let X be a set of colours and $\upsilon \in \mathcal{N}[L](X)$ be a multiset / urn of size L. For each $K \leq L$ and draw $\varphi \leq_K \upsilon$,*

$$hg[K](\upsilon)(\varphi) = \sum_{\vec{x} \in acc^{-1}(\varphi)} Flrn(\upsilon)(x_1) \cdot Flrn(\upsilon - 1|x_1\rangle)(x_2) \cdot \ldots \cdot Flrn(\upsilon - \textstyle\sum_{i<K} 1|x_i\rangle)(x_K).$$

Proof. Let's write the draw multiset as $\varphi = \sum_j n_j|y_j\rangle \leq_K \upsilon$. Then, for each list $\vec{x} = (x_1, \ldots, x_K) \in$

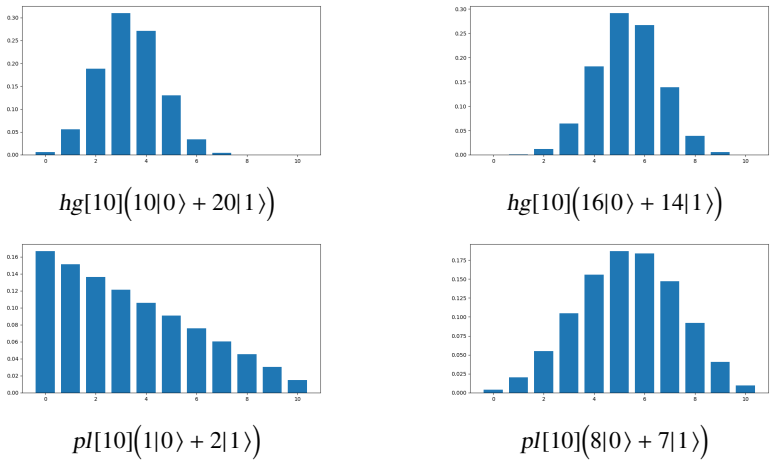


Figure 2.6 Plots of hypergeometric and Pólya distributions, giving the probability of drawing k-many balls of 'colour' 0 in a draw of size 10, see Definitions 2.6.1 and 2.6.3 for details.

X^K with $acc(\vec{x}) = \varphi$, each element y_j in φ occurs n_j times in $\vec{x}$. We first prove the following claim, for $\vec{x} \in acc^{-1}(\varphi)$.

$$\prod_{1 \leq i \leq K} \Big(\upsilon - \textstyle\sum_{j<i} 1|x_j\rangle\Big)(x_i) \;=\; \frac{\upsilon!}{(\upsilon - \varphi)!}. \tag{$*$}$$

The product on the left-hand-side in $(*)$ does not depend on the order of the elements in $\vec{x} \in acc^{-1}(\varphi)$: each element y_j in φ occurs n_j times in this product, with multiplicities $\upsilon(y_j), \upsilon(y_j) - 1, \ldots, \upsilon(y_j) - n_j + 1$, independently of the posititions of the y_j in $\vec{x}$. Thus:

$$\begin{aligned}
\prod_{1 \leq i \leq K} \Big(\upsilon - \textstyle\sum_{j<i} 1|x_j\rangle\Big)(x_i) \;&=\; \prod\nolimits_j \upsilon(y_j) \cdot (\upsilon(y_j) - 1) \cdot \ldots \cdot (\upsilon(y_j) - n_j + 1) \\
&=\; \prod\nolimits_j \upsilon(y_j) \cdot \ldots \cdot (\upsilon(y_j) - \varphi(y_j) + 1) \\
&=\; \prod\nolimits_j \frac{\upsilon(y_j)!}{(\upsilon(y_j) - \varphi(y_j))!} \;=\; \frac{\upsilon!}{(\upsilon - \varphi)!}.
\end{aligned}$$

Now we can finish the proof:

$$
\begin{aligned}
&\sum_{\vec{x}\in acc^{-1}(\varphi)} \prod_{1\le i\le K} Flrn\Big(\upsilon - \textstyle\sum_{j<i} 1|x_j\rangle\Big)(x_i) \\
&= \sum_{\vec{x}\in acc^{-1}(\varphi)} \prod_{1\le i\le K} \frac{(\upsilon - \sum_{j<i} 1|x_j\rangle)(x_i)}{L-i+1} \\
&\overset{(*)}{=} \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{\upsilon!}{(\upsilon-\varphi)!} \cdot \frac{1}{\prod_{1\le i\le K} L-i+1} \\
&= (\varphi) \cdot \frac{\upsilon!}{(\upsilon-\varphi)!} \cdot \frac{(L-K)!}{L!} && \text{by Proposition 1.7.2} \\
&= \frac{\upsilon!}{\varphi! \cdot (\upsilon-\varphi)!} \cdot \frac{K! \cdot (L-K)!}{L!} && \text{by Definition 1.7.1 (5)} \\
&\overset{(1.46)}{=} \frac{\binom{\upsilon}{\varphi}}{\binom{L}{K}} \overset{(2.48)}{=} hg[K](\upsilon)(\varphi).
\end{aligned}
$$

❑

2.6.2 Pólya draw-and-duplicate draws

In the hypergeometric distributions that we have just seen the balls that are drawn disappear from the urn. This seems obvious, certainly from a physical perspective. One can also organise things differently. In so-called Pólya urns one draws a ball, registers its colour, puts it back in the urn, together with a new ball of the same colour. One can envisage a situation where next to the urn at hand there is an infinite supply of balls, of each colour, from which one can take an extra ball to add to the urn. This leads to a new dynamics, where the size of the urn is growing — instead of decreasing, as in the hypergeometric case — and the (successive) drawing can go on indefinitely. Intuitively, the colours that occur often in the urn get reinforced. The additional ball has a strengthening effect that can capture situations with a cluster dynamics, like the spread of contagious diseases Hayhoe et al. (2017), the flow of tourists Lau et al. (2020), or topic classification Blei et al. (2003).

Let's elaborate this dynamics for the urn $\upsilon = 4|R\rangle + 6|B\rangle + 2|G\rangle \in \mathcal{N}[12](X)$, for set of colours $X = \{R, B, G\}$. We look at draws of size three, beginning with a draw $3|R\rangle$ of three red balls.

1 The probability of drawing a single red ball is $Flrn(\upsilon)(R) = \frac{\upsilon(R)}{\|\upsilon\|} = \frac{4}{12} = \frac{1}{3}$. The resulting urn now has an additional red ball: $\upsilon_1 = \upsilon + 1|R\rangle = 5|R\rangle + 6|B\rangle + 2|G\rangle$.
2 The draw of the next red ball happens with probability $Flrn(\upsilon_1)(R) = \frac{\upsilon_1(R)}{\|\upsilon_1\|} = \frac{5}{13}$, giving an urn $\upsilon_2 = \upsilon_1 + 1|R\rangle = 6|R\rangle + 6|B\rangle + 2|G\rangle$.
3 The probability of the third red ball is now $Flrn(\upsilon_2)(R) = \frac{\upsilon_2(R)}{\|\upsilon_2\|} = \frac{6}{14} = \frac{3}{7}$.

The Pólya probability assigned to the multiset draw $3|R\rangle$ is thus $\frac{1}{3} \cdot \frac{5}{13} \cdot \frac{3}{7} = \frac{5}{91}$. The probability assigned to a draw $2|R\rangle + 1|B\rangle$ has to take the three different orders of doing so into account and yields:

$$3 \cdot \frac{4 \cdot 5 \cdot 6}{12 \cdot 13 \cdot 14} = \frac{15}{91}.$$

We now introduce the general formulation.

Definition 2.6.3 Let $\upsilon \in \mathcal{N}[L](X)$ be a non-empty urn over a set X, of size $L = \|\upsilon\| > 0$, with a number $K \in \mathbb{N}$ describing the size of the draws from υ. Since Pólya urns grow in size, we do not have need to put any restrictions on K.

The *Pólya* distribution $pl[K](\upsilon)$ uses the multichoose coefficients from Definitions 1.2.2 and 1.8.4, both for numbers and for multisets.

$$pl[K](\upsilon) = \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} \frac{\binom{\upsilon}{\varphi}}{\binom{\|\upsilon\|}{K}} \big|\varphi\rangle. \tag{2.50}$$

Notice that the draws φ are required to be multisets over $supp(\upsilon)$. This guarantees that $supp(\varphi) \subseteq supp(\upsilon)$, so that we can only draw balls of colours that are present in the urn. In this situation it is most natural to use urns υ with full support, that is with $\upsilon(x) > 0$ for each $x \in X$. In that case the urn contains at least one ball of each colour.

The above Pólya formula (2.50) yields a proper distribution by the multichoose version of Vandermonde, see Proposition 1.8.6. A justification for the Pólya formula appears below, in Theorem 2.6.4.

Here is an example Pólya distribution, with the same urn as in the above explanations, and draws of size 3.

$$\begin{aligned}
&pl[3]\big(4|R\rangle + 6|B\rangle + 2|G\rangle\big) \\
&= \tfrac{5}{91}\Big|3|R\rangle\Big\rangle + \tfrac{15}{91}\Big|2|R\rangle + 1|B\rangle\Big\rangle + \tfrac{3}{13}\Big|1|R\rangle + 2|B\rangle\Big\rangle + \tfrac{2}{13}\Big|3|B\rangle\Big\rangle \\
&\quad + \tfrac{5}{91}\Big|2|R\rangle + 1|G\rangle\Big\rangle + \tfrac{12}{91}\Big|1|R\rangle + 1|B\rangle + 1|G\rangle\Big\rangle + \tfrac{3}{26}\Big|2|B\rangle + 1|G\rangle\Big\rangle \\
&\quad + \tfrac{3}{91}\Big|1|R\rangle + 2|G\rangle\Big\rangle + \tfrac{9}{182}\Big|1|B\rangle + 2|G\rangle\Big\rangle + \tfrac{1}{91}\Big|3|G\rangle\Big\rangle.
\end{aligned}$$

In general, we can describe Pólya distributions as a channel of the form:

$$\mathcal{N}_{fs}(X) \xrightarrow{pl[K]} \mathcal{N}[K](X).$$

For convenience, we assume here that urns are multisets with full support. Implicitly, X is a finite set (of colours). More generally, one can take the set $\mathcal{N}_*(X)$ of non-empty multisets as domain.

Figure 2.6 contains bar plots for 'bivariate' hypergeometric and Pólya distributions, with two colours, so with $X = \mathbf{2} = \{0, 1\}$, and with draws of size 10. These plots show the probabilities for numbers $0 \leq k \leq 10$ in a drawn multiset $k|0\rangle + (10-k)|1\rangle$.

There is an analogue of Theorem 2.6.2 for Pólya, in which drawn colours x are not removed (subtracted) but added to the urn. Frequentist learning is used as in Remark 2.2.1.

Theorem 2.6.4 *Let X be a set of colours and $\upsilon \in \mathcal{N}[L](X)$ be a multiset / urn of size L. For each draw $\varphi \in \mathcal{N}(X)$ with $supp(\varphi) \subseteq supp(\upsilon)$,*

$$pl[K](\upsilon)(\varphi) = \sum_{\vec{x} \in acc^{-1}(\varphi)} Flrn(\upsilon)(x_1) \cdot Flrn(\upsilon + 1|x_1\rangle)(x_2) \cdot \ldots \cdot Flrn(\upsilon + \textstyle\sum_{i<K} 1|x_i\rangle)(x_K).$$

Proof. We write the draw multiset as $\varphi = \sum_j n_j|y_j\rangle$. We now use, for $\vec{x} \in acc^{-1}(\varphi)$, the equation:

$$\prod_{1 \leq i \leq K} \big(\upsilon + \textstyle\sum_{j<i} 1|x_j\rangle\big)(x_i) = \frac{(\upsilon + \varphi - \mathbf{1})!}{(\upsilon - \mathbf{1})!}, \tag{$*$}$$

where $\mathbf{1} = \sum_{x \in supp(\upsilon)} 1|x\rangle$. It holds by the same reasoning as in the proof of Theorem 2.6.2:

$$\begin{aligned}\prod_{1 \leq i \leq K} \Big(\upsilon + \textstyle\sum_{j<i} 1|x_j\rangle\Big)(x_i) &= \prod_j \upsilon(y_j) \cdot (\upsilon(y_j) + 1) \cdot \ldots \cdot (\upsilon(y_j) + \varphi(y_j) - 1) \\ &= \prod_j \frac{(\upsilon(y_j) + \varphi(y_j) - 1)!}{(\upsilon(y_j) - 1)!} = \frac{(\upsilon + \varphi - \mathbf{1})!}{(\upsilon - \mathbf{1})!}.\end{aligned}$$

Now we can derive the Pólya probability distribution:

$$\begin{aligned}\sum_{\vec{x} \in acc^{-1}(\varphi)} \prod_{1 \leq i \leq K} Flrn\Big(\upsilon + \textstyle\sum_{j<i} 1|x_j\rangle\Big)(x_i) &= \sum_{\vec{x} \in acc^{-1}(\varphi)} \prod_{1 \leq i \leq K} \frac{(\upsilon + \sum_{j<i} 1|x_j\rangle)(x_i)}{L + i - 1} \\ &\overset{(*)}{=} \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{(\upsilon + \varphi - \mathbf{1})!}{(\upsilon - \mathbf{1})!} \cdot \frac{1}{\prod_{1 \leq i \leq K} L + i - 1} \\ &= (\varphi) \cdot \frac{(\upsilon + \varphi - \mathbf{1})!}{(\upsilon - \mathbf{1})!} \cdot \frac{(L-1)!}{(L + K - 1)!} \\ &= \frac{(\upsilon + \varphi - \mathbf{1})!}{\varphi! \cdot (\upsilon - \mathbf{1})!} \cdot \frac{K! \cdot (L-1)!}{(L + K - 1)!} \\ &= \frac{\left(\!\binom{\upsilon}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} \qquad \text{by Exercise 1.8.7} \\ &\overset{(2.50)}{=} pl[K](\upsilon)(\varphi).\end{aligned}$$

❑

Remark 2.6.5 We have described the hypergeometric and Pólya distributions with *natural* multisets as urns. In the Pólya case there is a way to formulate the distribution for arbitrary, not necessarily natural, multisets as urns. This works via the gamma funtion Γ, defined as:

$$\Gamma(z) := \int_0^\infty x^{z-1} \cdot e^x \, dx. \tag{2.51}$$

The variable z can be instantiated with any complex number with positive real part. We shall use it for positive real numbers only. For a positive *natural* number $n > 0$ the above gamma function satisfies:

$$\Gamma(n) = (n-1)! \tag{2.52}$$

This property guarantees that the formulation below really is a generalisation of the earlier formulation (2.50) of the Polya distribution.

Let $\upsilon \in \mathcal{M}_{fs}(X)$ be an urn, with full support, say of size $L = \|\upsilon\| \in \mathbb{R}_{>0}$, whose multiplicities $\upsilon(x)$ are thus positive real numbers. We (re)define:

$$pl[K](\upsilon) := \sum_{\varphi \in \mathcal{N}[K](X)} (\varphi) \cdot \frac{\Gamma(L)}{\Gamma(L+K)} \cdot \prod_{x \in X} \frac{\Gamma(\upsilon(x) + \varphi(x))}{\Gamma(\upsilon(x))} \, |\varphi\rangle. \tag{2.53}$$

This forms a distribution because of the equation:

$$\sum_{\varphi \in \mathcal{N}[K](X)} \prod_{x \in X} \frac{\Gamma(\upsilon(x) + \varphi(x))}{\Gamma(\upsilon(x)) \cdot \varphi(x)!} = \frac{\Gamma(L+K)}{\Gamma(L) \cdot K!}.$$

It is a Γ-analogue of Vandermonde's formulas, in Lemma 1.8.2 and Proposition 1.8.6.

Notice that the draws φ in (2.53) are still natural multisets. We shall occasionally use this general Pólya formulation. It is sometimes called the Dirichlet-multinomial distribution, since

it can be obtained via state tranformation of the (continuous) Dirichlet distribution along the multinomial channel, see Proposition **??**.

2.6.3 Multinomial draw-and-replace draws

We have seen that the hypergeometric distribution is based on a draw-and-delete mode, whereas Pólya uses a draw-and-duplicate mode. There is a third, intermediate draw-and-replace mode, where the drawn ball is restored to the urn (after inspection), and assumed to be mixed-in randomly. In this case the urn remains unchanged. The resulting distributions are called multinomial.

In this multinomial case, since the urn remains unchanged, we may describe it more abstractly as a probability distribution $\omega \in \mathcal{D}(X)$. The probability of drawing a ball with colour x is then simply given by the associated probability $\omega(x) \in [0, 1]$. Drawing an x-ball three times has probability $\omega(x)^3$. This is assigned to the multiset $3|x\rangle$. And the draw $2|x\rangle + 1|y\rangle$, for $x \neq y$, will get probability $3 \cdot \omega(x)^2 \cdot \omega(y)$. The number 3 appears since we have to take the three possible orders into account.

Definition 2.6.6 Let $\omega \in \mathcal{D}(X)$ be a distribution, representing an abstract urn. The multinomial distribution $mn[K](\omega)$ on $\mathcal{N}[K](X)$ is defined as:

$$mn[K](\omega) \;:=\; \sum_{\varphi \in \mathcal{N}[K](X)} \mathbf{(}\varphi\mathbf{)} \cdot \omega^{\varphi} \,|\varphi\rangle \;=\; \sum_{\varphi \in \mathcal{N}[K](X)} \mathbf{(}\varphi\mathbf{)} \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \,|\varphi\rangle. \tag{2.54}$$

For a channel $c\colon X \to \mathcal{D}(Y)$ we sometimes write more generally $mn[K](c) := mn[K] \circ c\colon X \to \mathcal{D}(\mathcal{N}[K](Y))$. This general pointwise form of multinomial, for a channel c, arises naturally in an axiomatic setting Jacobs (2021c).

We recall that the number $\mathbf{(}\varphi\mathbf{)}$ is the multiset coefficient $\frac{K!}{\prod_x \varphi(x)!}$, see Definition 1.7.1 (5). It takes care of the different lists of single draws that accumulate to φ, see Proposition 1.7.2. In Subsection 2.2.1 we have described a single sampling $x \leftarrow \omega$ from a distribution ω. The multinomial distribution $mn[K](\omega)$ gives the probabilities associated with K-fold sampling $\varphi \leftarrow \omega$ of a multiset φ of size K from ω, obtained by K-many single samples.

The Multinomial Theorem (1.43) ensures that the probabilities in (2.54) add up to one:

$$\sum_{\varphi \in \mathcal{N}[K](X)} \mathbf{(}\varphi\mathbf{)} \cdot \prod\nolimits_x \omega(x)^{\varphi(x)} \overset{(1.43)}{=} \Big(\textstyle\sum_x \omega(x)\Big)^K \;=\; 1^K \;=\; 1.$$

For space $X = \{R, B, G\}$ and urn $\omega = \frac{1}{3}|R\rangle + \frac{1}{2}|B\rangle + \frac{1}{6}|G\rangle$ the draws of size 3 form a distribution of the form:

$$\begin{aligned} mn[3](\omega) \;=\; & \tfrac{1}{27}\Big|3|R\rangle\Big\rangle + \tfrac{1}{6}\Big|2|R\rangle + 1|B\rangle\Big\rangle + \tfrac{1}{4}\Big|1|R\rangle + 2|B\rangle\Big\rangle + \tfrac{1}{8}\Big|3|B\rangle\Big\rangle \\ & + \tfrac{1}{18}\Big|2|R\rangle + 1|G\rangle\Big\rangle + \tfrac{1}{6}\Big|1|R\rangle + 1|B\rangle + 1|G\rangle\Big\rangle + \tfrac{1}{216}\Big|3|G\rangle\Big\rangle \\ & + \tfrac{1}{36}\Big|1|R\rangle + 2|G\rangle\Big\rangle + \tfrac{1}{24}\Big|1|B\rangle + 2|G\rangle\Big\rangle + \tfrac{1}{8}\Big|2|B\rangle + 1|G\rangle\Big\rangle. \end{aligned}$$

In general, we can describe multinomial distributions as a channel:

$$\mathcal{D}(X) \xrightarrow{\;mn[K]\;} \mathcal{N}[K](X).$$

In the end, we note that the draw multisets φ in (2.54) can be restricted to those with $supp(\varphi) \subseteq$

$supp(\omega)$. Indeed, if $\omega(x) = 0$, but $\varphi(x) \neq 0$, for some $x \in X$, then $\omega(x)^{\varphi(x)} = 0$, so that the whole product $\prod$ becomes zero, and so that φ does not contribute to the above multinomial distribution.

There is an analogue of Theorem 2.6.2 and 2.6.4 for multinomial distributions. For a distribution $\omega \in \mathcal{D}(X)$, considered as urn, the product distribution $iid[K](\omega) = \omega^K \in \mathcal{D}(X^K)$ assigns probablities $\omega^K(\vec{x})$ to K successive draws $\vec{x} \in X^K$, as lists, in which the order of the individual draws is taken into account. When we wish to ignore the order of the elements in a draw, we think of a draw (of size K) as a multiset over X, in $\mathcal{N}[K](X)$. We can obtain the probability of such draws via pushforward, see Lemma 2.1.3, along the accumulation function $acc\colon X^K \to \mathcal{M}[K](X)$ from (1.38). In this way we recover the multinomial distribution, in analogy with coupon and coincidence distributions from Definition 2.3.7.

Theorem 2.6.7 *The K-draw multinomial distribution arises by pushforward along accumulation of a K-fold product:*

$$mn[K](\omega) = \mathcal{D}(acc)\Big(iid[K](\omega)\Big) = \mathcal{D}(acc)\Big(\omega^K\Big). \tag{2.55}$$

Proof. By unfolding the relevant definitions:

$$\begin{aligned}
\mathcal{D}(acc)(iid[K](\omega)) &= \sum_{\vec{x}\in X^K} \omega^K(\vec{x})\,\big|acc(\vec{x})\big\rangle \\
&\overset{(*)}{=} \sum_{\vec{x}\in X^K} \prod_{y\in X} \omega(y)^{acc(\vec{x})(y)}\,\big|acc(\vec{x})\big\rangle \\
&= \sum_{\varphi\in\mathcal{M}[K](X)} \sum_{\vec{x}\in acc^{-1}(\varphi)} \prod_{y\in X} \omega(y)^{acc(\vec{x})(y)}\,\big|acc(\vec{x})\big\rangle \\
&= \sum_{\varphi\in\mathcal{M}[K](X)} \sum_{\vec{x}\in acc^{-1}(\varphi)} \prod_{y\in X} \omega(y)^{\varphi(y)}\,\big|\varphi\big\rangle \\
&= \sum_{\varphi\in\mathcal{M}[K](X)} \mathbf{(}\varphi\mathbf{)} \cdot \prod_{y\in X} \omega(x)^{\varphi(y)}\,\big|\varphi\big\rangle \quad \text{by Proposition 1.7.2} \\
&= mn[K](\omega).
\end{aligned}$$

The marked equation $\overset{(*)}{=}$ is the crucial step:

$$\begin{aligned}
\omega^K(\vec{x}) &= \omega(x_1)\cdot\ldots\cdot\omega(x_K) \\
&= \prod_{y\in X} \omega(y)^{n_y} \quad \text{where } n_y \text{ is the number of occurrences of } y \text{ in } \vec{x} \\
&= \prod_{y\in X} \omega(y)^{\varphi(y)} \quad \text{where } \varphi = acc(\vec{x}).
\end{aligned}$$

❑

The literature sometimes distinguishes 'bivariate' and 'multivariate' draw distributions. In the bivariate case there are only two colours (of balls), corresponding to having $\mathbf{2} = \{0, 1\}$ as set of colours. The multivariate case involves multiple (more than two) colours. In this book we typically work with the general multivariate form, and treat the bivariate situation as a special case.

In Example 2.1.2 (2) we have seen the binomial distribution. It is this special bivariate case of the multinomial distribution, as we shall see, after a bit of massaging. The two results below make this precise.

Lemma 2.6.8 *The binomial distribution $bn[K]\colon [0, 1] \to \mathcal{D}(\{0, \ldots, K\})$ is a bivariate version*

of the multinomial distribution $\mathrm{mn}[K]\colon \mathcal{D}(X) \to \mathcal{D}(\mathcal{N}[K](X))$*, namely for* $X = \mathbf{2} = \{0,1\}$*, via the following two isomorphisms.*

- *The unit interval* $[0,1]$ *is isomorphic to the set of distributions* $\mathcal{D}(\mathbf{2})$ *on the two-element set* $\mathbf{2}$*, via the flip function from Example 2.1.2 (1), given by* $\mathit{flip}(r) = r|1\rangle + (1-r)|0\rangle$*.*
- *The set* $\{0,1,\ldots,K\}$ *of the first* $K+1 = \binom{\binom{2}{K}}{}$ *natural numbers is isomorphic to the set of natural multisets* $\mathcal{N}[K](\mathbf{2})$ *of size* K*, over the two-element set* $\mathbf{2}$*. This works via the isomorphism* $\mathit{ones}(k) = k|1\rangle + (K-k)|0\rangle$*, of type* $\mathit{ones}\colon \{0,1,\ldots,K\} \xrightarrow{\cong} \mathcal{N}[K](\mathbf{2})$*.*

Via these two isomorphisms we get a commuting diagram that connects binomial and multinomial distributions:

$$\begin{tikzcd} {[0,1]} \arrow[r, "bn[K]"] \arrow[d, "flip"', "\cong"] & \mathcal{D}(\{0,\ldots,K\}) \arrow[d, "\cong"', "\mathcal{D}(ones)"] \\ \mathcal{D}(\mathbf{2}) \arrow[r, "mn[K]"] & \mathcal{D}(\mathcal{N}[K](\mathbf{2})) \end{tikzcd} \tag{2.56}$$

Proof. For $r \in [0,1]$,

$$\begin{aligned} (mn[K] \circ flip)(r) &= \sum_{\varphi \in \mathcal{N}[K](\mathbf{2})} (\varphi) \cdot \prod_{i \in \mathbf{2}} flip(r)(i)^{\varphi(i)} \,\big|\varphi\rangle \\ &= \sum_{k \in \{0,\ldots,K\}} \frac{K!}{k! \cdot (K-k)!} \cdot r^k \cdot (1-r)^{K-k} \,\Big|\, k|1\rangle + (K-k)|0\rangle \Big\rangle \\ &= \sum_{k \in \{0,\ldots,K\}} bn[K](r)(k) \,\big|ones(k)\rangle \;=\; (\mathcal{D}(ones) \circ bn[K])(r). \end{aligned}$$

❑

Via stick breaking one can express multinomial distributions in terms of multiple binomial distributions. This is a folklore result that occurs for instance in Jacobs (2022d).

Proposition 2.6.9 *For a sequence of probabilities* $\vec{r} = (r_0,\ldots,r_{n-2}) \in (0,1)^{n-1}$ *of length* $n > 1$*, and for a draw multiset* $\varphi \in \mathcal{N}[K](\boldsymbol{n})$ *over* $\boldsymbol{n} = \{0,\ldots,n-1\}$*,*

$$\begin{aligned} mn[K](stbr(\vec{r}))(\varphi) \;=\; & bn[K](r_0)(\varphi(0)) \cdot bn[K-\varphi(0)](r_1)(\varphi(1)) \cdot bn[K-\varphi(0)-\varphi(1)](r_2)(\varphi(2)) \\ & \cdot \ldots \cdot\; bn[K-\textstyle\sum_{i<n-2}\varphi(i)](\varphi(n-2)). \end{aligned}$$

Proof. We use induction on $n > 1$. When $n = 2$ one has $stbr(r_0) = r_0|0\rangle + (1-r_0)|1\rangle$, so that for $\varphi \in \mathcal{N}[K](\mathbf{2})$,

$$mn[K](stbr(r_0))(\varphi) \;=\; \binom{K}{\varphi(0)} \cdot r_0^{\varphi(0)} \cdot (1-r_0)^{\varphi(1)} \;=\; bn[K](r_0)(\varphi(0)).$$

Next, let $\varphi = \sum_{i \le n} k_i|i\rangle \in \mathcal{M}[K](\boldsymbol{n+1})$ and $\vec{r} = r_0,\ldots,r_{n-1} \in (0,1)^n$ be given. We use a shifted multiset $\varphi' = \sum_{i<n-1} k_{i+1}|i\rangle$ of size $K-k_0$. Then:

$$\begin{aligned} & bn[K](r_0)(k_0) \cdot bn[K-k_0](r_1)(k_1) \cdot \ldots \cdot bn[K-\textstyle\sum_{i<n-1} k_i](r_{n-1})(k_{n-1}) \\ &\overset{\text{(IH)}}{=} bn[K](r_0)(k_0) \cdot mn[K-k_0](stbr(r_1,\ldots,r_{n-1}))(\varphi') \\ &= \binom{K}{k_0} \cdot r_0^{k_0} \cdot (1-r_0)^{K-k_0} \cdot (\varphi') \cdot \prod_{i>0} stbr(r_1,\ldots,r_{n-1})(i)^{k_i} \\ &= \frac{K!}{k_0! \cdot (K-k_0)!} \cdot \frac{(K-k_0)!}{k_1! \cdots k_{n-1}!} \cdot r_0^{k_0} \cdot \prod_{i>0} \Big(stbr(r_1,\ldots,r_{n-1})(i) \cdot (1-r_0)\Big)^{k_i} \\ &= (\varphi) \cdot \textstyle\prod_{i \ge 0} stbr(r_0,\ldots,r_{n-1})(i)^{k_i} \;=\; mn[K](stbr(\vec{r}))(\varphi). \end{aligned}$$

❑

Exercises

2.6.1 Let $X = \{R, B, G\}$ with draw $\varphi = 2|R\rangle + 3|B\rangle + 2|G\rangle \in \mathcal{N}[7](X)$.

1 Consider distribution $\omega = \frac{1}{3}|R\rangle + \frac{1}{2}|B\rangle + \frac{1}{6}|G\rangle$ and show that:

$$mn[7](\omega)(\varphi) \;=\; \tfrac{35}{432}.$$

Explain this outcome in terms of iterated single draws.

2 Consider urn $\upsilon = 4|R\rangle + 6|B\rangle + 2|G\rangle \in \mathcal{N}[12](X)$ and compute:

$$hg[7](\upsilon)(\varphi) \;=\; \tfrac{5}{33}.$$

3 Check that the Pólya probability of the same draw from the same urn is given by:

$$pl[7](\upsilon)(\varphi) \;=\; \tfrac{35}{663}.$$

2.6.2 Let X be a non-empty finite set with n elements, carrying the uniform distribution $unif_X = \sum_{x \in X} \frac{1}{n}|x\rangle$. Show that:

$$mn[K](unif_X) \;=\; \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\binom{\varphi}{}}{n^K}\,|\varphi\rangle.$$

Describe this distribution for $X = \{R, B, G\}$ and $K = 4$. Relate it to Exercise 1.7.9.

2.6.3 Let X be a finite set with $n \geq 1$ elements. Write $\mathbf{1} = \sum_{x \in X} 1|x\rangle$ for the multiset of singletons over X. Show that Pólya of this singleton multiset is the uniform distribution on $\mathcal{N}[K](X)$, that is:

$$pl[K](\mathbf{1}) \;=\; unif_{\mathcal{N}[K](X)} \;=\; \sum_{\varphi \in \mathcal{N}[K](X)} \frac{1}{\binom{\binom{n}{K}}{}}\,|\varphi\rangle.$$

2.6.4 Let υ be a natural multiset. Use (2.52) to show that the two Pólya formulations (2.50) and (2.53) coincide.

2.6.5 Consider the multinomial computation from Exercise 2.6.1, with outcome $\frac{35}{432}$. Obtain this same outcome via successive binomials, following Proposition 2.6.9.

2.6.6 Check that:

$$mn[1](\omega) \;=\; \sum_{x \in supp(\omega)} \omega(x)\,\big|\,1|x\rangle\rangle.$$

2.6.7 Use the formulation of the multinomial distribution in Theorem 2.6.7 together with the naturality of accumulation, in Exercise 1.7.15, and of iid, in Lemma 2.3.5 (1), to obtain the naturality of the multinomial channel: for each function $f\colon X \to Y$ the following diagram commutes.

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "mn[K]"] \arrow[d, "\mathcal{D}(f)"'] & \mathcal{D}(\mathcal{N}[K](X)) \arrow[d, "\mathcal{D}(\mathcal{N}(f))"] \\
\mathcal{D}(X) \arrow[r, "mn[K]"] & \mathcal{D}(\mathcal{N}[K](X))
\end{tikzcd}$$

2.6.8 For an element $x \in X$ consider the 'evaluation at x' function $ev_x\colon \mathcal{N}[K](X) \to \{0, 1, \ldots, K\}$, for $K \in \mathbb{N}$, given by $ev_x(\varphi) = \varphi(x)$. Show that for a distribution $\omega \in \mathcal{D}(X)$,

$$\mathcal{D}(ev_x)\big(mn[K](\omega)\big) \;=\; bn[K](\omega(x)).$$

2.6.9 Let X be a non-empty finiset set of size $N = |X| > 0$ and let $K \in \mathbb{N}_{>0}$. We consider the uniform, multiset coefficient, and support-size distributions on $\mathcal{N}[K](X)$, defined as:

$$\begin{aligned} ums[X,K] &:= \sum_{\varphi \in \mathcal{N}[K](X)} \frac{1}{\left(\!\binom{N}{K}\!\right)} \big|\varphi\rangle \\ mc[X,K] &:= \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\binom{\varphi}{}}{N^K} \big|\varphi\rangle \\ ssms[X,K] &:= \sum_{\varphi \in \mathcal{N}[K](X)} \frac{|supp(\varphi)|}{N \cdot \left(\!\binom{N}{K-1}\!\right)} \big|\varphi\rangle. \end{aligned} \tag{2.57}$$

The multiset coefficient distribution already occurred in (2.5). These distributions are well-defined by, respectively, Proposition 1.8.7 and Exercises 1.7.9 and 1.8.15. The last two of these distributions arise in Jacobs and Széles (2025) as stationary distributions for certain Markov chains on $\mathcal{N}[K](X)$.

Prove that for each $x \in X$, using the evaluation function from the previous exercise,

$$\begin{aligned} \mathcal{D}(ev_x)\big(ums[X,K]\big) &= \sum_{0 \le i \le K} \frac{\left(\!\binom{N-1}{K-i}\!\right)}{\left(\!\binom{N}{K}\!\right)} \big|i\rangle \\ \mathcal{D}(ev_x)\big(mc[X,K]\big) &= bn[K](\tfrac{1}{N}) \\ \mathcal{D}(ev_x)\big(ssms[X,K]\big) &= \frac{(N-1) \cdot \left(\!\binom{N-1}{K-1}\!\right)}{N \cdot \left(\!\binom{N}{K-1}\!\right)} \big|0\rangle + \sum_{1 \le i \le K} \frac{\left(\!\binom{N-1}{K-i}\!\right) + (N-1) \cdot \left(\!\binom{N-1}{K-i-1}\!\right)}{N \cdot \left(\!\binom{N}{K-1}\!\right)} \big|i\rangle. \end{aligned} \tag{2.58}$$

Notice that the distributions on the right-hand-side do not depend on the element $x \in X$. Exercise 4.2.2 states that their 'means' are all equal to $\frac{K}{N}$.

2.6.10 Show that one can obtain the coupon distribution from Definition 2.3.7 via pushforward of the multionial distribution, *i.e.* show that:

$$\mathcal{D}(supp)(mn[K](\omega)) = cpn[K](\omega).$$

2.6.11 Can you generalise Exercise 2.1.11 to numbers $r_1, \ldots, r_n$ and set partitions of X of size n?

2.6.12 Define yourself 'sequence' versions of the hypergeometric and Pólya distributions, forming channels:

$$\mathcal{N}[L](X) \xrightarrow{seqhg[K]} X^K \xleftarrow{seqpl[K]} \mathcal{N}_*(X)$$

where X is finite and $K \le L$. Do this in such a way that:

$$acc \circ seqhg[K] = hg[K] \qquad \text{and} \qquad acc \circ seqpl[K] = pl[K].$$

Hint: Use the formulations of Theorem 2.6.2 and 2.6.4, or more abstractly, the arrangement channel from (2.38).

2.6.13 In Dunn (1967) a 'bivariate negative binomial' distribution is introduced on $\mathbb{N}^2 \cong \mathcal{N}(\mathbf{2})$. It is generalised to the multivariate case below. Alternative negative generalisations of multinomial distributions (and hypergeometric and Pólya), as infinite distributions on $\mathbb{N}$, occur in Jacobs (2022e).

Let $\omega \in \mathcal{D}(X)$ be a distribution with a special 'success element' $s \in supp(\omega) \subseteq X$ and

a 'success number' $L \in \mathbb{N}_{>0}$. We write $X \setminus s = X \setminus \{s\}$ for the set X without the element $s \in X$. We define (one form of) the negative multinomial $nmn[s, L](\omega) \in \mathcal{D}_\infty(\mathcal{N}(X))$ as:

$$nmn[s, L](\omega) \coloneqq \sum_{\varphi \in \mathcal{N}(X \setminus s)} \omega(s)^L \cdot \left(\!\!\binom{L}{\|\varphi\|}\!\!\right) \cdot (\varphi) \cdot \prod_{x \neq s} \omega(x)^{\varphi(x)} \big| \varphi + L|s\rangle\rangle.$$

(One can discuss whether the summand $L|s\rangle$ should be there, but we have used a similar formulation for the negative binomial in Example 2.1.8 (3).)

1 Give an interpretation of the probability that this distribution associates with a multiset $\varphi \in \mathcal{N}(X \setminus s)$.
2 Show that $nmn[s, L](\omega)$ is indeed a distribution: its probabilities add up to one.
3 Consider the function $size \colon \mathcal{N}(X) \to \mathbb{N}$ defined by $size(\varphi) = \|\varphi\|$. Show that one obtains the negative binomial distribution as:

$$\mathcal{D}(size)\big(nmn[s, L](\omega)\big) = nbn[L](\omega(s)).$$

4 Recall that the geometric distribution from Exercise 2.1.15 corresponds to the negative binomial with one success occurrence. Simplify the above distribution $nmn[s, L](\omega)$ for $L = 1$. It may be called the multivariate geometric distribution.

2.7 Convolution

A convolution is a particular kind of (binary) operation in mathematics. In general, a convolution of parallel maps $f, g \colon X \to Y$ is a composite of the form:

$$X \xrightarrow{split} X \times X \xrightarrow{f \otimes g} Y \times Y \xrightarrow{join} Y \tag{2.59}$$

The *split* and *join* operations depend on the situation. In the sequel they are typically copy and sum.

This section describes convolution for probability distributions and channels. It does not work for all distributions, but only for those whose underlying space is a commutative monoid, like natural numbers $\mathbb{N}$ with addition (or multiplication). In the description given below, this monoid structure is 'lifted' to distributions. The construction makes essential use of parallel products (tensors $\otimes$) of distributions, like in the above diagram (2.59).

Definition 2.7.1 Let $M = (M, +, 0)$ be a commutative monoid. One defines a sum and zero element on $\mathcal{D}(M)$ via:

$$\begin{aligned} \omega + \rho &\coloneqq \mathcal{D}(+)(\omega \otimes \rho) \\ 0 &\coloneqq 1|0\rangle \end{aligned} \qquad \text{using} \qquad \begin{cases} +\colon M \times M \to M \\ 0 \in M. \end{cases} \tag{2.60}$$

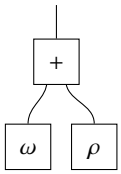
This structure $+, 0$ on $\mathcal{D}(M)$ is called *convolution*. As we shall see, it turns the set $\mathcal{D}(M)$ of distributions on M into a commutative monoid.

As a result, for $K \in \mathbb{N}$ and a distribution $\omega \in \mathcal{D}(M)$ we can write $K \cdot \omega$ for the K-fold convolution $K \cdot \omega = \omega + \cdots + \omega \in \mathcal{D}(M)$.

Alternative, equivalent descriptions of this sum of distributions are:

$$\begin{aligned} \omega + \rho &= \sum_{a,b \in M} \omega(a) \cdot \rho(b) \,|a+b\rangle \\ \Big(\sum_i r_i \,|a_i\rangle\Big) + \Big(\sum_j s_j \,|b_j\rangle\Big) &= \sum_{i,j} r_i \cdot s_j \,|a_i + b_j\rangle. \end{aligned}$$

We may describe this convolution $\omega + \rho \in \mathcal{D}(M)$ also as a string diagram, on the left below, or as a probabilistic program (with sampling), on the right.




```
a ← ω
b ← ρ
return a+b
```

(2.61)

The sum of discrete distributions $\omega + \rho$ occurs in (Keimel and Plotkin, 2017, p.82), but does not seem to be widely used and/or familiar. The construction is often implicit in a sum of random variables. It is an instance of an abstract form of convolution in (Kock, 2012, §10), like in Diagram (2.59), as in:

$$\mathbf{1} \xrightarrow{\Delta} \mathbf{1} \times \mathbf{1} \xrightarrow{\omega \otimes \rho} M \times M \xrightarrow{+} M$$

The copy Δ and sum $+$ are deterministic channels. This convolution easily extends to parallel channels $c, d \colon X \rightsquigarrow M$ as in:

$$c + d \;:=\; \Big(X \xrightarrow{\Delta} X \times X \xrightarrow{c \otimes d} M \times M \xrightarrow{+} M\Big) \tag{2.62}$$

This gives a pointwise convolution: $(c + d)(x) = c(x) + d(x)$.

Here is an example, for distributions $\omega, \rho \in \mathcal{D}(\mathbb{N})$ given by:

$$\omega \;=\; \tfrac{1}{3}|2\rangle + \tfrac{2}{3}|4\rangle \qquad\qquad \rho \;=\; \tfrac{4}{7}|0\rangle + \tfrac{1}{7}|1\rangle + \tfrac{2}{7}|2\rangle.$$

We use $\mathbb{N}$ as additive commutative monoid. Then, according to (2.60),

$$\begin{aligned} \omega + \rho &= \mathcal{D}(+)\Big(\big(\tfrac{1}{3}|2\rangle + \tfrac{2}{3}|4\rangle\big) \otimes \big(\tfrac{4}{7}|0\rangle + \tfrac{1}{7}|1\rangle + \tfrac{2}{7}|2\rangle\big)\Big) \\ &= \mathcal{D}(+)\Big(\tfrac{4}{21}|2,0\rangle + \tfrac{1}{21}|2,1\rangle + \tfrac{2}{21}|2,2\rangle + \tfrac{8}{21}|4,0\rangle + \tfrac{2}{21}|4,1\rangle + \tfrac{4}{21}|4,2\rangle\Big) \\ &= \tfrac{4}{21}|2{+}0\rangle + \tfrac{1}{21}|2{+}1\rangle + \tfrac{2}{21}|2{+}2\rangle + \tfrac{8}{21}|4{+}0\rangle + \tfrac{2}{21}|4{+}1\rangle + \tfrac{4}{21}|4{+}2\rangle \\ &= \tfrac{4}{21}|2\rangle + \tfrac{1}{21}|3\rangle + \tfrac{2}{21}|4\rangle + \tfrac{8}{21}|4\rangle + \tfrac{2}{21}|5\rangle + \tfrac{4}{21}|6\rangle \\ &= \tfrac{4}{21}|2\rangle + \tfrac{1}{21}|3\rangle + \tfrac{10}{21}|4\rangle + \tfrac{2}{21}|5\rangle + \tfrac{4}{21}|6\rangle. \end{aligned}$$

As announced, convolution gives a commutative monoid structure. But there is more to say.

Proposition 2.7.2 *1 Via sums $+, 0$ of distributions (2.60), the set $\mathcal{D}(M)$ forms a commutative monoid.*

2 If $f \colon M \to N$ is a homomorphism of commutative monoids, then so is $\mathcal{D}(f) \colon \mathcal{D}(M) \to \mathcal{D}(N)$. Let **CMon** *be the category with commutative monoids as objects, and with monoid homomorphisms as arrows between them. The above two items tell that the distribution functor $\mathcal{D} \colon$* **Sets** $\to$

Sets *can be restricted to a functor* $\mathcal{D}\colon$ **CMon** $\to$ **CMon** *in a commuting diagram:*

$$\begin{tikzcd} \mathbf{CMon} \arrow[r, "\mathcal{D}"] \arrow[d] & \mathbf{CMon} \arrow[d] \\ \mathbf{Sets} \arrow[r, "\mathcal{D}"] & \mathbf{Sets} \end{tikzcd} \tag{2.63}$$

The vertical arrows 'forget' the monoid structure, by sending a monoid to its underlying set.

Proof. 1 Commutativity and associativity of $+$ on $\mathcal{D}(M)$ follow from commutativity and associativity of $+$ on M, and of multiplication $\cdot$ on $[0,1]$. Next,

$$\omega + 0 \;=\; \omega + 1|0\rangle \;=\; \sum_{a\in M} \omega(a)\cdot 1\,\big|a+0\big\rangle \;=\; \sum_{a\in M} \omega(a)\,\big|a\big\rangle \;=\; \omega.$$

2 The unit $1|0\rangle$ of the monoid on $\mathcal{D}(M)$ is preserved since:

$$\mathcal{D}(f)(1|0\rangle) \;=\; 1|f(0)\rangle \;=\; 1|0\rangle.$$

By assumption $f\colon M \to N$ is a homomorphism of monoids. We then have an equation $f \circ + = + \circ (f \times f)$, see Diagram (1.9). Hence:

$$\begin{aligned} \mathcal{D}(f)(\omega+\rho) &\overset{(2.60)}{=} \big(\mathcal{D}(f)\circ\mathcal{D}(+)\big)(\omega\otimes\rho) \\ &= \mathcal{D}(f\circ +)(\omega\otimes\rho) \\ &= \mathcal{D}(+\circ(f\times f))(\omega\otimes\rho) \qquad f \text{ is a homomorphism} \\ &= \big(\mathcal{D}(+)\circ\mathcal{D}(f\times f)\big)(\omega\otimes\rho) \\ &= \mathcal{D}(+)\big(\mathcal{D}(f)(\omega)\otimes\mathcal{D}(f)(\rho)\big) \quad \text{by Lemma 2.3.2 (3)} \\ &\overset{(2.60)}{=} \mathcal{D}(f)(\omega) + \mathcal{D}(f)(\rho). \end{aligned}$$

❑

The lifting of distributions to commutative monoids can *not* be extended to vector spaces: scalar multiplication can be defined for distributions on vector spaces, but it does not satisfy all of the properties required to get a vector space of distributions.

Proposition 2.7.3 *Let V be a vector space (over $\mathbb{R}$). Using its commutive monoid structure $(V,+,0)$ we can form the commutative monoid $\mathcal{D}(V)$. The scalar multiplication $s\cdot(-)\colon V\to V$, for $s\in\mathbb{R}$, can be extended to $\mathcal{D}(V)$. We write it as $s * (-)\colon \mathcal{D}(V)\to\mathcal{D}(V)$ and define it on a distribution $\omega\in\mathcal{D}(V)$ as a generalisation of the scaling of Example 2.1.4 (1):*

$$s * \omega \;:=\; \mathrm{scale}(s,\omega) \overset{(2.7)}{=} \mathcal{D}(s\cdot(-))(\omega) \;\in\; \mathcal{D}(V). \tag{2.64}$$

1 *This scalar multiplication $*$ is a monoid action $\mathbb{R}\times\mathcal{D}(V)\to\mathcal{D}(V)$, for the multiplicative monoid structure $(\mathbb{R},\cdot,1)$. In addition, for a linear map $f\colon V\to W$ between vector spaces, the pushforward function $\mathcal{D}(f)\colon\mathcal{D}(V)\to\mathcal{D}(W)$ preserves scalar multiplication $*$.*
2 *It is a map of monoids $s * (-)\colon\mathcal{D}(V)\to\mathcal{D}(V)$, for each $s\in\mathbb{R}$.*
3 *However, it is* not *a map of monoids $(-) * \omega\colon\mathbb{R}\to\mathcal{D}(V)$, for the additive monoid structure $(\mathbb{R},+,0)$, and for $\omega\in\mathcal{D}(V)$.*

Because of point (3), the commutative monoid $(\mathcal{D}(V),+,0)$ is not *a vector space with the scalar multiplication $*$ from (2.64).*

Although scalar multiplication $*$ fails to satisfy an obvious property for vector spaces, it is still a useful operations that will (re)appear in many places.

Proof. 1 This is a consequence of Proposition 2.1.5.

2 Next, for $s \in \mathbb{R}$, the function $s * (-)\colon \mathcal{D}(V) \to \mathcal{D}(V)$ is a map of monoids, as expressed by the equations:

$$s * 1|0\rangle = 1|0\rangle \qquad\qquad s * (\omega + \rho) = (s * \omega) + (s * \rho).$$

We use that scalar multiplication $s\colon (-)\colon V \to V$ is by definition a map of monoids. Hence $s * (-) = \mathcal{D}(s \cdot (-))$ is also a map of monoids $\mathcal{D}(V) \to \mathcal{D}(V)$, by Proposition 2.7.2 (2).

3 Here is a simple counterexample. Take $\omega = \frac{1}{3}|2\rangle + \frac{2}{3}|5\rangle \in \mathcal{D}(\mathbb{R})$. Then:

$$2 * \omega = \tfrac{1}{3}|4\rangle + \tfrac{2}{3}|10\rangle \qquad\qquad \omega + \omega = \tfrac{1}{9}|4\rangle + \tfrac{4}{9}|7\rangle + \tfrac{4}{9}|10\rangle.$$

Hence: $2 * \omega = (1 + 1) * \omega \neq (1 * \omega) + (1 * \omega) = \omega + \omega$. ❑

In Exercise 2.7.17 below we shall see that the restricted (or lifted) functor $\mathcal{D}\colon \mathbf{CMon} \to \mathbf{CMon}$ is also a monad. The same construction works for $\mathcal{D}_\infty$.

Remark 2.7.4 Let $\omega \in \mathcal{D}(\mathbb{R})$ be a distribution on the reals, considered as vector space. Recall from Fact (2.3.10) that in general $\mathcal{D}(\Delta)(\omega) \neq \omega \otimes \omega$, using the diagonal $\Delta\colon \mathbb{R} \to \mathbb{R} \times \mathbb{R}$. This is related to the inequality $2 * \omega \neq \omega + \omega$ that we have seen above, in the proof of Proposition 2.7.3 (3). Using the sum function $+\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ we have, in general:

$$\begin{aligned} 2 * \omega = \sum_{x \in \mathbb{R}} \omega(x)\big|2 \cdot x\rangle = \sum_{x \in \mathbb{R}} \omega(x)\big|x + x\rangle &= \mathcal{D}(+)\big(\mathcal{D}(\Delta)(\omega)\big) \\ &\neq \mathcal{D}(+)\big(\omega \otimes \omega\big) = \omega + \omega. \end{aligned}$$

The next example illustrates how we sometimes have to stretch the underlying space a bit to make it into a monoid so that the monoid structure can be used for convolution.

Example 2.7.5 Recall that we write *dice* for the uniform (fair) dice distribution $dice = unif_{pips} = \sum_{1 \leq i \leq 6} \frac{1}{6}|i\rangle \in \mathcal{D}(pips)$, where $pips = \{1, 2, 3, 4, 5, 6\}$. When we throw two dices and we wish to look at the sum of outcomes, we have go beyond *pips*. It turns out to be convenient to use the inclusion $pips \subseteq \mathbb{N}$ and to consider a dice as a distribution $dice \in \mathcal{D}(\mathbb{N})$ on the natural numbers. This change is a formality, since the support of dice is still finite: $supp(dice) = pips$.

A big benefit of this change of sample space, from *pips* to $\mathbb{N}$, is that we can exploit that the set $\mathbb{N}$ of natural numbers carries a commutative monoid structure, in this case via addition. Hence we can form the convolution (sum) of two dices as:

$$\begin{aligned} 2 \cdot dice = dice + dice &\overset{(2.60)}{=} \mathcal{D}(+)(dice \otimes dice) \\ &= \mathcal{D}(+)\left(\sum_{i,j \in pips} \tfrac{1}{36}\big|i, j\rangle\right) \\ &= \sum_{i,j \in pips} \tfrac{1}{36}\big|i + j\rangle \\ &= \tfrac{1}{36}\big|2\rangle + \tfrac{1}{18}\big|3\rangle + \tfrac{1}{12}\big|4\rangle + \tfrac{1}{9}\big|5\rangle + \tfrac{5}{36}\big|6\rangle + \tfrac{1}{6}\big|7\rangle \\ &\quad + \tfrac{5}{36}\big|8\rangle + \tfrac{1}{9}\big|9\rangle + \tfrac{1}{12}\big|10\rangle + \tfrac{1}{18}\big|11\rangle + \tfrac{1}{36}\big|12\rangle. \end{aligned}$$

This gives a new distribution in $\mathcal{D}(\mathbb{N})$. The construction can be generalised easily to an n-throw $n \cdot dice = dice + \cdots + dice \in \mathcal{D}(\mathbb{N})$.

In Example 2.1.4 (3) we have described the maximum of two dices as an image distribution.

It can also be described as a convolution, by using the maximum as a commutative monoid operation on the set *pips*. In that case we do not need to go beyond *pips* to $\mathbb{N}$.

Exercise 2.7.7 contains another variation, with modular arithmetic.

The general structure of Diagram (2.59) becomes clear when we apply convolution in the context of channels. We do a simple example first, in which we consider flips and binomials as distributions on $\mathbb{N}$.

Proposition 2.7.6 *Fix a probability $r \in [0,1]$.*

1 *The convolution sum of multiple flip distributions is a binomial distribution:*

$$\mathit{flip}(r) + \mathit{flip}(r) \;=\; \mathit{bn}[2](r).$$

In a diagram:

$$\begin{tikzcd} {[0,1]} \arrow[r, "\Delta"] \arrow[drr, bend right, "\mathit{bn}[2]"'] & {[0,1]\times[0,1]} \arrow[r, "\mathit{flip}\otimes\mathit{flip}"] & \mathbb{N}\times\mathbb{N} \arrow[d, "+"] \\ & & \mathbb{N} \end{tikzcd}$$

More generally, in K-ary form:

$$K \cdot \mathit{flip}(r) \;=\; \underbrace{\mathit{flip}(r) + \cdots + \mathit{flip}(r)}_{K \text{ times}} \;=\; \mathit{bn}[K](r).$$

This equation already occurred in Corollary 2.5.3 but not in terms of an iterated convolution sum.

2 *Binomials themselves are closed under convolution:*

$$\begin{tikzcd} {[0,1]} \arrow[r, "\Delta"] \arrow[drr, bend right, "\mathit{bn}[K+L]"'] & {[0,1]\times[0,1]} \arrow[r, "\mathit{bn}[K]\otimes\mathit{bn}[L]"] & \mathbb{N}\times\mathbb{N} \arrow[d, "+"] \\ & & \mathbb{N} \end{tikzcd}$$

Using the convolution of channels notation (2.62) we can simply write $\mathrm{bn}[K]+\mathrm{bn}[L] = \mathrm{bn}[K+L]$.

This last equation can also be obtained by marginalising out the wire on the right in Theorem 2.5.2 (2). We shall see later in Exercise 3.3.10 that multinomial distributions are also closed under convolution.

Proof. 1 We do the binary case:

$$\begin{aligned} \mathit{flip}(r) + \mathit{flip}(r) &= \mathcal{D}(+)\big(\mathit{flip}(r) \otimes \mathit{flip}(r)\big) \\ &= \mathcal{D}(+)\big((r|1\rangle + (1-r)|0\rangle) \otimes (r|1\rangle + (1-r)|0\rangle)\big) \\ &= \mathcal{D}(+)\big(r^2\big|1,1\rangle + r\cdot(1-r)\big|1,0\rangle + (1-r)\cdot r\big|0,1\rangle + (1-r)^2\big|0,0\rangle\big) \\ &= r^2\big|2\rangle + 2\cdot r\cdot(1-r)\big|1\rangle + (1-r)^2\big|0\rangle \\ &= \sum_{0\leq i\leq 2} \binom{2}{i} \cdot r^i \cdot (1-r)^{2-i}\,\big|i\rangle \;=\; \mathit{bn}[2](r). \end{aligned}$$

2 For $r \in [0,1]$ and $K, L \in \mathbb{N}$ we compute, via Vandermonde's binary formula (1.48):

$$\begin{aligned}
&\mathcal{D}(+)\Big(bn[K](r) \otimes bn[L](r)\Big) \\
&= \mathcal{D}(+)\left(\sum_{0\leq i\leq K}\sum_{0\leq j\leq L}\binom{K}{i}\cdot r^i\cdot(1-r)^{K-i}\cdot\binom{L}{j}\cdot r^j\cdot(1-r)^{L-j}\,\big|i,j\rangle\right) \\
&= \sum_{0\leq i\leq K}\sum_{0\leq j\leq L}\binom{K}{i}\cdot\binom{L}{j}\cdot r^{i+j}\cdot(1-r)^{(K+L)-(i+j)}\,\big|i+j\rangle \\
&= \sum_{0\leq m\leq K+L}\left(\sum_{0\leq i\leq K,\,0\leq j\leq L,\,i+j=m}\binom{K}{i}\cdot\binom{L}{j}\right)\cdot r^m\cdot(1-r)^{(K+L)-m}\,\big|m\rangle \\
&\overset{(1.48)}{=} \sum_{0\leq m\leq K+L}\binom{K+L}{m}\cdot r^m\cdot(1-r)^{(K+L)-m}\,\big|m\rangle \;=\; bn[K+L](r).
\end{aligned}$$

❑

The next illustration involves distributions on the commutative monoids of multisets $\mathcal{N}(X)$ and of natural numbers, with a homomorphism of monoids $sum\colon \mathcal{N}(X) \to \mathbb{N}$ between them, as in Proposition 2.7.2 (2).

Proposition 2.7.7 *1 Fix a set X and consider the multiset coefficient distribution $mc[X,K]$ from (2.5) and (2.57), as a distribution on the commutative monoid $\mathcal{N}(X)$ of natural multisets. These distributions are closed under convolution:*

$$mc[X,K_1] \;+\; mc[X,K_2] \;=\; mc[X,K_1+K_2].$$

2 Fix a number $N \geq 1$. The nomial distributions from (2.3), considered as distributions on the commutative (additive) monoid $\mathbb{N}$, are also closed under convolution, in the following way.

$$nom[N,K_1] \;+\; nom[N,K_2] \;=\; nom[N,K_1+K_2].$$

Proof. 1 Via Lemma 1.8.3 (2):

$$\begin{aligned}
mc[X,K_1] + mc[X,K_2] &= \sum_{\varphi_1\in\mathcal{N}[K_1](X)}\;\sum_{\varphi_2\in\mathcal{N}[K_2](X)}\frac{(\varphi_1)}{N^{K_1}}\cdot\frac{(\varphi_2)}{N^{K_2}}\,\big|\varphi_1+\varphi_2\rangle \\
&= \sum_{\psi\in\mathcal{N}[K_1+K_2](X)}\;\sum_{\varphi\leq_{K_1}\psi}\frac{(\varphi)\cdot(\psi-\varphi)}{N^{(K_1+K_2)}}\,\big|\psi\rangle \\
&= \sum_{\psi\in\mathcal{N}[K_1+K_2](X)}\frac{(\psi)}{N^{(K_1+K_2)}}\,\big|\psi\rangle \;=\; mc[X,K_1+K_2].
\end{aligned}$$

2 We present two proofs. First, direct calculation works via the Vandermonde property from Lemma 1.10.2 (2):

$$\begin{aligned}
&nom[N,K_1] + nom[N,K_2] \\
&= \sum_{0\leq i_1\leq(N-1)\cdot K_1}\;\sum_{0\leq i_2\leq(N-1)\cdot K_2} nom[N,K_1](i_1)\cdot nom[N,K_2](i_2)\,\big|i_1+i_2\rangle \\
&= \sum_{0\leq i\leq(N-1)\cdot(K_1+K_2)}\;\sum_{0\leq i_1\leq(N-1)\cdot K_1,\,0\leq i_2\leq(N-1)\cdot K_2,\,i_1+i_2=i}\frac{C_N(K_1,i_1)}{N^{K_1}}\cdot\frac{C_N(K_1,i_2)}{N^{K_2}}\,\big|i\rangle \\
&= \sum_{0\leq i\leq(N-1)\cdot(K_1+K_2)}\frac{C_N(K_1+K_2,i)}{N^{K_1+K_2}}\,\big|i\rangle \;=\; nom[N,K_1+K_2].
\end{aligned}$$

Alternatively, we can use the first item, the description of the nomial distribution as image under *sum* from Exercise 2.1.4, and sum-preservation from Proposition 2.7.2 (2):

$$\begin{aligned} nom[N, K_1] + nom[N, K_2] &= \mathcal{D}(sum)\big(mc[N, K_1]\big) + \mathcal{D}(sum)\big(mc[N, K_2]\big) \\ &= \mathcal{D}(sum)\big(mc[N, K_1] + mc[N, K_2]\big) \\ &= \mathcal{D}(sum)\big(mc[N, K_1 + K_2]\big) \\ &= nom[N, K_1 + K_2]. \end{aligned}$$

❑

The following result for independent and identical distributions has a definite convolution-flavour, but does not fully fit in the convolution mold. We can see inclusions $X^K \subseteq \mathcal{L}(X)$ at work, but the monoid of lists $\mathcal{L}(X)$, with concatenation $+\!\!+$, is not commutative. Still, the following result makes sense.

Lemma 2.7.8 *For a distribution $\omega \in \mathcal{D}(X)$ and numbers K, L,*

$$\mathcal{D}(+\!\!+)\big(iid[K](\omega) \otimes iid[L](\omega)\big) = iid[K+L].$$

This means that the following convolution-style diagram commutes.

$$\begin{tikzcd} \mathcal{D}(X) \arrow[r, "\Delta"] \arrow[drr, bend right, "iid[K+L]"'] & \mathcal{D}(X) \times \mathcal{D}(X) \arrow[r, "iid[K]\otimes iid[L]"] & X^K \times X^L \arrow[d, "\cong" ', "+\!\!+"] \\ & & X^{K+L} \end{tikzcd}$$

Proof. Since:

$$\begin{aligned} (+\!\!+ \circ (iid[K] \otimes iid[L]) \circ \Delta)(\omega) &= \sum_{\vec{x} \in X^K, \vec{y} \in X^L} \omega^K(\vec{x}) \cdot \omega^L(\vec{y}) \,\big|\, \vec{x} +\!\!+ \vec{y} \rangle \\ &= \sum_{\vec{x} \in X^K, \vec{y} \in X^L} \omega^{K+L}(\vec{x} +\!\!+ \vec{y}) \,\big|\, \vec{x} +\!\!+ \vec{y} \rangle \\ &= \sum_{\vec{z} \in X^{K+L}} \omega^{K+L}(\vec{z}) \,\big|\, \vec{z} \rangle = \omega^{K+L} = iid[K+L](\omega). \end{aligned}$$

❑

We continue with a basic property of Poisson distributions $pois[\lambda]$, see (2.10). This property is commonly expressed in terms of random variables X_i as: if $X_1 \sim pois[\lambda_1]$ and $X_2 \sim pois[\lambda_2]$ then $X_1 + X_2 \sim pois[\lambda_1 + \lambda_2]$. We have not discussed random variables yet, but we do not need them for the channel-based reformulation that we use below. It involves a parallel product $pois \otimes pois$ of channels, as introduced in Definition 2.4.4 (2).

Recall that the Poisson distribution has infinite support, so that we need to use $\mathcal{D}_\infty$ instead of $\mathcal{D}$, see Definition 2.1.6, but that difference is immaterial here. We now use the mapping $\lambda \mapsto pois[\lambda]$ as a function $\mathbb{R}_{\geq 0} \to \mathcal{D}_\infty(\mathbb{N})$ and as a $\mathcal{D}_\infty$-channel $pois \colon \mathbb{R}_{\geq 0} \rightsquigarrow \mathbb{N}$.

Proposition 2.7.9 *The Poisson channel $pois \colon \mathbb{R}_{\geq 0} \to \mathcal{D}_\infty(\mathbb{N})$ is a homomorphism of monoids. Indeed, for $\lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}$,*

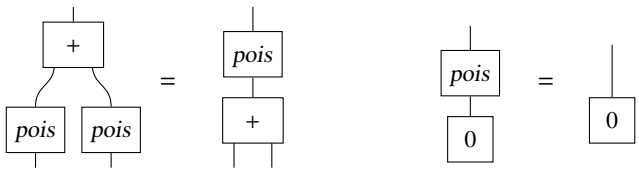
$$pois[\lambda_1 + \lambda_2] = pois[\lambda_1] + pois[\lambda_2] \qquad \textit{and} \qquad pois[0] = 1|0\rangle.$$

This monoid structure on $\mathcal{D}_\infty(\mathbb{N})$ is the one from Definition 2.7.1, building on the additive monoid structure of $\mathbb{N}$.

One can express the above two equations via commutation of the following two diagrams of channels.

$$\begin{tikzcd} \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \arrow[r, "+"] \arrow[d, "pois \otimes pois"'] & \mathbb{R}_{\geq 0} \arrow[d, "pois"] \\ \mathbb{N} \times \mathbb{N} \arrow[r, "+"] & \mathbb{N} \end{tikzcd} \qquad \begin{tikzcd} 1 \arrow[r, "0"] \arrow[d, equal, "unit"'] & \mathbb{R}_{\geq 0} \arrow[d, "pois"] \\ 1 \arrow[r, "0"] & \mathbb{N} \end{tikzcd}$$

Alternatively, these two equations can be expressed via equations of string diagrams:



Proof. We reason equationally and first do preservation of sums +, for which we pick arbitrary $\lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}$ and $k \in \mathbb{N}$.

$$\begin{aligned} (pois[\lambda_1] + pois[\lambda_2])(k) &\overset{(2.60)}{=} \mathcal{D}(+)(pois[\lambda_1] \otimes pois[\lambda_2])(k) \\ &= \sum_{k_1, k_2, k_1+k_2=k} (pois[\lambda_1] \otimes pois[\lambda_2])(k_1, k_2) \\ &= \sum_{0 \leq m \leq k} pois[\lambda_1](m) \cdot pois[\lambda_2](k-m) \\ &\overset{(2.10)}{=} \sum_{0 \leq m \leq k} \left(e^{-\lambda_1} \cdot \frac{\lambda_1^m}{m!}\right) \cdot \left(e^{-\lambda_2} \cdot \frac{\lambda_2^{k-m}}{(k-m)!}\right) \\ &= \sum_{0 \leq m \leq k} \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot \frac{k!}{m! \cdot (k-m)!} \cdot \lambda_1^m \cdot \lambda_2^{k-m} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot \sum_{0 \leq m \leq k} \binom{k}{m} \cdot \lambda_1^m \cdot \lambda_2^{k-m} \\ &\overset{(1.42)}{=} \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \cdot (\lambda_1+\lambda_2)^k \overset{(2.10)}{=} pois[\lambda_1+\lambda_2](k). \end{aligned}$$

Finally, in the expression $pois[0] = \sum_k e^0 \cdot \frac{0^k}{k!}|k\rangle$ everything vanishes except for $k = 0$, since only $0^0 = 1$. Hence $pois[0] = 1|0\rangle$. ❑

We close this section by showing that K-fold convolutions can also be expressed via multionial distributions.

Proposition 2.7.10 *Let $\omega \in \mathcal{D}(M)$ be a distribution over a commutative monoid M. For a multiset $\varphi \in \mathcal{N}[K](M)$ we write $sum(\varphi) = \sum_{x \in M} \varphi(x) \cdot x \in M$. Then, for $K \geq 1$,*

$$K \cdot \omega = \omega + \cdots + \omega = \mathcal{D}(sum)\big(mn[K](\omega)\big).$$

Proof. We also write $sum\colon M^K \to M$ for $sum(x_1, \ldots, x_K) = x_1 + \cdots + x_K$. Clearly, $sum(\vec{x}) =$

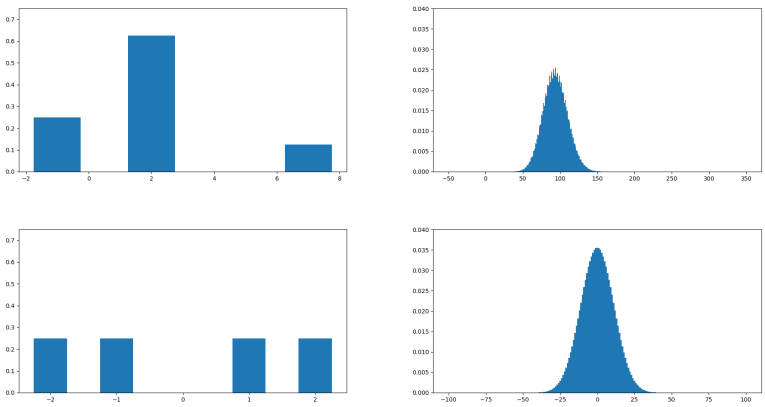


Figure 2.7 Illustrations of two distributions $\omega = \frac{1}{4}|-1\rangle + \frac{5}{8}|2\rangle + \frac{1}{8}|7\rangle$ and $\rho = \frac{1}{4}|-2\rangle + \frac{1}{4}|-1\rangle + \frac{1}{4}|1\rangle + \frac{1}{4}|2\rangle$ on the left and their 50-fold iterated convolutions $50 \cdot \omega$ and $50 \cdot \rho$ on the right.

$sum(acc(\vec{x}))$, where the latter sum works on multisets. Then, by Proposition 1.7.2:

$$\begin{aligned} K \cdot \omega \;=\; \mathcal{D}(sum)(\omega^K) \;&=\; \sum_{\vec{x} \in M^K} \omega^K(\vec{x}) \,\big|\, sum(\vec{x}) \rangle \\ &=\; \sum_{\varphi \in \mathcal{N}[K](M)} \; \sum_{\vec{x} \in acc^{-1}(\varphi)} \omega^K(\vec{x}) \,\big|\, sum(\varphi) \rangle \\ &=\; \sum_{\varphi \in \mathcal{N}[K](M)} \; \sum_{\vec{x} \in acc^{-1}(\varphi)} \prod_{x \in M} \omega(x)^{\varphi(x)} \,\big|\, sum(\varphi) \rangle \\ &=\; \sum_{\varphi \in \mathcal{N}[K](M)} \mathbf{(}\varphi\mathbf{)} \cdot \prod_{x \in M} \omega(x)^{\varphi(x)} \,\big|\, sum(\varphi) \rangle \;=\; \mathcal{D}(sum)\big(mn[K](\omega)\big). \qquad \square \end{aligned}$$

We will look closer at these iterated convolutions $K \cdot \omega$, later on, in Section 5.5. When K goes to infinity, such iterations take the form of a Bell curve, as illustrated in Figure 2.7. The Bell shapes always arise, independent of the discrete distribution ω that we start with. The so-called mean (center location) and variance (width) of the Bell curve are entirely determined by the (mean and variance of the) orginal distribution ω. The emergence of Bell curves from such iterations is a deep fact, captured by the so-called Central Limit Theorem, see *e.g.* Blitzstein and Hwang (2019).

Exercises

2.7.1 Describe $3 \cdot dice \in \mathcal{D}(\mathbb{N})$ in detail, following Example 2.7.5.

2.7.2 Describe the conjunction and disjunction channels *ac* and *oc* from Exercise 2.5.9 as convolutions $\mathbf{2} \times \mathbf{2} \rightarrowtail \mathbf{2}$.

2.7.3 Show that convolution preserves convex sums in its arguments. Explicitly, for a commutative monoid M with distributions $\omega_1, \ldots, \omega_n, \rho \in \mathcal{D}(M)$ and with probabilities $r_1, \ldots, r_n \in$

$[0,1]$ satisfying $\sum_i r_i = 1$, one has:

$$\Big(r_1 \cdot \omega_1 + \cdots + r_n \cdot \omega_n\Big) + \rho \;=\; r_1 \cdot (\omega_1 + \rho) + \cdots + r_n \cdot (\omega_n + \rho).$$

2.7.4 The set of natural numbers $\mathbb{N}$ is a commutative semiring: it has two commutative monoid structures, one additive with $+, 0$, and one multiplicative with $\cdot, 1$, where multiplication distributes over sum. Accordingly, Definiton 2.7.1 gives two commutative monoid structures on the set of distributions $\mathcal{D}(\mathbb{N})$, namely:

$$\omega + \rho \;=\; \mathcal{D}(+)(\omega \otimes \rho) \qquad \text{and} \qquad \omega \star \rho \;=\; \mathcal{D}(\cdot)(\omega \otimes \rho).$$

Consider the following three distributions on $\mathbb{N}$.

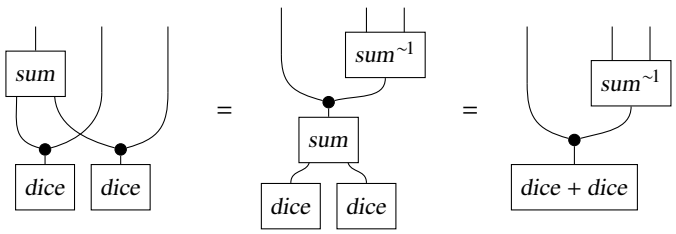
$$\rho_1 \;=\; \tfrac{1}{2}|0\rangle + \tfrac{1}{3}|1\rangle + \tfrac{1}{6}|2\rangle \quad \rho_2 \;=\; \tfrac{1}{2}|0\rangle + \tfrac{1}{2}|1\rangle \quad \omega \;=\; \tfrac{2}{3}|0\rangle + \tfrac{1}{3}|1\rangle.$$

Show consecutively:

1 $\rho_1 + \rho_2 = \frac{1}{4}|0\rangle + \frac{5}{12}|1\rangle + \frac{1}{4}|2\rangle + \frac{1}{12}|3\rangle$;
2 $\omega \star (\rho_1 + \rho_2) = \frac{3}{4}|0\rangle + \frac{5}{36}|1\rangle + \frac{1}{12}|2\rangle + \frac{1}{36}|3\rangle$;
3 $\omega \star \rho_1 = \frac{5}{6}|0\rangle + \frac{1}{9}|1\rangle + \frac{1}{18}|2\rangle$;
4 $\omega \star \rho_2 = \frac{5}{6}|0\rangle + \frac{1}{6}|1\rangle$;
5 $(\omega \star \rho_1) + (\omega \star \rho_2) = \frac{25}{36}|0\rangle + \frac{25}{108}|1\rangle + \frac{7}{108}|2\rangle + \frac{1}{108}|3\rangle$.

Observe that $\star$ does not distribute over $+$ on $\mathcal{D}(\mathbb{N})$. More generally, conclude that the functor $\mathcal{D}\colon \mathbf{Sets} \to \mathbf{Sets}$ restricts to commutative monoids but not to commutative (semi)rings. Similarly, $\mathcal{D}$ does not restrict to Abelian groups.

2.7.5 Recall $pips = \{1,2,3,4,5,6\}$ and $dice = unif_{pips}$. Consider the (surjective) addition function $sum\colon pips \times pips \to \{1,\ldots,12\}$. Determine what its probabilistic inverse channel $sum^{\sim 1}\colon \{1,\ldots,12\} \rightsquigarrow pips \times pips$ is and check the following series of equations between string diagrams, where the left one can be seen as an instance of Exercise 2.5.6.



2.7.6 Show that *negative* binomials *nbn*, see Example 2.1.8 (3), are also closed under convolution:

$$nbn[K](s) + nbn[L](s) \;=\; nbn[K+L](s),$$

where the sum $+$ of distributions on the left-hand-side is the one from Definition 2.7.1 for $\mathcal{D}_\infty$.

2.7.7 Recall that for $N \in \mathbb{N}_{>0}$ we write $\boldsymbol{N} = \{0,1,\ldots,N-1\}$ for the set of natural numbers (strictly) below N. It is an additive monoid, via addition modulo N. As such it is sometimes written as $\mathbb{Z}_N$ or as $\mathbb{Z}/N\mathbb{Z}$. Prove that:

$$unif_N + unif_N \;=\; unif_N, \qquad \text{with } + \text{ from Definition 2.7.1.}$$

You may wish to check this equation first for $N = 4$ or $N = 5$. It works for the modular sum,

not for the ordinary sum (on $\mathbb{N}$), as one can see from the sum *dice*+*dice*, see Example 2.7.5. See Scozzafava (1993) for more info.

2.7.8 Let M be a commutative monoid. There is a shift function:

$$M \times \mathcal{D}(M) \xrightarrow{shift} \mathcal{D}(M) \qquad \text{given by} \qquad shift(m, \omega) \coloneqq 1|m\rangle + \omega.$$

1 Check that $shift(m, \omega) = \mathcal{D}(m + (-))(\omega)$ and thus that the above shift operation generalises Example 2.1.4 (1) from $\mathbb{R}$ to a commutative monoid. Conclude from Proposition 2.1.5 that *shift* is a monoid action.

2 Show also that:

$$shift(m_1, \omega_1) + \cdots + shift(m_K, \omega_K) = shift(m_1 + \cdots + m_K, \omega_1 + \cdots + \omega_K).$$

3 Conclude that:

$$K \cdot shift(m, \omega) = shift(K \cdot m, K \cdot \omega).$$

4 Let $\omega, \rho \in \mathcal{D}(M)$ be given. Consider the channel $shift(-, \rho)\colon M \to \mathcal{D}(M)$ and show that:

$$shift(-, \rho)_*(\omega) = \omega + \rho.$$

5 Finally, assume now that M is a vector space and that $s \in \mathbb{R}$ is a scalar. Show that for the scalar multiplication $*$ from (2.64) one has:

$$s * shift(m, \omega) = shift(s \cdot m, s * \omega).$$

2.7.9 Let $\omega \in \mathcal{D}(V)$, where V is a vector space, with numbers $K \in \mathbb{N}_{>0}$ and $s \in \mathbb{R}$.

1 Prove that K-fold convolution and scalar multiplication (2.64) commute:

$$s * (K \cdot \omega) = K \cdot (s * \omega).$$

2 Deduce that:

$$\tfrac{1}{K} * \Big(K \cdot \omega\Big) = K \cdot \Big(\tfrac{1}{K} * \omega\Big).$$

Check that this distribution differs from ω.

2.7.10 Let M, L be two commutative monoids and let $c\colon M \to \mathcal{D}(L)$ be a channel that forms a map of monoids — like in for instance in Proposition 2.7.9. Show that pushforward $c_*\colon \mathcal{D}(M) \to \mathcal{D}(L)$ is then also a map of monoids.

2.7.11 Let M_1, M_2 be two commutative monoids. The product $M_1 \times M_2$ is then also a commutative monoid, via component-wise operations. Show that the tensor $\otimes$ commutes with convolution $+$ in the follow way. For distributions $\omega_i, \rho_i \in \mathcal{D}(M_i)$ one has:

$$(\omega_1 + \rho_1) \otimes (\omega_2 + \rho_2) = (\omega_1 \otimes \omega_2) + (\rho_1 \otimes \rho_2).$$

This can be expressed via the following diagram, where the isomorphism is $\langle \pi_1, \pi_3, \pi_2, \pi_4 \rangle$.

$$\begin{tikzcd}
\mathcal{D}(M_1)\times\mathcal{D}(M_1)\times\mathcal{D}(M_2)\times\mathcal{D}(M_2) \arrow[r, "+\times+"] \arrow[d, equal, "\wr"'] & \mathcal{D}(M_1)\times\mathcal{D}(M_2) \arrow[d, "\otimes"] \\
\mathcal{D}(M_1)\times\mathcal{D}(M_2)\times\mathcal{D}(M_1)\times\mathcal{D}(M_2) \arrow[r, "\otimes\times\otimes"'] & \mathcal{D}(M_1\times M_2)\times\mathcal{D}(M_1\times M_2) \arrow[u, "+"']
\end{tikzcd}$$

2.7.12 Consider a function $f\colon X \to M$ where X is an ordinary set and M is a commutative monoid. We can *add noise* to the function f via a channel $c\colon X \rightsquigarrow M$. The result is a channel $noise(f, c)\colon X \rightsquigarrow M$ given by pointwise convolution:

$$noise(f, c) \coloneqq \langle f \rangle + c.$$

Check that we can concretely describe this noise channel as:

$$noise(f, c)(x) = \sum_{y \in Y} c(x)(y) \,\big|\, f(x) + y \rangle.$$

2.7.13 Convolution as described in Definition 2.7.1 for distributions $\mathcal{D}$ can also be formulated for multisets $\mathcal{N}$ and $\mathcal{M}$, via their tensors. For this purpose, let $M = (M, +, 0)$ be a commutative monoid.

1 Write the convolution operation multiplicatively as $\star$ in:

$$\star \coloneqq \Big(\mathcal{N}(M) \times \mathcal{N}(M) \xrightarrow{\otimes} \mathcal{N}(M \times M) \xrightarrow{\mathcal{N}(+)} \mathcal{N}(M)\Big)$$

and describe concretely what $\varphi \star \psi$ is for multisets $\varphi, \psi \in \mathcal{N}(M)$.

2 Check that $\star$ preserves $+, 0$ in each coordinate and thus turns $\mathcal{N}(M)$ into a commutative semiring with unit.

3 Take $M = \mathbb{N}$ and use the identification of elements of $\mathcal{N}(\mathbb{N})$ with polynomials from Exercise 1.6.6. Compute both:

$$\Big(2|1\rangle + 3|2\rangle\Big) \star \Big(1|0\rangle + 1|1\rangle + 1|2\rangle\Big) \qquad\qquad \Big(2x + 3x^2\Big) \cdot \Big(1 + x + x^2\Big).$$

Show in general that $\star$ corresponds to multiplication of polynomials.

2.7.14 Check that Proposition 2.7.6 (2) can be reformulated as: for each $r \in [0, 1]$ the channel $bn[-](r)\colon \mathbb{N} \to \mathcal{D}(\mathbb{N})$ is a map of (commutative) monoids, making diagrams commute as in Proposition 2.7.9.

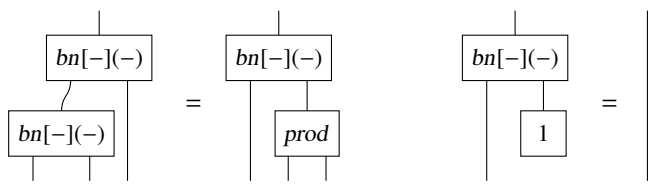
2.7.15 Recall Exercise 2.4.2, with its disease and infection rates $dr, tr \in [0, 1]$.

1 Suppose on weekdays you meet 20 people per day and in the weekend only 10. Check that the distribution of the number of people with the disease that you encounter per week is given by the convolution sum of binomials:

$$5 \cdot bn[20](dr) + 2 \cdot bn[10](dr) = bn[120](dr).$$

2 Conclude that for you the infection distribution after a week is $ic[tr]_*(bn[120](dr))$, where the channel $ic[tr]$ is as defined in Exercise 2.4.2.

2.7.16 We consider the binomial distribution from Example 2.1.2 (2) as a channel $bn[-](-)\colon \mathbb{N} \times [0, 1] \rightsquigarrow \mathbb{N}$ with two arguments / inputs. We further write $prod\colon [0, 1] \times [0, 1] \to [0, 1]$ for the obvious multiplication function $prod(r, s) = r \cdot s$, and consider it as a deterministic channel. Together with the number 1 it forms a (commutative) monoid. Consider the equations of string diagrams:



1 Check that these equations corresponds to the equations:

$$bn[-](s)_*(bn[K](r)) \;=\; bn[K](r \cdot s) \qquad\qquad bn[K](1) \;=\; 1|K\rangle.$$

2 Prove these equations.

3 Check that they express that $bn[-](-)\colon \mathbb{N} \times [0,1] \rightsquigarrow \mathbb{N}$ is an action of the multiplicative monoid $[0,1]$ on $\mathbb{N}$, in the category of channels.

(Commonly this situation is not described in terms of a monoid action, but as a 'conditional binomial'.)

2.7.17 Consider the situation described in Proposition 2.7.2, with a commutative monoid M, and induced monoid structure on $\mathcal{D}(M)$.

1 Check that $1|a\rangle + 1|B\rangle = 1|a+b\rangle$, for all $a, b \in M$. This says that $unit\colon M \to \mathcal{D}(M)$ is a homomorphism of monoids, which can also be expressed via commutation of the diagram:

$$\begin{tikzcd} M \times M \arrow[r, "unit \times unit"] \arrow[d, "+"'] & \mathcal{D}(M) \times \mathcal{D}(M) \arrow[d, "+"] \\ M \arrow[r, "unit"] & \mathcal{D}(M) \end{tikzcd}$$

2 Check also that $flat\colon \mathcal{D}(\mathcal{D}(M)) \to \mathcal{D}(M)$ is a monoid homomorphism. This means that for $\Omega, \Theta \in \mathcal{D}(\mathcal{D}(M))$ one has:

$$flat(\Omega + \Theta) \;=\; flat(\Omega) + flat(\Theta) \qquad \text{and} \qquad flat\Big(1\Big|1|0\rangle\rangle\Big) \;=\; 1|0\rangle.$$

The sum $+$ on the on the right-hand-side of the (first) equation is the one in $\mathcal{D}(M)$, from the beginning of this exercise. The sum $+$ on the left-hand-side is the one in $\mathcal{D}(\mathcal{D}(M))$, using that $\mathcal{D}(M)$ is a commutative monoid — and thus $\mathcal{D}(\mathcal{D}(M))$ too.

3 Check that the functor $\mathcal{D}\colon \mathbf{CMon} \to \mathbf{CMon}$ in (2.63) is also a monad.

4 Show that if V is a vector space, then the unit and flatten functions $V \to \mathcal{D}(V)$ and $\mathcal{D}^2(V) \to \mathcal{D}(V)$ preserve scalar multiplication.

2.8 A glimpse of probabilistic logic programming

Logic programming is one of the paradigms for writing software, based on logic. Computation proceeds by applying logical rules. Theses rules can be defined by the programmer in general form, containing variables that can be suitably instantiated to match other rules. Logic programming gained popularity especially in the area of artificial intelligence. There are probabilistic versions of logic programming, like PRISM Sato and Kameya (2008), ICL Poole (2008), or ProbLog Raedt et al. (2007); Raedt and Kimmig (2015). In this section we take a closer look at ProbLog, to illustrate how distributions and channels can be represented in probabilistic logic programming, and also to see how the essentials of this language can be represented as distributions and channels. The ProbLog languages is implemented as Python package. This review is in no way complete. It ignores many aspects of the language, notably the underlying inference algorithms. There is also no claim to generality: this section contains several examples, which serve to give an impression. Later on we shall use ProbLog again for illustrating validity of predicates and observables and for probabilistic updating. In the next section we take a similar brief look at the probabilistic programming language WebPPL.

Typical examples of distributions that we use are of the form $\omega = \frac{1}{2}|a\rangle + \frac{1}{3}|b\rangle + \frac{1}{6}|c\rangle$ and $\rho = \frac{1}{8}|a\rangle + \frac{1}{4}|b\rangle + \frac{5}{8}|c\rangle$. In ProbLog one can write them as follows.

```
1/2 :: omega(a); 1/3 :: omega(b); 1/6 :: omega(c).
1/8 :: rho(a); 1/4 :: rho(b); 5/8 :: rho(c).

query(omega(_)).
query(rho(_)).
```
(2.65)

The expressions in the first two lines for distributions are called annotated disjunctions in ProbLog. This terminology reflects the logical perspective. The `query` expressions in the last two lines yield the relevant probabilities as output, when the code fragment is executed. This output is of the form:

```
omega(a): 0.5                    rho(a): 0.125
omega(b): 0.33333333             rho(b): 0.25
omega(c): 0.16666667             rho(c): 0.625
```

There are more general forms to define distributions than in (2.65), for instance as families of distributions, via parametrisation using variables. But first we look at how to obtain a tensor / parallel product distribution $\omega\otimes\rho \in \mathcal{D}(\{a,b\}\times\{a,b\})$ and a convex sum $s\cdot\omega+(1-s)\cdot\rho \in \mathcal{D}(\{a,b\})$, for $s = \frac{1}{5}$. This also involves variables `X` and `Y`.

```
omega_tensor_rho(X,Y) :- omega(X), rho(Y).

1/5 :: scalar.
omega_convexsum_rho(X) :- scalar, omega(X).
omega_convexsum_rho(X) :- not scalar, rho(X).

query(omega_tensor_rho(_,_)).
query(omega_convexsum_rho(_)).
```
(2.66)

The expression `1/5 :: scalar` is called a probabilistic fact, corresponding to a flip distribution with bias $\frac{1}{5}$. The prefix `not` is used for the 'other' label of a binary distribution — in this case `scalar`. Thus, the probabilistic fact `1/5 :: scalar` introduces a binary distribution of the form $\frac{1}{5}\big|\texttt{scalar}\big\rangle + \frac{4}{5}\big|\texttt{not scalar}\big\rangle$.

In general, if r is the probability associated with label `a`, then $1-r$ is associated with `not a`. In examples we have been writing $a^{\perp}$ for the 'other' label of a. In ProbLog one can also write `\+a` instead `not a`.

The two `query`'s at the end of (2.66) produce:

```
omega_tensor_rho(a,a): 0.0625
omega_tensor_rho(a,b): 0.125
omega_tensor_rho(a,c): 0.3125
omega_tensor_rho(b,a): 0.041666667
omega_tensor_rho(b,b): 0.083333333
omega_tensor_rho(b,c): 0.20833333
omega_tensor_rho(c,a): 0.020833333
omega_tensor_rho(c,b): 0.041666667
omega_tensor_rho(c,c): 0.10416667
```

```
omega_convexsum_rho(a):
    0.2
omega_convexsum_rho(b):
    0.26666667
omega_convexsum_rho(c):
    0.53333333
```

This corresponds to, in fractional form:

$$\begin{aligned}\omega\otimes\rho &= \tfrac{1}{16}|a,a\rangle+\tfrac{1}{8}|a,b\rangle+\tfrac{5}{16}|a,c\rangle+\tfrac{1}{24}|b,a\rangle+\tfrac{1}{12}|b,b\rangle+\tfrac{5}{24}|b,c\rangle\\ &\quad+\tfrac{1}{48}|c,a\rangle+\tfrac{1}{24}|c,b\rangle+\tfrac{5}{48}|c,c\rangle+\\ s\cdot\omega+(1-s)\cdot\rho &= \tfrac{1}{5}|a\rangle+\tfrac{4}{15}|b\rangle+\tfrac{8}{15}|c\rangle.\end{aligned}$$

The code fragment (2.66) uses two separate mechanisms for defining products and convex sums. An expression on the left (before) the 'neck' operator `:-` looks like a product when there are multiple expressions on the right, as for `omega_tensor_rho`. In fact, this is not really a product but a conjunction, with its own dynamics, see also Remark 2.8.4. If there are multiple clauses with the same expression before `:-`, this expression is defined as a sum, or more accurately as a disjunction. This logical perspective can give unexpected behaviour. For instance, we may use the formulation of (2.66) to define the tensor product $\omega\otimes\omega$ of ω with itselfs, as in:

```
omega_tensor_attempt_omega(X,Y) :- omega(X), omega(Y).
```
(2.67)

Querying it gives:

```
omega_tensor_attempt_omega(a,a): 0.5
omega_tensor_attempt_omega(a,b): 0
omega_tensor_attempt_omega(a,c): 0
omega_tensor_attempt_omega(b,a): 0
omega_tensor_attempt_omega(b,b): 0.33333333
omega_tensor_attempt_omega(b,c): 0
omega_tensor_attempt_omega(c,a): 0
omega_tensor_attempt_omega(c,b): 0
omega_tensor_attempt_omega(c,c): 0.16666667
```

Somewhat surprisingly, this is not the tensor product distribution $\omega\otimes\omega$. It is the pushforward $\mathcal{D}(\Delta)(\omega)$. These two distributions are really different, see Fact 2.3.10 and should not be mixed up. However, the tensor definitions in (2.66) and (2.67) look very much the same in ProbLog. We shall go deeper into what is happening here in Remark 2.8.4 below.

The next code fragment, on the left below, illustrates how to define a joint distribution and its

marginals. Its output is on the right.

```
1/8 :: joint(1,1);
3/8 :: joint(1,2);
1/3 :: joint(2,1);
1/6 :: joint(2,2).

marginal1(X) :- joint(X,Y).
marginal2(Y) :- joint(X,Y).

query(marginal1(_)).
query(marginal2(_)).
```

```
marginal1(1): 0.5
marginal1(2): 0.5
marginal2(1): 0.45833333
marginal2(2): 0.54166667
```

One can also define the convolution sums from Definition 2.7.1 in ProbLog, for distributions defined on commutative monoids. For instance, for distributions `a` and `b` defined on the natural numbers:

```
1/2 :: a(1); 1/3 :: a(2); 1/6 :: a(3).
1/5 :: b(0); 1/5 :: b(1); 1/5 :: b(2); 1/5 :: b(3); 1/5 :: b(4).

a_tensor_b(X,Y) :- a(X), b(Y).
a_convolution_b(Z) :- a_tensor_b(X,Y), Z is X+Y.

query(a_convolution_b(_)).
```
(2.68)

The output takes the form:

```
a_convolution_b(1): 0.1
a_convolution_b(2): 0.1666667
a_convolution_b(3): 0.2
a_convolution_b(4): 0.2
a_convolution_b(5): 0.2
a_convolution_b(6): 0.1
a_convolution_b(7): 0.03333333
```

It corresponds to the convolution sum (2.60):

$$\begin{aligned}&\left(\tfrac{1}{2}|1\rangle + \tfrac{1}{3}|2\rangle + \tfrac{1}{6}|3\rangle\right) + \left(\tfrac{1}{5}|0\rangle + \tfrac{1}{5}|1\rangle + \tfrac{1}{5}|2\rangle + \tfrac{1}{5}|3\rangle + \tfrac{1}{5}|4\rangle\right)\\ &= \tfrac{1}{10}|1\rangle + \tfrac{1}{6}|2\rangle + \tfrac{1}{5}|3\rangle + \tfrac{1}{5}|4\rangle + \tfrac{1}{5}|5\rangle + \tfrac{1}{10}|6\rangle + \tfrac{1}{30}|7\rangle.\end{aligned}$$

We turn to channels and illustrate how they are represented in ProbLog in a quite natural manner via probabilitic clauses. In fact, these clauses can be understood as descriptions of channels. We reconsider the distribution $\omega = \frac{1}{6}|x\rangle + \frac{1}{2}|y\rangle + \frac{1}{3}|z\rangle$ from Exercise 2.4.1 with channels $c\colon \{x,y,z\} \rightsquigarrow \{u,v\}$ and $d\colon \{u,v\} \rightsquigarrow \{1,2,3,4\}$. The distribution and channels are described in the

following way.

```
1/6 :: omega(x); 1/2 :: omega(y); 1/3 :: omega(z).

1/2 :: c(u); 1/2 :: c(v) :- omega(x).
1   :: c(u); 0   :: c(v) :- omega(y).
3/4 :: c(u); 1/4 :: c(v) :- omega(z).

1/4 :: d(1); 1/8 :: d(2); 1/2 :: d(3); 1/8 :: d(4) :- c(u).
1/4 :: d(1); 0   :: d(2); 1/8 :: d(3); 5/8 :: d(4) :- c(v).

query(c(_)).
query(d(_)).
```
(2.69)

What is noteworthy is that the channel `c` is defined in combination with the distribution `omega`, to which it can be applied, via pushforward of the distribution, that is, via distribution transformation / prediction. The application to `omega` is built into the channel clauses for `c`. The same applies to the channel `d`, which is defined in combination with `c`. Thus, channels are not defined as separate entities, but as combinations (c, ω), together with a distribution ω. To be precise, the channel definition uses the names from the distribution, not its probabilities. Parametrisation offers a bit more flexibility, but the coupling of a channel with the (naming of the) distribution on which it acts is an integral part of the ProbLog language. It limits the abstraction level of the language, like we have seen above above for tensors.

The output of the first query in the code fragment (2.69) describes the pushforward $c_*(\omega)$, via probabilities:

```
c(u): 0.83333333          c(v): 0.16666667
```

The second query `query(d(_))` at the end of (2.69) performs the next pushforward $d_*(c_*(\omega)) = (d \circ c)_*(\omega)$, with outcomes:

```
d(1): 0.25                d(3): 0.4375
d(2): 0.10416667          d(4): 0.20833333
```

This corresponds to the predictions $c_*(\omega) = \frac{5}{6}|u\rangle + \frac{1}{6}|v\rangle$ and $d_*(c_*(\omega)) = \frac{1}{4}|1\rangle + \frac{5}{48}|2\rangle + \frac{7}{16}|3\rangle + \frac{5}{24}|4\rangle$. We thus see how channels and pushforward allong channels is an integral part of ProbLog, albeit in slightly different form.

Having seen these illustrations we come to a few more general observations about ProbLog.

Remarks 2.8.1 1 The probabilities $\frac{1}{2}, \frac{1}{3}, \frac{1}{6}$ for the distribution `omega` in (2.65) obviously add up to 1. ProbLog gives an error message if the sum of the listed numbers is strictly bigger than 1. But it does allow the sum to be less than 1. Thus, ProbLog is really about *subdistributions*. However, in our illustrations we generally use proper distributions, with probabilities adding up to 1.

2 A general rule is that missing cases get probability 0. This is the so-called closed world assumption of logic programming, see *e.g.* Fierens et al. (2015): what is not known to be true must be false. For instance, in (2.69), the cases `0 :: c(v)` and `0 :: d(2)` may be omitted, without affecting the outcome. Even stronger, one can drop the entire line second line defining

the channel `d` in (2.69). Then all its cases are set to 0, so that `d` on `v` becomes the zero distribution. Ommitting this second line does affect the outcome, making $d_*(c_*(\omega))$ a subdistribution, with probabilities:

```
d(1): 0.20833333          d(3): 0.41666667
d(2): 0.10416667          d(4): 0.10416667
```

In general, we do write the 0-cases in order to be explicit and prevent possible confusion.

3 ProbLog has some special arrangements for binary distributions, that is, for distributions in $\mathcal{D}(B)$ where B is a two-element set, for instance $B = \mathbf{2} = \{0,1\}$ or $B = \{b, b^\perp\}$. In these situations the distribution is determined by one of its probabilities, say r, since the other one is necessarily $1-r$. This works since there is an isomorphism $\mathcal{D}(\mathbf{2}) \cong [0,1]$. Where we often write $b^\perp$ to denote the 'other' element than b, ProbLog uses the prefix `not` as we have seen in Example (2.66). Accordingly, the two statements below have the same meaning in ProbLog.

```
1/4 :: b.          1/4 :: b; 3/4 :: not b.
```

They both define[3] the binary distribution $\frac{1}{4}|\texttt{b}\rangle + \frac{3}{4}|\texttt{not b}\rangle$, often written as $\frac{1}{4}|b\rangle + \frac{3}{4}|b^\perp\rangle$. These binary distributions are convenient, as will be illustrated in the Bayesian network below. But one has to be a bit careful. For instance, omitting the case for `b` does not set everything to 0, like for non-binary distributions as just illustrated for `d` in the previous item. Omitting the case for `b` does set `b` to 0 but it also sets `not b` to 1. This applies, for instance, to the last lines in the definition of `wet_grass` and `slippery_road` in Figure 2.8.

Another point of care is the sum, that we used in (2.66) to define the convex sum. For binary distributions this sum is a 'noisy or', as illustrated in Exercise 2.8.1 below.

These binary distributions in ProbLog can also be seen as 'fuzzy' predicates, since channels $X \twoheadrightarrow \mathbf{2}$ can be identified with functions $X \to [0,1]$. Such functions to $[0,1]$ are studied systematically in Chapter 4 as predicates in probabilistic logic. There we shall see more ProbLog examples.

The wetness Bayesian network from Subsection 2.5.2 can be reformulated in ProbLog, see Figure 2.8 (and also Raedt and Kimmig (2015)). It uses binary distributions only. The conditional probability tables of the Bayesian network translate almost immediately into the probabilistic clauses in ProbLog (defining channels). The queries at the end in this figure immediately produce the probabilties that we computed via pushforwards in Subsection 2.5.2:

```
winter: 0.6          sprinkler: 0.42          rain: 0.52
    slippery_road: 0.364          wet_grass: 0.6995
```

We see that ProbLog is convenient for writing probabilistic models and for computing outcomes. The mechanics, however, for computing these outcomes, remains hidden.

[3] There are some peculiarities: writing `1/4 :: b; 3/2 :: not b.` gives an error message, since the sum exceeds 1. This makes sense. Writing `1/4 :: b; 1/2 :: not b.` does not give an error, but subsequently querying `\+b` gives $\frac{3}{4}$ as output.

```
3/5 :: winter.

1/5 :: sprinkler :- winter.
3/4 :: sprinkler :- not winter.

4/5  :: rain :- winter.
1/10 :: rain :- not winter.

19/20 :: wet_grass :- sprinkler,     rain.
9/10  :: wet_grass :- sprinkler,     not rain.
4/5   :: wet_grass :- not sprinkler, rain.
0     :: wet_grass :- not sprinkler, not rain.

7/10 :: slippery_road :- rain.
0    :: slippery_road :- not rain.

query(winter).
query(sprinkler).
query(rain).
query(wet_grass).
query(slippery_road).
```

Figure 2.8 ProbLog version of the wetness Bayesian network from Subsection 2.5.2.

2.8.1 Variables in ProbLog

Definitions in ProbLog may be parametrised by — or indexed by — variables. These variables are written in capital letters. An example is the variable `V` describing a family of distributions in:

```
1/9 :: rho(x,V); 1/3 :: rho(y,V); 5/9 :: rho(z,V) :- rho(V).

rho(a).
rho(b).

query(rho(_,_)).
```

The lines `rho(a).` and `rho(b).` form concrete instantiations, with constants `a` and `b` put in place of the variable `V`. There may be arbitrarily many such instantiations. The query in the last line produces the following output.

```
rho(x,a): 0.11111111        rho(x,b): 0.11111111
rho(y,a): 0.33333333        rho(y,b): 0.33333333
rho(z,a): 0.55555556        rho(z,b): 0.55555556
```

The family of distributions `rho(V)` defined above is of the same form, in the sense that all distributions have three elements in their support with the same three probabilities. ProbLog offers

more flexibility, as can be seen from the following illustration.

```
1/2 :: tau(a,1); 1/3 :: tau(b,1); 1/6 :: tau(c,1) :- tau(1).
1/3 :: tau(u,2); 2/3 :: tau(v,2) :- tau(2).
1/4 :: tau(w,3) :- tau(3).

tau(1).
tau(2).
tau(3).

query(tau(_,_)).
```

We now get as output:

```
tau(a,1): 0.5                tau(u,2): 0.33333333
tau(b,1): 0.33333333         tau(v,2): 0.66666667
tau(c,1): 0.16666667         tau(w,3): 0.25
```

Thus, we have different supports of the three distributions $\tau(\_, i)$, with their support on the left, and the distribution itself on the right.

$$\begin{array}{rcl@{\qquad}rcl} supp(\tau(\_,1)) & = & \{x,y,z\} & \tau(\_,1) & = & \frac{1}{2}|x\rangle + \frac{1}{3}|y\rangle + \frac{1}{6}|z\rangle \\ supp(\tau(\_,2)) & = & \{u,v\} & \tau(\_,2) & = & \frac{1}{3}|u\rangle + \frac{2}{3}|v\rangle \\ supp(\tau(\_,3)) & = & \{w,w^{\perp}\} & \tau(\_,3) & = & \frac{1}{4}|w\rangle + \frac{3}{4}|w^{\perp}\rangle. \end{array}$$

In order to accomodate this variation semantically we need to work with indexed sets, of the form $(X_i)_{i\in I}$, where I is an index set and where X_i is a set for each element $i \in I$. In the above situation we would have $I = \{1,2,3\}$ with set X_i given by the support of $\tau(\_, i)$.

$$X_1 = \{x,y,z\} \qquad X_2 = \{u,v\} \qquad X_3 = \{w,w^{\perp}\}.$$

The next definition introduces such indexing in categorical form, first as $\mathbb{C}^I$ for an arbitrary category $\mathbb{C}$. This construction is then used where $\mathbb{C}$ is a category of probabilistic channels. When I is a finite set, say with n elements, then $\mathbb{C}^I$ is the n-fold product of categories $\mathbb{C} \times \cdots \times \mathbb{C}$.

Definition 2.8.2 Let I be an arbitrary set and $\mathbb{C}$ an arbitrary category. We write $\mathbb{C}^I$ for the I-indexed product of $\mathbb{C}$. It is a new category with:

- objects of $\mathbb{C}^I$ are I-indexed collections $(X_i)_{i\in I}$ of objects $X_i \in \mathbb{C}$;
- a morphism $f\colon (X_i)_{i\in I} \to (Y_i)_{i\in I}$ in $\mathbb{C}^I$ is an I-indexed collection $f = (f_i)_{i\in I}$ of morphisms $f_i\colon X_i \to Y_i$ in $\mathbb{C}$.

Composition in $\mathbb{C}^I$ is done pointwise: $g \circ f = (g_i \circ f_i)_{i\in I}$. Clearly, the identity map $id\colon (X_i)_{i\in I} \to (X_i)_{i\in I}$ in $\mathbb{C}^I$ is the I-indexed collection of identity maps $(id\colon X_i \to X_i)_{i\in I}$.

The mapping $I \mapsto \mathbb{C}^I$ yields a functor $\mathbf{Sets}^{\mathrm{op}} \to \mathbf{Cat}$, since for a function $h\colon I \to J$ between index sets there is a 'substitution' function $h^{-1}\colon \mathbb{C}^J \to \mathbb{C}^I$. It is given by $h^{-1} = (-)\colon h$, both on objections and on morphisms, as in:

$$h^{-1}\big((X_j)_{j\in J}\big) = (X_{h(i)})_{i\in I} \qquad h^{-1}\big((f_j)_{j\in J}\big) = (f_{h(i)})_{i\in I}.$$

Example 2.8.3 Let's elaborate what the I-indexed product $\mathbf{Chan}^I$ of the category of (probabilistic) channels is. Its maps turn out to be 'dependent' channels. But first, an obect of the category $\mathbf{Chan}^I$ is an I-indexed collection $(X_i)_{i\in I}$ of sets X_i. A morphism $c\colon (X_i)_{i\in I} \to (Y_i)_{i\in I}$ in $\mathbf{Chan}^I$ is a collection of morphisms $c_i\colon X_i \rightsquigarrow Y_i$ in $\mathbf{Chan}$, and thus of functions $c_i\colon X_i \to \mathcal{D}(Y_i)$. In dependent type theory one would write that such a morphism c has type:

$$c : \prod_{i\in I} \big(X_i \to \mathcal{D}(Y_i)\big).$$

Composition in $\mathbf{Chan}^I$, also written as $\odot$, is pointwise channel composition: $d \odot c = (d_i \odot c_i)_{i\in I}$.

As ProbLog uses subdistributions $\mathcal{D}_{\leq 1}$ instead of proper distributions $\mathcal{D}$, we will have to use the category $\mathbf{Chan}^I_{\leq 1}$ instead of $\mathbf{Chan}^I$, which channels of the form $X_i \to \mathcal{D}_{\leq 1}(Y_i)$.

Our aim in the remainder of this section is to sketch how such I-indexed categories $\mathbf{Chan}^I_{\leq 1}$ can be used to give meaning to ProbLog programs with variables. We do so by elaborating an example, not by giving a systematic interpretation of all language constructs. In fact, we concentrate on distributions, channels and composition. An alternative, more comprehensive account, using coalgebras can be found in Gu and Zanasi (2019), building on Komendantskaya and Power (2011).

We consider the following example from Raedt et al. (2007).

```
1.0:: likes(X,Y):- friendof(X,Y).
0.8:: likes(X,Z):- friendof(X,Y), likes(Y,Z).

0.5:: friendof(john,mary).
0.5:: friendof(mary,pedro).
0.5:: friendof(mary,tom).
0.5:: friendof(pedro,tom).

query(likes(john,tom)).
```
(2.70)

The query in this program yields as probability 0.24. Our aim is to reconstruct this probability via indexed semantics.

The `likes` and `friendof` expressions correspond to $I \times I$-indexed binary sets, written as L and F, of the form:

$$L = \big(\{\ell_{i,j}, \ell^{\perp}_{i,j}\}\big)_{i,j\in I} \qquad \text{and} \qquad F = \big(\{f_{i,j}, f^{\perp}_{i,j}\}\big)_{i,j\in I}.$$

We leave it open what the set I is, but we will assume that it has certain elements, corresponding to the names in the program fragment (2.70).

We interprete the first two lines the program as indexed 'sub' channels. The first line gives a map $c\colon L \to F$ in $Chan^{I\times I}$ consisting of channels $c_{i,j}\colon L_{i,j} \to \mathcal{D}(F_{i,j})$. Concretely as:

$$c_{i,j}(\ell_{i,j}) := 1\big|f_{i,j}\rangle \qquad\qquad c_{i,j}(\ell^{\perp}_{i,j}) := 1\big|f^{\perp}_{i,j}\rangle.$$

The second equation follows since there is a missing binary clause.

The interpretation of the second line is more subtle, since it involves substitution. We write $\pi_{1,2}, \pi_{2,3}, \pi_{1,3}\colon I \times I \times I \to I \times I$ for the obvious projection maps:

$$\pi_{1,2} = \langle \pi_1, \pi_2 \rangle \qquad \pi_{2,3} = \langle \pi_2, \pi_3 \rangle \qquad \pi_{1,3} = \langle \pi_1, \pi_3 \rangle.$$

We can now interprete the second line in (2.70) as a channel d in the category $\mathbf{Chan}^{I\times I\times I}$ of the form:

$$\pi_{1,2}^{-1}(F)\times\pi_{2,3}^{-1}(L)\xrightarrow{\quad d\quad}\pi_{2,3}^{-1}(L)$$

Thus, this d is given by channels, for $i,j,k\in I$,

$$L_{i,j}\times F_{j,k}\xrightarrow{d_{i,j,k}}\mathcal{D}(L_{i,k})\quad\text{where}\quad\begin{cases}d_{i,j,k}(\ell_{i,j},f_{j,k}) := \frac{4}{5}\big|\ell_{i,k}\rangle+\frac{1}{5}\big|\ell^{\perp}_{i,k}\rangle\\ d_{i,j,k}(x,y) := 1\big|\ell^{\perp}_{i,k}\rangle,\quad\text{otherwise.}\end{cases}$$

The four `friendsof` lines in (2.70) are translated as follows. We assume elements jo, ma, pe, to in the index set I, together with uniform distributions:

$$\begin{aligned}\omega_1 &:= \tfrac{1}{2}\big|f_{jo,ma}\rangle+\tfrac{1}{2}\big|f^{\perp}_{jo,ma}\rangle & \omega_3 &:= \tfrac{1}{2}\big|f_{ma,to}\rangle+\tfrac{1}{2}\big|f^{\perp}_{ma,to}\rangle\\ \omega_2 &:= \tfrac{1}{2}\big|f_{ma,pe}\rangle+\tfrac{1}{2}\big|f^{\perp}_{ma,pe}\rangle & \omega_4 &:= \tfrac{1}{2}\big|f_{pe,to}\rangle+\tfrac{1}{2}\big|f^{\perp}_{pe,to}\rangle.\end{aligned}$$

There are now two routes to arrive at the liking between john and tom in the query, both going via john and mary liking each other. Such a situation is common in (probabilistic) logic programming. The two routes will be expressed via two distributions $\sigma_1,\sigma_2\in\mathcal{D}(\{\ell_{ma,to},\ell^{\perp}_{ma,to}\})$ about mary and tom liking each other.

- We can directly apply the channel c to the friendship distribution between mary and tom, giving:

$$\sigma_1 = (c_{ma,to})_*(\omega_3) = (c_{ma,to})_*(\tfrac{1}{2}\big|f_{ma,to}\rangle+\tfrac{1}{2}\big|f^{\perp}_{ma,to}\rangle) = \tfrac{1}{2}\big|\ell_{ma,to}\rangle+\tfrac{1}{2}\big|\ell^{\perp}_{ma,to}\rangle.$$

- There is an alternative route via pedro:

$$\begin{aligned}\sigma_2 &:= (d_{ma,pe,to})_*\Big(\omega_2\otimes(c_{pe,to})_*(\omega_4)\Big)\\ &= (d_{ma,pe,to})_*\Big((\tfrac{1}{2}\big|f_{ma,pe}\rangle+\tfrac{1}{2}\big|f^{\perp}_{ma,pe}\rangle)\otimes(c_{pe,to})_*(\tfrac{1}{2}\big|f_{pe,to}\rangle+\tfrac{1}{2}\big|f^{\perp}_{pe,to}\rangle)\Big)\\ &= (d_{ma,pe,to})_*\Big((\tfrac{1}{2}\big|f_{ma,ma}\rangle+\tfrac{1}{2}\big|f^{\perp}_{ma,ma}\rangle)\otimes(\tfrac{1}{2}\big|\ell_{pe,to}\rangle+\tfrac{1}{2}\big|\ell^{\perp}_{pe,to}\rangle)\Big)\\ &= (d_{ma,pe,to})_*\Big(\tfrac{1}{4}\big|f_{ma,pe},\ell_{pe,to}\rangle+\tfrac{1}{4}\big|f_{ma,pe},\ell^{\perp}_{pe,to}\rangle+\tfrac{1}{4}\big|f^{\perp}_{ma,pe},\ell_{pe,to}\rangle+\tfrac{1}{4}\big|f^{\perp}_{ma,pe},\ell^{\perp}_{pe,to}\rangle\Big)\\ &= \tfrac{1}{5}\big|\ell_{ma,to}\rangle+\tfrac{4}{5}\big|\ell^{\perp}_{ma,to}\rangle.\end{aligned}$$

This or channel oc sends a pair x,y to ℓ as soon as either x or y is ℓ.

Now that we have two binary distribution $\sigma_1,\sigma_2\in\mathcal{D}(\{\ell_{ma,to},\ell^{\perp}_{ma,to}\})$ about mary and tom liking each other, we combine them via the 'noisy' or channel oc from Exercises 2.5.9 and 2.8.1, adapted to the two-element space of likes and non-likes:

$$\begin{aligned}\sigma := oc_*\Big(\sigma_1\otimes\sigma_2\Big) &= oc_*\Big((\tfrac{1}{2}\big|\ell_{ma,to}\rangle+\tfrac{1}{2}\big|\ell^{\perp}_{ma,to}\rangle)\otimes(\tfrac{1}{5}\big|\ell_{ma,to}\rangle+\tfrac{4}{5}\big|\ell^{\perp}_{ma,to}\rangle)\Big)\\ &= oc_*\Big(\tfrac{1}{10}\big|\ell_{ma,to},\ell_{ma,to}\rangle+\tfrac{2}{5}\big|\ell_{ma,to},\ell^{\perp}_{ma,to}\rangle\\ &\qquad+\tfrac{1}{10}\big|\ell^{\perp}_{ma,to},\ell_{ma,to}\rangle+\tfrac{2}{5}\big|\ell^{\perp}_{ma,to},\ell^{\perp}_{ma,to}\rangle\Big)\\ &= \tfrac{3}{5}\big|\ell_{ma,to}\rangle+\tfrac{2}{5}\big|\ell^{\perp}_{ma,to}\rangle.\end{aligned}$$

We can now determine the liking distribution between john and tom via a final application of the

d channel:

$$\begin{aligned}
\tau \;\coloneqq\; (d_{jo,ma,to})_*\big(\,\omega_1 \otimes \sigma\,\big) &= (d_{jo,ma,to})_*\Big(\,(\tfrac{1}{2}\big|f_{jo,ma}\rangle + \tfrac{1}{2}\big|f^{\perp}_{jo,ma}\rangle) \otimes (\tfrac{3}{5}\big|\ell_{ma,to}\rangle + \tfrac{2}{5}\big|\ell^{\perp}_{ma,to}\rangle)\Big) \\
&= (d_{jo,ma,to})_*\Big(\,\tfrac{3}{10}\big|f_{jo,ma}, \ell_{ma,to}\rangle + \tfrac{1}{5}\big|f_{jo,ma}, \ell^{\perp}_{ma,to}\rangle \\
&\qquad\qquad + \tfrac{3}{10}\big|f^{\perp}_{jo,ma}, \ell_{ma,to}\rangle + \tfrac{1}{5}\big|f^{\perp}_{jo,ma}, \ell^{\perp}_{ma,to}\rangle\Big) \\
&= \tfrac{6}{25}\big|\ell_{jo,to}\rangle + \tfrac{19}{25}\big|\ell^{\perp}_{jo,to}\rangle \\
&= 0.24\big|\ell_{jo,to}\rangle + 0.76\big|\ell^{\perp}_{jo,to}\rangle.
\end{aligned}$$

This 0.24 probability that john likes tom is indeed the outcome of (2.70) that ProbLog computes.

There is one thing that we like to make explicit, namely about when to apply the noisy-or *oc*. ProbLog does this *early*, as soon as two distributions, namely σ_1 and σ_2 on the same set are available. One can also apply noisy-or in a *late* manner, when all the liking distributions about john and tom are available. This gives a different outcome, namely:

$$\begin{aligned}
\tau' \;\coloneqq\; & oc_*\Big((d_{jo,ma,to})_*(\,\omega_1 \otimes \sigma_1\,) \otimes (d_{jo,ma,to})_*(\,\omega_1 \otimes \sigma_2\,)\Big) \\
&= \tfrac{33}{125}\big|\ell_{jo,to}\rangle + \tfrac{92}{125}\big|\ell^{\perp}_{jo,to}\rangle \\
&= 0.264\big|\ell_{jo,to}\rangle + 0.736\big|\ell^{\perp}_{jo,to}\rangle.
\end{aligned}$$

It is part of the underlying semantical model that ProbLog favours early application of noisy-or over late application. It is a topic that will return later on, see Chapter 8.

Remark 2.8.4 We have seen different behaviour for the seemingly equal tensor descriptions in the code fragements (2.66) and (2.67), for $\omega \otimes \rho$ and $\omega \otimes \omega$ We give a brief account about what is going on, with references to the literature Fierens et al. (2015); Derkinderen et al. (2024) on ProbLog semantics for more details[4]. In general, a ProbLog program specifies a (sub)distribution over possible worlds, consisting of subsets of 'ground facts', given by atomic expressions, possibly with clauses (on the right of a neck `:-`). The assignment of probabilities happens modulo logical entailment, written as $\models$, see Derkinderen et al. (2024), see esp. Eqn. (4). A subset F of ground facts gets as probability the product of probabilities $P(f)$ of $f \in F$ multiplied with the product of $1 - P(g)$ for $g \notin F$. The latter product corresponds to the closed world assumption: what is absent is assumed to be false.

We first look at the following simple examples with Booleans `a` and `b` with probabilities.

```
3/10 :: a.  2/5  :: b.

test1 :- not a, b, b.

test2 :- a.
test2 :- not b.

test3 :- a, not b.
test3 :- not a, b.
```

We are interested in the outcomes of querying the above tests. We shall describe how their interpretations work in ProbLog via truth tables.

[4] Thanks are due to Vincent Derkinderen from the ProbLog team for kindly providing additional explanation in personal communication (nov. 2024).

The first test deals with the conjunction $\neg a \wedge b \wedge b$ which can be interpreted via the following table with possible worlds.

a	b	`test1`	**probabilities**
1	1	0	0
1	0	0	0
0	1	1	$(1-\frac{3}{10})\cdot\frac{2}{5} = \frac{7}{25}$
0	0	0	0
			total: $\frac{7}{25}$

This total probability $\frac{7}{25} = 0.28$ is the outcome of querying `test1`.

The second test involves a disjunction $\neg a \vee b$, which gives rise to the following table.

a	b	`test2`	**probabilities**
1	1	1	$\frac{3}{10}\cdot\frac{2}{5} = \frac{6}{50}$
1	0	1	0
0	1	1	$(1-\frac{3}{10})\cdot\frac{2}{5} = \frac{14}{50}$
0	0	0	$(1-\frac{3}{10})\cdot(1-\frac{2}{5}) = \frac{21}{50}$
			total: $\frac{41}{50}$

Similarly, for the thrid test we have a formula $(a \wedge \neg b) \vee (\neg a \wedge b)$, with table:

a	b	`test3`	**probabilities**
1	1	0	0
1	0	1	$\frac{3}{10}\cdot(1-\frac{2}{5}) = \frac{9}{50}$
0	1	1	$(1-\frac{3}{10})\cdot\frac{2}{5} = \frac{14}{50}$
0	0	0	
			total: $\frac{23}{50}$

We turn to probability distributions, where we recondider the distribution $\omega = \frac{1}{2}|a\rangle + \frac{1}{3}|b\rangle + \frac{1}{6}|c\rangle$ from (2.66).

```
1/2 :: omega(a); 1/3 :: omega(b); 1/6 :: omega(c).

test4 :- omega(a), omega(b).

test5 :- omega(a), not omega(b), not omega(c).

test6 :- omega(a), omega(a), omega(c).
test6 :- not omega(b).

test7 :- not omega(a), not omega(b).

test8 :- not omega(a).
test8 :- not omega(b).
```

When we query the fourth test we get zero as outcome. This may be somewhat surprising, since

the probability $\frac{1}{2} \cdot \frac{1}{3}$ seems in line with the outcomes of the first four tests. However, the facts `omega(a)`, `omega(b)` (and `omega(c)`) of the distribution are logically incompatible in ProbLog, since they are understood as annotated disjunctions (see Mcareavey et al. (2017), esp. Ex. 1). As a result, the fifth test yields the probability $\frac{1}{2}$ of `omega(a)` since the negation of the other options does not add anything. The sixth test yields the orthocomplement $1 - \frac{1}{3} = \frac{2}{3}$ of `omega(b)`, which includes the option `omega(a)`. The double occurrence of `omega(a)` has no effect, in line with the above observations. Removing `omega(c)` in the first line has no effect on this test outcome. The seventh test yields the probability $\frac{1}{6}$ of the remaining option `omega(c)`. The eight and last test gives probability 1, since it covers all three options of the distribution ω.

In light of these observations it becomes clearer why the attempted definition of the tensor product $\omega \otimes \omega$ in (2.67) fails: the combinations `omega(X), omega(Y)` are incompatible unless `X` and `Y` are equal. Thus the outcome is $\mathcal{D}(\Delta)(\omega)$.

One can properly define a tensor product of a distribution with itself in ProbLog via parametrisation, as described below.

```
1/8 :: tau(a,V); 5/8 :: tau(b,V); 1/4 :: tau(c,V) :- tau(V).

tau(1).  tau(2).

tau_tensor_tau(X,Y) :- tau(X,1), tau(Y,2).
```

Within the setting of this work the elements in the support of a tensor product distribution are pairs (x_i, y_j) occurring in ket expressions $|x_i, y_j\rangle$ in the product. Interaction between these elements x_i and y_j via logical dynamics is undesired. Hence we have to be careful with models in ProbLog, especially in situations where distributions or channels are used multiple times. Then we should use parametrised definitions with different instances, like for `tau` above, in order to prevent unintended interactions (and outcomes).

Exercises

2.8.1 Consider the following simple ProbLog program involving four binary flip distribution with bias as indicated, and with new distributions `flip23` and `flip2345` defined as sum and product:

```
1/2 :: flip2.
1/3 :: flip3.
1/4 :: flip4.
1/5 :: flip5.

flip23 :- flip2.
flip23 :- flip3.

flip2345 :- flip23, flip4, not flip5.

query(flip23).
query(flip2345).
```

It prints:

```
flip23: 0.66666667          flip2345: 0.13333333
```

Let's write $\omega_i = \frac{1}{i}|1\rangle + (1 - \frac{1}{i})|0\rangle \in \mathcal{D}(\mathbf{2})$ for the distribution corresponding to `flipi` in this program, for $i = 2, 3, 4, 5$.

1 Show that the above probability of `flip23` is obtained via the 'noisy' or-channel $oc\colon \mathbf{2} \times \mathbf{2} \rightsquigarrow \mathbf{2}$ from Exercise 2.5.9, as:

$$\omega_{23} \coloneqq oc_*(\omega_2 \otimes \omega_3).$$

2 Show also that the probability for `flip2345` is given by:

$$\omega_{2345} \coloneqq (ac \circ (ac \otimes nc))_*(\omega_{23} \otimes \omega_4 \otimes \omega_5),$$

where $ac\colon \mathbf{2} \times \mathbf{2} \rightsquigarrow \mathbf{2}$ and $nc\colon \mathbf{2} \rightsquigarrow \mathbf{2}$ are the and-channel and negation-channel from Exercise 2.5.9.

2.8.2 Recall the first fault tree from Exercise 2.5.9 and convince yourself that its outcome is given by the following ProbLog program.

```
1/2 :: leaf1. 1/2 :: leaf2. 1/2 :: leaf3. 1/2 :: leaf4.

node1 :- leaf1.
node1 :- leaf2.
node2 :- leaf2.
node2 :- leaf3.
node3 :- node1, node2.
node4 :- node3, leaf4.

query(node4).
```

2.9 A glimpse of probabilistic programming using WebPPL

In this section we take a brief look at a second probabilistic programming language, namely WebPPL Goodman and Stuhlmüller (2014); Goodman and Tenenbaum (2016). It is a functional language that is embedded in the popular general-purpose programming language Javascript. We shall assume some basic level of familiarity with Javascript. As for ProbLog in the previous section, we only scratch the surface and concentrate on how the notions and ideas in this book appear (or can be represented) in the language WebPPL.

WebPPL arises via a special Javascript library in which various probability distributions — both discrete and continuous — are pre-defined, together with suitable built-in inference mechanisms. The most relevant type of distribution is called 'Categorical', corresponding to what is a finite discrete probability distribution in this setting. Such a distribution is defined in WebPPL as:

```
var w = Categorical({ ps:[r1,r2,r3,r4], vs:[x1,x2,x3,x4] })
```

In more mathemtical style, this corresponds to defining a distribution of the form:

$$\omega = r_1|x_1\rangle + r_2|x_2\rangle + r_3|x_3\rangle + r_4|x_4\rangle.$$

In the above WebPPL definition, the expression `[r1,r2,r3,r4]` is a list of numbers `ri`. These numbers should be non-negative (*i.e.* ≥ 0), but need not add up to one: they are automatically normalised to form a distribution — like in frequentist learning. The expression `[x1,x2,x3,x4]` is a list of elements, acting as support. The lists of probabilities and elements are combined into what is called a record object in Javascript, of the form `{ps: .., vs: ..}`.

An operation for obtaining the support of such a categorical distribution exists, as illustrated below. The probabilities associated with elements in the support may be obtained via the `score` function, except that it produces the natural logarithm of the probabilities involved. For instance, the following WebPPL code snippet writes the relevant support and one probability (in two forms) to the console — which correponds to printing.

```
var w = Categorical({ps:[1/2, 1/3, 1/6], vs:['a','b','c']});

console.log( w.support() );
console.log( w.score('a') );
console.log( Math.exp(w.score('a')) );
```

This prints:

```
[ 'a', 'b', 'c' ]
-0.6931471805599451
0.5000000000000001
```

The conversion of the probability $1/2$ to a (natural) logarithm $\ln(1/2) = -0.6931...$ and back leads to some loss of precision.

A big advantage of the fact that WebPPL is embedded in a programming language (namely Javascript) means that we may use this language to add our own functions. For instance, the code snippet below introduces our favourite ket-notation for (discrete) WebPPL distributions.

```
var as_kets = function (dist) {
    // number of decimals
    var precision = 3;
    var supp = dist.support();
    var kets = map(function(s){
        var prob = Math.exp(dist.score(s));
        var str = s.toString();
        var rounded_prob =
           Math.round(Math.pow(10, precision) * prob)
              / Math.pow(10, precision);
        return rounded_prob.toString() + '|' + str + '>'},
                  supp)
    return kets.join(' + ');
```

(2.71)

This Javascript snippet defines the function `as_kets` that takes a distribution `dist` as argument and produces a string, obtained by applying a function, as first argument of `map`, to the support

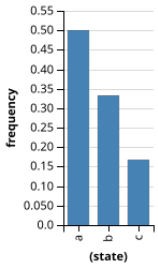
list of the distribution. The latter function sends an element `s` from the support to the string `r|s>`, where `r` is the rounded probability associated with `s`. The precision of the rounding is included in the beginning of the function. In the end the individual strings are joined via the plus sign.

In the above code fragment defining the categorical distribution `w` we can now add the following two lines:

```
console.log( as_kets(w) )
viz( w )
```

This leads to the outcomes given below, with the ket-representation on the left and WebPPL's visualisation (via the pre-defined `viz`) on the right.

```
0.5|a> + 0.333|b> + 0.167|c>
```



In the sequel we shall use both forms to describe WebPPL distributions.

Given a WebPPL distribution one can sample from it, that is, obtain an element from the support, where the choice is determined by the probabilities of the all elements in the support, see Subsection 2.2.1. Again for the above distribution `w` we can write:

```
console.log( sample(w) )
console.log( repeat(10, function() { sample(w) }) )
```

It gives a single sample together with a list of 10 samples — not a multiset — which may be of the form:

```
a
[ 'a', 'a', 'b', 'b', 'c', 'a', 'a', 'b', 'b', 'a' ]
```

Different runs typically give different outcomes. When we accumulate this particular list of 10 elements to a multiset we get $5|a\rangle + 4|b\rangle + 1|c\rangle$; when we then apply frequentist learning we obtain the distribution $\frac{1}{2}|a\rangle + \frac{2}{5}|b\rangle + \frac{1}{10}|c\rangle$. It approximates the original distribution `w`. This approximation gets better as the number of samples grows.

A typical feature probabilistic programming languages is the combination of sampling and frequentist learning. In WebPPL this is done via `Infer`, as illustrated below.

```
var flrn_acc = function(list) { Infer(function () {
        return uniformDraw(list); })}

console.log( as_kets(flrn_acc(['a','b','a','c','b','a'])) )
```

This inference produces `0.5|a> + 0.333|b> + 0.167|c>`, which happens to be the distribution `w`. In the words of (Goodman and Tenenbaum, 2016, §2), "`Infer` lets us reify the distribution implicitly described by a sampling process".

If we first sample and then reify, as below, we recover the original distribution.

```
console.log( as_kets(Infer( function() { sample(w) })) )
```

Conversely, if we sample from a reified distribution, we may as well use the original sampling process.

Inference in WebPPL via `Infer` is a powerful mechanism, see Remark 2.9.1 below for some more information. It can be used to combine distributions into new ones. For instance, the convolution sum of two distributions can be defined in WebPPL very much like in (2.61), on the right. Notice that WebPPL allows us to define convolution as a function, with distributions as inputs. In contract, in ProbLog, convolution is defined specifically for particular distributions, see (2.68).

```
var convolution = function(dist1, dist2){ Infer(function(){
        var s1 = sample(dist1);
        var s2 = sample(dist2);
        return s1 + s2; }) }

var v = Categorical({ps: [1/4, 3/4], vs: [2,3]});
var b = Binomial({n:5, p:0.5})
console.log( as_kets( convolution(v,b) ) )
```

The last line gives:

```
0.008|2> + 0.063|3> + 0.195|4> + 0.313|5> + 0.273|6> + 0.125|7> + 0.023|8>
```

In the notation of this book this same outcome can be calculated precisely, using the convolution sum + from Definition 2.7.1, namely as:

$$
\begin{aligned}
&\left(\tfrac{1}{4}|2\rangle + \tfrac{3}{4}|3\rangle\right) + bn[5](\tfrac{1}{2}) \\
&= \left(\tfrac{1}{4}|2\rangle + \tfrac{3}{4}|3\rangle\right) + \left(\tfrac{1}{32}|0\rangle + \tfrac{5}{32}|1\rangle + \tfrac{5}{16}|2\rangle + \tfrac{5}{16}|3\rangle + \tfrac{5}{32}|4\rangle + \tfrac{1}{32}|5\rangle\right) \\
&\overset{(2.60)}{=} \tfrac{1}{128}|2{+}0\rangle + \tfrac{5}{128}|2{+}1\rangle + \tfrac{5}{64}|2{+}2\rangle + \tfrac{5}{64}|2{+}3\rangle + \tfrac{5}{128}|2{+}4\rangle + \tfrac{1}{128}|2{+}5\rangle \\
&\qquad + \tfrac{3}{128}|3{+}0\rangle + \tfrac{15}{128}|3{+}1\rangle + \tfrac{15}{64}|3{+}2\rangle + \tfrac{15}{64}|3{+}3\rangle + \tfrac{15}{128}|3{+}4\rangle + \tfrac{3}{128}|3{+}5\rangle \\
&= \tfrac{1}{128}|2\rangle + \tfrac{1}{16}|3\rangle + \tfrac{25}{128}|4\rangle + \tfrac{5}{16}|5\rangle + \tfrac{35}{128}|6\rangle + \tfrac{1}{8}|7\rangle + \tfrac{3}{128}|8\rangle.
\end{aligned}
$$

A convex sum of two or more distributions can be formed in WebPPL via the pre-defined `Mixture` operation. We illustrate this by re-using the ProbLog distributions from (2.65) with

```
var distproduct1 = function(dist1, dist2) {Infer(function(){
       var s1 = sample(dist1);
       var s2 = sample(dist2);
       return [s1, s2]; }) }

var listproduct = function(list1, list2) {
   var pairlist = map(function(x1) {
       map(function(x2) { return [x1, x2] }, list2) }, list1);
   return pairlist.flat() }

var distproduct2 = function(dist1, dist2) {
    var supp1 = dist1.support();
    var supp2 = dist2.support();
    var prod_supp = listproduct(supp1, supp2);
    var prod_probs = map(function(x1) {
        map(function(x2) { Math.exp(dist1.score(x1) + dist2.score(x2)) },
            supp2) },    supp1);
    return Categorical({ps: prod_probs.flat(), vs: prod_supp}) }
```

Figure 2.9 The parallel tensor product of two distribution in WebPPL, in two forms, via inference above, and via pointwise products below, where, recall, the `score` produces the logarithm of the probability.

their convex sum in (2.66).

```
var supp = ['a','b','c']
var omega = Categorical({ps: [1/2, 1/3, 1/6], vs: supp})
var rho   = Categorical({ps: [1/8, 1/4, 5/8], vs: supp})

console.log( as_kets( Mixture( {dists: [omega, rho],
                                ps: [1/5, 4/5]}) ) )
```

It prints `0.2|a> + 0.267|b> + 0.533|c>`, essentially as in the previous section.

This `Infer` operation can also be used to define the parallel (tensor) product of two distributions, see the first product function `distproduct1` in Figure 2.9. This definition works quite generally, for instance also for continuous distributions. In the discrete case, when supports are available, as lists, one can define a product distribution basically as in Definition 2.3.1. This is implemented the second product function `distproduct2` in Figure 2.9.

When we re-use the above two distributions `omega` and `rho` and take their product, as in:

```
var pp1 = distproduct1(omega, rho)
var pp2 = distproduct2(omega, rho)

console.log( as_kets(pp1) )
console.log( as_kets(pp2) )
viz( pp1 )
```

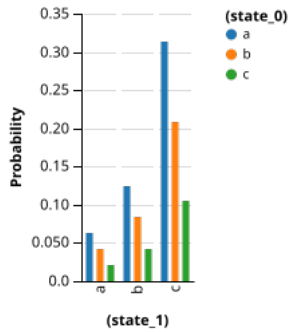
```
var push1 = function(chan, dist) { Infer(function () {
    var x = sample(dist);
    var y = sample(chan(x));
    return y; }) }

var push2 = function(chan, dist) {
    var supp = dist.support();
    var dist_probs = map(function(s){Math.exp(dist.score(s))}, supp);
    return Mixture({ dists: map(chan, supp), ps : dist_probs }) }
```

Figure 2.10 The pushforward of a distribution `dist` along a channel `chan`, in two forms, first via `Infer`, and next via a mixture of the distribution of the channel.

The parallel products `pp1` and `pp2` both produce the distribution on the left below, and (hence) also the same WebPPL visualisation on the right.

```
0.063|a,a> + 0.125|a,b> + 0.313|a,c>
+ 0.042|b,a> + 0.083|b,b> + 0.208|b,c>
+ 0.021|c,a> + 0.042|c,b> + 0.104|c,c>
```



Notice that the picture is two-dimensional in contrast to the three-dimensional bar chart that we used in (2.25) for a joint distribution.

Next we look at pushforward of a distribution along a channel. Distributions are formalised in WebPPL, as we have seen, but channels are not. One could extend the WebPPL language with a Javascript class for channels, but here we take a pragmatic approach and represent a channel as a *ad hoc* function sending elements of a (support) list to distributions, see Figure 2.10. It contains two ways to implement pushforward in WebPPL, first via `Infer` and next via `Mixture`. This last approach corresponds to the convex sum description that we use in (2.34). The first implementation using sampling and `Infer` corresponds to the description that we have seen in (2.35).

We illustrate more concretely how this works for the channels in Exercise 2.4.1 — also used

in the ProbLog example 2.69.

```
var w = Categorical({ps: [1/6, 1/2, 1/3], vs: ['x','y','z']})
var c = function(input) {
    var probs = (input == 'x') ? [1/2, 1/2] :
                (input == 'y') ? [1,   0  ] : [3/4, 1/4];
    return Categorical({ps: probs, vs: ['u', 'v']}) }
var d = function(input) {
    var probs = (input == 'u') ? [1/4, 1/8, 1/2, 1/8] :
                                 [1/4, 0,   1/8, 5/8];
    return Categorical({ps: probs, vs: [1,2,3,4]}) }

var v1 = push1(c, w)
var v2 = push2(c, w)
console.log( as_kets(v1) )
console.log( as_kets(v2) )
console.log( as_kets(push1(d, v1)) )
console.log( as_kets(push2(d, v2)) )
```

The printing of the variables `v1` and `v2` in the first two `console` lines gives the distribution `0.833|u> + 0.167|v>`, which we write as $c_*(\omega)$. Similarly, the next two outputs both yield `0.25|1> + 0.104|2> + 0.437|3> + 0.208|4>`, which is $d_*(c_*(\omega))$.

Remark 2.9.1 The `Infer{ - }` operation is the inference workhorse of the language WebPPL. It executes the program block inside its braces `{ - }` multiple times, collects the outputs given in the `return` statement, and turns these outputs into a distribution object — via frequentist learning. The `Infer` function can take be supplied with different options, like 'enumerate' or 'rejection' to select the form of inference, see the literature Wingate et al. (2011); Goodman and Stuhlmüller (2014); Goodman and Tenenbaum (2016) for further details[5]. It may include conditioning, as will be illustrated later on in Section 6.6. This inference is also described as marginalisation, since the resulting distribution does not cover all variables in the enclosed code block but only the ones in the return statement. Concretely, in the `push1` function definition Figure 2.10 the variable `x` is marginalised out.

In Figures 2.9 and 2.10 we have seen two different implementations of the same constructions, namely parallel product and pushforward. One of these versions works via sampling and `Infer`, and the other versions via direct calculations — in the style of this book — for the discrete case. The latter approach seems to be more efficient (and also more accurate) in larger programs since it avoids multiple, possibly nested inference runs. Although the language WebPPL does allow nested inference, see *e.g.* Figure 7.6 later on, it seems that the common programming pattern in WebPPL is to have one single `Infer` around the entire program (model).

Nested inference is discussed explicitly in Zhang and Amin (2022) and in (Goodman and Tenenbaum, 2016, §6) as part of a theory of mind.

We conclude this introductory section on WebPPL with the wetness Bayesian network from Subsection 2.5.2, see Figure 2.8 for its ProbLog version. The WebPPL version is in Figure 2.11. It makes use of the `flip` expression, which is a draw from the Bernoulli distribution; it is useful for representing fuzzy boolean predicates, see Chapter 4. Thus, `flip` in WebPPL is not a distribution,

[5] Or the WebPPL documentation at https://webppl.readthedocs.io/en/dev/inference/index.html.

```
var wetness = function() {
    var winter        = flip(3/5);
    var sprinkler     = winter ? flip(1/5) : flip(3/4);
    var rain          = winter ? flip(4/5) : flip(1/10);
    var wet_grass     = sprinkler  & rain  ? flip(19/20) :
                        sprinkler  & !rain ? flip(9/10)  :
                        !sprinkler & rain  ? flip(4/5)
: flip(0);
    var slippery_road = rain ? flip(7/10) : flip(0);
    return { Winter : winter, Sprinkler : sprinkler, Rain : rain,
             Wet_grass : wet_grass,  Slippery_road : slippery_road };
}
```

Figure 2.11 WebPPL version of the wetness Bayesian network from Subsection 2.5.2.

as used here of the form $flip(r) = r|1\rangle + (1-r)|0\rangle$. The latter is represented in WebPPL as the distribution `Bernoulli({p: r})`.

The output / return in Figure 2.11 is given a by a record object. This makes it easier to compute the various distributions, namely as:

```
console.log( as_kets(Infer(function() { wetness().Winter })) )
console.log( as_kets(Infer(function() { wetness().Sprinkler })) )
console.log( as_kets(Infer(function() { wetness().Rain })) )
console.log( as_kets(Infer(function() { wetness().Wet_grass })) )
console.log( as_kets(Infer(function() { wetness().Slippery_road })) )
```

The output is, consecutively:

```
0.4|false> + 0.6|true>   0.58|false> + 0.42|true>   0.48|false> + 0.52|true>
         0.301|false> + 0.699|true>     0.636|false> + 0.364|true>
```

Exercises

2.9.1 In Figure 2.9 we have seen the tensor product defined in two ways. Define marginalisation yourself, also in two such ways.

2.10 Divergence between distributions

In many situations it is useful to know how unequal / different / apart probability distributions are. This can be used for instance in learning, where one can try to bring a distribution closer to a target via iterative adaptations. Such comparison of distributions can be defined via a metric / distance function, like 'total variation', see Section 4.8 later on. At this stage, we describe a different comparison, called *divergence*, or more fully *Kullback-Leibler divergence*, written as

D_{KL}. It is a not a proper distance function since it is not symmetric: $D_{KL}(\omega,\rho) \neq D_{KL}(\rho,\omega)$, in general. But it does satisfy $D_{KL}(\omega,\rho) = 0$ iff $\omega = \rho$, and other useful properties.

In the sequel we shall make frequent use of the Kullback-Leibler divergence D_{KL}. This section collects the definition and some basic (standard) facts, see Kullback (1959); Cover and Thomas (2006) for more information. It assumes rudimentary familiarity with the natural logarithm $\ln = \log_e$, with Euler's number $e = 2.718\ldots$ as base. Crucial properties are that the logarithm sends multiplications to sums and that it preserves and reflects the order, see Exercise 1.4.2.

One can also use different bases than e, via logarithm functions $\log_b \colon \mathbb{R}_{>0} \to \mathbb{R}$ with basis b. The crucial property is $\log_b(x) = y$ iff $x = b^y$. These logarithms with different bases are related via a constant, as $\log_b(x) = \frac{\ln(x)}{\ln(b)}$.

Definition 2.10.1 Let ω,ρ be two distributions/states on the same set X with $supp(\omega) \subseteq supp(\rho)$. The *Kullback-Leibler divergence*, or *KL-divergence*, or simply *divergence*, of ω from ρ is:

$$D_{KL}(\omega,\rho) \coloneqq \sum_{x \in supp(\omega)} \omega(x) \cdot \ln\left(\frac{\omega(x)}{\rho(x)}\right). \tag{2.72}$$

With the convention that $r \cdot \ln(r) = 0$ when $r = 0$, we can also take the sum over all $x \in X$.

When the logarithm $\log_2$ is used the KL-divergence number is expressed in terms of 'bits'. When the natural logarithm ln is used, the term 'nats' is sometimes used, see *e.g.* Murphy (2023).

The inclusion $supp(\omega) \subseteq supp(\rho)$ is equivalent to: $\rho(x) = 0$ implies $\omega(x) = 0$. This requirement immediately implies that divergence is not symmetric. But even when ω and ρ do have the same support, the divergences $D_{KL}(\omega,\rho)$ and $D_{KL}(\rho,\omega)$ are different, in general, see Exercise 2.10.1 below for an easy illustration. To emphasise the difference, some people write $D_{KL}(\rho \,\|\, \omega)$ instead of $D_{KL}(\rho,\omega)$.

Whenever we write an expression $D_{KL}(\omega,\rho)$ we will implicitly assume an inclusion $supp(\omega) \subseteq supp(\rho)$.

Divergence behaves well with respect to tensor products and graph distributions.

Proposition 2.10.2 *Let X, Y be arbitrary sets.*

1 *Divergence of tensors is a sum of divergences: for distributions $\omega,\rho \in \mathcal{D}(X)$ and $\omega',\rho' \in \mathcal{D}(Y)$,*

$$D_{KL}(\omega \otimes \omega',\, \rho \otimes \rho') \;=\; D_{KL}(\omega,\rho) + D_{KL}(\omega',\rho').$$

2 *Let $\omega,\rho \in \mathcal{D}(X)$ be two distributions, and let $c,d \colon X \rightsquigarrow Y$ be two channels. Then:*

$$D_{KL}\Big(\langle id, d\rangle_*(\omega),\, \langle id, c\rangle_*(\rho)\Big) \;=\; D_{KL}\Big(\langle id, d\rangle_*(\omega),\, \langle id, c\rangle_*(\omega)\Big) \;+\; D_{KL}(\omega,\rho).$$

As a special case: $D_{KL}\Big(\langle id, c\rangle_*(\omega),\, \langle id, c\rangle_*(\rho)\Big) \;=\; D_{KL}(\omega,\rho)$.

Proof. 1 By unwrapping the relevant definitions and using that the logarithm ln sends multipli-

cations to sums:

$$\begin{aligned}
D_{KL}(\omega\otimes\omega',\,\rho\otimes\rho') &= \sum_{x\in X,\,y\in Y}(\omega\otimes\omega')(x,y)\cdot\ln\left(\frac{(\omega\otimes\omega')(x,y)}{(\rho\otimes\rho')(x,y)}\right)\\
&= \sum_{x\in X,\,y\in Y}\omega(x)\cdot\omega'(y)\cdot\ln\left(\frac{\omega(x)}{\rho(x)}\cdot\frac{\omega'(y)}{\rho'(y)}\right)\\
&= \sum_{x\in X,\,y\in Y}\omega(x)\cdot\omega'(y)\cdot\left(\ln\left(\frac{\omega(x)}{\rho(x)}\right)+\ln\left(\frac{\omega'(y)}{\rho'(y)}\right)\right)\\
&= \sum_{x\in X,\,y\in Y}\omega(x)\cdot\omega'(y)\cdot\ln\left(\frac{\omega(x)}{\rho(x)}\right)\;+\;\sum_{x\in X,\,y\in Y}\omega(x)\cdot\omega'(y)\cdot\ln\left(\frac{\omega'(y)}{\rho'(y)}\right)\\
&= \sum_{x\in X}\omega(x)\cdot\ln\left(\frac{\omega(x)}{\rho(x)}\right)\;+\;\sum_{y\in Y}\omega'(y)\cdot\ln\left(\frac{\omega'(y)}{\rho'(y)}\right)\\
&= D_{KL}(\omega,\rho)+D_{KL}(\omega',\rho').
\end{aligned}$$

2 We compute the subtraction:

$$\begin{aligned}
&D_{KL}\Big(\langle id,d\rangle_*(\omega),\,\langle id,c\rangle_*(\rho)\Big)\;-\;D_{KL}(\omega,\rho)\\
&= \sum_{x\in X,\,y\in Y}\omega(x)\cdot d(x)(y)\cdot\ln\left(\frac{\omega(x)\cdot d(x)(y)}{\rho(x)\cdot c(x)(y)}\right)\;-\;\sum_{x\in X}\omega(x)\cdot\ln\left(\frac{\omega(x)}{\rho(x)}\right)\\
&= \sum_{x\in X,\,y\in Y}\omega(x)\cdot d(x)(y)\cdot\ln\left(\frac{\omega(x)\cdot d(x)(y)}{\rho(x)\cdot c(x)(y)}\right)\\
&\qquad -\sum_{x\in X,\,y\in Y}\omega(x)\cdot d(x)(y)\cdot\ln\left(\frac{\omega(x)}{\rho(x)}\right)\\
&= \sum_{x\in X,\,y\in Y}\omega(x)\cdot d(x)(y)\cdot\ln\left(\frac{\omega(x)\cdot d(x)(y)}{\rho(x)\cdot c(x)(y)}\cdot\frac{\rho(x)}{\omega(x)}\right)\\
&= \sum_{x\in X,\,y\in Y}\omega(x)\cdot d(x)(y)\cdot\ln\left(\frac{d(x)(y)}{c(x)(y)}\right)\\
&= \sum_{x\in X,\,y\in Y}\omega(x)\cdot d(x)(y)\cdot\ln\left(\frac{\omega(x)\cdot d(x)(y)}{\omega(x)\cdot c(x)(y)}\right)\\
&= D_{KL}\Big(\langle id,d\rangle_*(\omega),\,\langle id,c\rangle_*(\omega)\Big).
\end{aligned}$$

❑

In order to prove further properties about divergence we need a powerful classical result called *Jensen's inequality* about functions acting on convex combinations of non-negative reals. We shall use Jensen's inequality here, in this section, but also later on at various stages. We start with a standard result about concave functions.

Lemma 2.10.3 *Let $f\colon\mathbb{R}_{>0}\to\mathbb{R}$ be a function whose second derivative is negative: $f''<0$. Then f is stricly concave, as expressed by the first two points below.*

1 For $a,b\in\mathbb{R}_{>0}$ and $r\in[0,1]$,

$$f(r\cdot a+(1-r)\cdot b)\;\geq\;r\cdot f(a)+(1-r)\cdot f(b).$$

2 Strictness means that if there is an equality $f(r\cdot a+(1-r)\cdot b)=r\cdot f(a)+(1-r)\cdot f(b)$, then either $r=0$ or $r=1$ or $a=b$.

3 *The inequality in item (1) extends to arbitrary (finite) convex sums: for numbers* $a_1, \dots, a_n \in \mathbb{R}_{>0}$ *and* $r_1, \dots, r_n \in [0,1]$ *with* $\sum_i r_i = 1$*, one has:*

$$f(\textstyle\sum_i r_i \cdot a_i) \geq \sum_i r_i \cdot f(a_i). \tag{2.73}$$

These results hold in particular for logarithms, when $f = \ln$*, or* $f = \log_b$ *for a basis* b*. We shall refer to (2.73) as Jensen's inequality, especially when* f *is such a logarithm.*

The proof is standard but is included for convenience.

Proof. 1 Let $a, b \in \mathbb{R}_{>0}$ be given, with $r \in [0,1]$. We write $c := ra + (1-r)b = b - r(b-a)$, so that $a \leq c \leq b$. We need to prove $f(ra + (1-r)b) \geq rf(a) + (1-r)f(b)$. The result is trivial if $r = 0$ or $r = 1$ or $a = b$. So let, without loss of generality, $a < b$ and $r \in (0,1)$, so that $a < c < b$. By the mean value theorem we can find $a < u < c$ and $c < v < b$ with:

$$\frac{f(c) - f(a)}{c - a} = f'(u) \qquad \text{and} \qquad \frac{f(b) - f(c)}{b - c} = f'(v)$$

Since $f'' < 0$ we have that f' is strictly decreasing, so $f'(u) > f'(v)$ because $u > v$. We can write:

$$c - a = (r-1)a + (1-r)b = (1-r)(b-a) \quad \text{and} \quad b - c = r(b-a).$$

From $f'(u) > f'(v)$ we deduce inequalities:

$$\frac{f(c) - f(a)}{(1-r)(b-a)} > \frac{f(b) - f(c)}{r(b-a)} \qquad \textit{i.e.} \qquad r(f(c) - f(a)) > (1-r)(f(b) - f(c).$$

By reorganising the latter inequality we get $f(c) > rf(a) + (1-r)f(b)$, as required.

2 In case we do have an equality $f(c) = rf(a) + (1-r)f(b)$, in the proof of the previous point, we have to give up one of our assumptions $a < b$ or $r \in (0,1)$.

3 We use induction on $n \geq 1$. If $n = 1$ the result is trivial. The case $n = 2$ is given by Lemma 2.10.3 (1). Next, let, for $n \geq 2$, numbers $a_1, \dots, a_n, a_{n+1} \in \mathbb{R}_{>0}$ and $r_1, \dots, r_n, r_{n+1} \in [0,1]$ be given with $\sum_{i \leq n+1} r_i = 1$. If $r_{n+1} = 1$, then $r_1 = \cdots = r_n = 0$ and the result trivialises. Hence we may assume $r_{n+1} \neq 0$.

For $i \leq n$ we write $r'_i := \frac{r_i}{1 - r_{n+1}}$ so that $\sum_{i \leq n} r'_i = 1$. Then:

$$\begin{aligned} f\left(\sum_{i \leq n+1} r_i \cdot a_i\right) &= f\left((1 - r_{n+1}) \cdot \left(\sum_{i \leq n} r'_i \cdot a_i\right) + r_{n+1} \cdot a_{n+1}\right) \\ &\geq (1 - r_{n+1}) \cdot f\left(\sum_{i \leq n} r'_i \cdot a_i\right) + r_{n+1} \cdot f(a_{n+1}) \\ &\geq (1 - r_{n+1}) \cdot \left(\sum_{i \leq n} r'_i \cdot f(a_i)\right) + r_{n+1} \cdot f(a_{n+1}) = \sum_{i \leq n+1} r_i \cdot f(a_i). \end{aligned}$$

The first inequality ($\geq$) follows from Lemma 2.10.3 (1) and the second one by the induction hypothesis. ❑

A strict version of this result, for arbitrary convex combinations, is rather subtle. We formulate — and need — it only for the natural logarithm.

Lemma 2.10.4 *Let numbers $a_1, \ldots, a_n \in \mathbb{R}_{>0}$ and $r_1, \ldots, r_n \in (0,1)$, so $r_i \neq 0$ and $r_i \neq 1$, for all i, be given with $\sum_i r_i = 1$ and with $\sum_i r_i \cdot a_i \leq 1$. Then strictness holds of the following form.*

$$\textstyle\sum_i r_i \cdot \ln(a_i) = 0 \implies a_1 = \cdots = a_n = 1.$$

Proof. We use induction on $n \geq 2$, starting with the case $n = 2$. When $a_1 \neq a_2$ we have, by Lemma 2.10.3 (2), a strict inequality $<$ in:

$$0 = r_1 \cdot \ln(a_1) + r_2 \cdot \ln(a_2) < \ln(r_1 \cdot a_1 + r_2 \cdot a_2) \leq \ln(1) = 0.$$

This is impossible, and thus $a_1 = a_2$. Then:

$$0 = r_1 \cdot \ln(a_1) + r_2 \cdot \ln(a_2) = (r_1 + r_2) \cdot \ln(a_1) = 1 \cdot \ln(a_1) = \ln(a_1).$$

This yields $a_1 = a_2 = 1$.

For the induction step we use a result known as the inequality of arithmetic and geometric means: for $s_i, b_i \in \mathbb{R}_{>0}$ with $\sum_i s_i = 1$,

$$\textstyle\sum_i s_i \cdot b_i \geq \prod_i b_i^{s_i}. \tag{2.74}$$

The proof uses that ln is concave:

$$\ln\big(\textstyle\sum_i s_i \cdot b_i\big) \geq \sum_i s_i \cdot \ln(b_i) = \sum_i \ln(b_i^{s_i}) = \ln\big(\prod_i b_i^{s_i}\big).$$

Hence by taking an e-power on both sides we are done.

We proceed with the induction step, for $n \geq 2$, and assume $\sum_{i \leq n+1} r_i \cdot \ln(a_i) = 0$, where $\sum_{i \leq n+1} r_i \cdot a_i \leq 1$. Since $r_{n+1} \neq 1$, for $i \leq n$ we can write $r'_i = \frac{r_i}{1 - r_{n+1}}$, so that $\sum_{i \leq n} r'_i = 1$. Then:

$$\begin{aligned} 0 = \sum_{i \leq n+1} r_i \cdot \ln(a_i) &= (1 - r_{n+1}) \cdot \sum_{1 \leq i \leq n} r'_i \cdot \ln(a_i) + r_{n+1} \cdot \ln(a_{n+1}) \\ &= (1 - r_{n+1}) \cdot \ln\left(\prod_{i \leq n} a_i^{r'_i}\right) + r_{n+1} \cdot \ln(a_{n+1}). \end{aligned}$$

We also have:

$$\begin{aligned} (1 - r_{n+1}) \cdot \left(\prod_{i \leq n} a_i^{r'_i}\right) + r_{n+1} \cdot a_{n+1} &\leq (1 - r_{n+1}) \cdot \left(\sum_{i \leq n} r'_i \cdot a_i\right) + r_{n+1} \cdot a_{n+1} \qquad \text{by (2.74)} \\ &= \sum_{i \leq n+1} r_i \cdot a_i \leq 1. \end{aligned}$$

We can now apply the implication in the theorem in binary form — which we already proved as base case. It yields $\prod_i a_i^{r'_i} = a_{n+1} = 1$. Then:

$$0 = \ln(1) = \ln\left(\prod_{i \leq n} a_i^{r'_i}\right) = \sum_{i \leq n} r'_i \cdot \ln(a_i).$$

In order to apply the induction hypothesis, we still need to prove $\sum_{i \leq n} r'_i \cdot a_i \leq 1$. This follows from the argument below that uses the assumption $\sum_{i \leq n+1} r_i \cdot a_i \leq 1$ together with the equation $a_{n+1} = 1$ that we just obtained.

$$\sum_{i \leq n} r'_i \cdot a_i = \tfrac{1}{1 - r_{n+1}} \cdot \left(\sum_{i \leq n} r_i \cdot a_i\right) \leq \tfrac{1}{1 - r_{n+1}} \cdot (1 - r_{n+1} \cdot a_{n+1}) = \tfrac{1}{1 - r_{n+1}} \cdot (1 - r_{n+1}) = 1.$$

The induction hypothesis now gives equalities: $a_1 = \cdots = a_n = 1$. Since we already knew $a_{n+1} = 1$, we are done. ❑

Remark 2.10.5 In the future we often use Jensen's inequality, of the form $\ln(\sum_i r_i \cdot a_i) \geq \sum_i r_i \cdot \ln(a_i)$, where finitely many $r_i \in [0,1]$ satisfy $\sum_i r_i = 1$. There are some other inequalities related to e-powers and natural logarithms that we sometimes use, such as:

$$e^x \geq 1 + x \quad \text{with equality} \quad e^x = 1 + x \iff x = 0. \tag{2.75}$$

We shall not prove this property; it is obvious when one looks at the graphs of the two functions e^x and $x + 1$.

From this property (2.75) we derive, by using $\ln(x)$ in place of x.

$$\ln(x) \leq x - 1 \quad \text{with equality} \quad \ln(x) = x - 1 \iff x = 1. \tag{2.76}$$

And by inserting $\frac{1}{x}$ for x in (2.76) we obtain:

$$\ln(x) \geq 1 - \tfrac{1}{x} \quad \text{with equality} \quad \ln(x) = 1 - \tfrac{1}{x} \iff x = 1. \tag{2.77}$$

We can now say a bit more about divergence. For instance, that it is non-negative, as one expects.

Proposition 2.10.6 *Let $\omega, \rho \in \mathcal{D}(X)$ be distributions on the same space X.*

1 $D_{KL}(\omega, \rho) \geq 0$.
2 $D_{KL}(\omega, \rho) = 0$ *implies* $\omega = \rho$.
3 *State transformation is D_{KL}-non-expansive (or short): for a channel $c\colon X \rightsquigarrow Y$ one has:*

$$D_{KL}\big(c_*(\omega),\, c_*(\rho)\big) \leq D_{KL}(\omega,\, \rho).$$

4 *In particular, for a function $f\colon X \to Y$,*

$$D_{KL}\big(\mathcal{D}(f)(\omega),\, \mathcal{D}(f)(\rho)\big) \leq D_{KL}(\omega,\, \rho).$$

This inequality $\leq$ is an equality when f is an isomorphism.

The first item shows that Kullback-Leibler divergence is non-negative. This statement is also known as Gibb's inequality, especially when it is written out explicitly, for instance as:

$$\sum_{x \in X} \omega(x) \cdot \ln(\omega(x)) \geq \sum_{x \in X} \omega(x) \cdot \ln(\rho(x)), \qquad \text{for } \omega, \rho \in \mathcal{D}(X). \tag{2.78}$$

Item (1) tells that KL-divergence is a number in $\mathbb{R}_{\geq 0}$; it can become arbitrarily large, see Exercise 2.10.11 below. A more informative lower bound for KL-divergence, known as Pinsker's inequality, occurs in Theorem 4.8.8 (2), in terms of the total variation distance. An alternative formulation of KL-divergence in terms of validity, known as the Donsker-Varadhan duality formula, occurs in Proposition 6.12.2.

Proof. 1 We recall the assumption $supp(\omega) \subseteq supp(\rho)$. The inequality (2.77) gives:

$$\begin{aligned} D_{KL}(\omega, \rho) &= \sum_{x \in supp(\omega)} \omega(x) \cdot \ln\left(\frac{\omega(x)}{\rho(x)}\right) \geq \sum_{x \in supp(\omega)} \omega(x) \cdot \left(1 - \frac{\rho(x)}{\omega(x)}\right) \\ &= \sum_{x \in supp(\omega)} \omega(x) - \rho(x) = 1 - \sum_{x \in supp(\omega)} \rho(x) \geq 0. \end{aligned}$$

2 The direction ($\Leftarrow$) is easy, by the equation $\ln(1) = 0$. For the reverse direction ($\Rightarrow$), let $D_{KL}(\omega,\rho) = 0$. We use the strictness of the natural logarithm, as concave function, see Lemma 2.10.4. We have:

$$0 = -D_{KL}(\omega,\rho) = \sum_{x\in supp(\omega)} \omega(x)\cdot \ln\left(\frac{\rho(x)}{\omega(x)}\right).$$

Lemma 2.10.4 applies since:

$$\sum_{x\in supp(\omega)} \omega(x) = 1 \quad \text{and} \quad \sum_{x\in supp(\omega)} \omega(x)\cdot\frac{\rho(x)}{\omega(x)} = \sum_{x\in supp(\omega)} \rho(x) \leq 1.$$

We may conclude $\frac{\rho(x)}{\omega(x)} = 1$, for each $x \in supp(\omega)$, and thus $\omega(x) = \rho(x)$. In particular:

$$1 = \sum_{x\in supp(\omega)} \omega(x) = \sum_{x\in supp(\omega)} \rho(x).$$

We now write $supp(\rho)$ as disjoint union $supp(\omega) \cup U$ for $U = supp(\rho) \setminus supp(\omega)$. It suffices to show $U = \emptyset$. We have:

$$1 = \sum_{x\in supp(\rho)} \rho(x) = \sum_{x\in supp(\omega)} \rho(x) + \sum_{x\in U} \rho(x) = 1 + \sum_{x\in U} \rho(x).$$

This means that $\rho(x) = 0$ for each $x \in U \subseteq supp(\rho)$, and thus $U = \emptyset$. This allows us to conclude $\omega = \rho$.

3 Via Jensen's inequality (2.73):

$$\begin{aligned} D_{KL}(c_*(\omega), c_*(\rho)) &= \sum_{y\in Y} c_*(\omega)(y)\cdot \ln\left(\frac{c_*(\omega)(y)}{c_*(\rho)(y)}\right) = \sum_{x\in X,\, y\in Y} \omega(x)\cdot c(x)(y)\cdot \ln\left(\frac{c_*(\omega)(y)}{c_*(\rho)(y)}\right) \\ &\leq \sum_{x\in X} \omega(x)\cdot \ln\left(\sum_{y\in Y} c(x)(y)\cdot \frac{c_*(\omega)(y)}{c_*(\rho)(y)}\right). \end{aligned}$$

Hence it suffices to prove an inequality:

$$\sum_{x\in X} \omega(x)\cdot \ln\left(\sum_{y\in Y} c(x)(y)\cdot \frac{c_*(\omega)(y)}{c_*(\rho)(y)}\right) \leq D_{KL}(\omega,\rho) = \sum_{x\in X} \omega(x)\cdot \ln\left(\frac{\omega(x)}{\rho(x)}\right).$$

This inequality $\leq$ follows from another application of Jensen's inequality:

$$\begin{aligned} &\sum_{x\in X} \omega(x)\cdot \ln\left(\sum_{y\in Y} c(x)(y)\cdot \frac{c_*(\omega)(y)}{c_*(\rho)(y)}\right) - \sum_{x\in X} \omega(x)\cdot \ln\left(\frac{\omega(x)}{\rho(x)}\right) \\ &= \sum_{x\in X} \omega(x)\cdot \left[\ln\left(\sum_{y\in Y} c(x)(y)\cdot \frac{c_*(\omega)(y)}{c_*(\rho)(y)}\right) - \ln\left(\frac{\omega(x)}{\rho(x)}\right)\right] \\ &= \sum_{x\in X} \omega(x)\cdot \ln\left(\sum_{y\in Y} c(x)(y)\cdot \frac{c_*(\omega)(y)}{c_*(\rho)(y)}\cdot\frac{\rho(x)}{\omega(x)}\right) \\ &\leq \ln\left(\sum_{x\in X,\, y\in Y} \omega(x)\cdot c(x)(y)\cdot \frac{c_*(\omega)(y)}{c_*(\rho)(y)}\cdot\frac{\rho(x)}{\omega(x)}\right) \\ &= \ln\left(\sum_{y\in Y}\left(\sum_{x\in X} c(x)(y)\cdot\rho(x)\right)\cdot \frac{c_*(\omega)(y)}{c_*(\rho)(y)}\right) = \ln\left(\sum_{y\in Y} c_*(\omega)(y)\right) = \ln(1) = 0. \end{aligned}$$

4 The inequality follows by applying the previous item to the deterministic channel $\langle f \rangle = \mathit{unit} \circ f$ and using $\mathcal{D}(f)(\omega) = \langle f \rangle_*(\omega)$, and similiarly for ρ. When f is an isomorphism we have:

$$\begin{aligned} D_{KL}(\omega, \rho) &= D_{KL}\Big(\mathcal{D}(f^{-1})(\mathcal{D}(f)(\omega)), \mathcal{D}(f^{-1})(\mathcal{D}(f)(\rho))\Big) \\ &\leq D_{KL}\Big(\mathcal{D}(f)(\omega), \mathcal{D}(f)(\rho)\Big). \end{aligned}$$

❑

2.10.1 Divergence, and mutual information and entropy

A joint distribution $\tau \in \mathcal{D}(X \times Y)$ is typically 'entwined', that is, it is different from the product $\tau[1,0] \otimes \tau[0,1]$ of its marginals, see Example 2.3.9 (2). The Kullback-Leibler divergence between the two allows us to assign a number to this entwinedness. This divergence is called the *mutual information* of the joint distribution. It can become arbitrarily large, see Exercise 2.10.12.

There is a related difference for joint distributions. We recall from Fact 2.3.10 that for an arbitrary distribution $\omega \in \mathcal{D}(X)$ the joint distributions $\mathcal{D}(\Delta)(\omega)$ and $\omega \otimes \omega$ in $\mathcal{D}(X \times X)$ are different, in general. The information theoretic notion of (Shannon) entropy can be expressed as their divergence (see *e.g.* Perrone (2024a)). Intuitively, the entropy of a distribution is the avarage uncertainty. It plays a crucial role in information theory Cover and Thomas (2006) and in statistical mechanics Dill and Bromberg (2010), where it captures a tendency of matter toward disorder, as maximal entropy. As we shall see below, point distributions give much certainty and have the lowest possible entropy, namely zero. Uniform distributions, on the other hand, give maximal uncertainty and have the highest entropy. Entropy also exists for (natural) multisets, see Exercise 3.6.6.

Definition 2.10.7 Mutual information MI and entropy H are defined as follows.

1 The mutual information $MI(\tau) \in \mathbb{R}_{\geq 0}$ of a joint distribution $\tau \in \mathcal{D}(X \times Y)$ is its divergence with the product of its two marginals:

$$MI(\tau) := D_{KL}\Big(\tau, \tau[1,0] \otimes \tau[0,1]\Big). \tag{2.79}$$

2 The entropy $H(\omega) \in \mathbb{R}_{\geq 0}$ of a distribution $\omega \in \mathcal{D}(X)$ is defined as:

$$H(\omega) := D_{KL}\Big(\mathcal{D}(\Delta)(\omega), \omega \otimes \omega\Big) = -\sum_{x \in X} \omega(x) \cdot \ln(\omega(x)). \tag{2.80}$$

It is easy to see that the second equation in (2.80) holds:

$$\begin{aligned} D_{KL}\Big(\mathcal{D}(\Delta)(\omega), \omega \otimes \omega\Big) &= \sum_{x,x' \in X} \mathcal{D}(\Delta)(\omega)(x,x') \cdot \ln\left(\frac{\mathcal{D}(\Delta)(\omega)(x,x')}{(\omega \otimes \omega)(x,x')}\right) \\ &= \sum_{x \in X} \omega(x) \cdot \ln\left(\frac{\omega(x)}{\omega(x) \cdot \omega(x)}\right) \\ &= -\sum_{x \in X} \omega(x) \cdot \ln(\omega(x)). \end{aligned}$$

We first establish some basic properties of Shannon entropy.

Lemma 2.10.8 *Let $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ be arbitrary distributions.*

1 Precisely the point distributions have zero entropy: $H(\omega) = 0$ if and only if $\omega = 1|x\rangle$, for a unique element $x \in X$.

2 $H(\omega \otimes \rho) = H(\omega) + H(\rho)$.

3 *For a joint distribution* $\tau \in \mathcal{D}(X \times Y)$*, write* $\tau_1 := \tau[1,0] \in \mathcal{D}(X)$ *and* $\tau_2 := \tau[0,1] \in \mathcal{D}(Y)$ *for its two marginals. Then:*

$$H(\tau) \leq H(\tau_1) + H(\tau_2) \overset{(2)}{=} H(\tau_1 \otimes \tau_2).$$

4 *For a joint distribution* $\tau \in \mathcal{D}(X \times Y)$ *with marginals* τ_i *as in the previous point, mutual information, on the left below, can be expressed in terms of entropy as:*

$$D_{KL}(\tau,\, \tau_1 \otimes \tau_2) = H(\tau_1) + H(\tau_2) - H(\tau) \overset{(2)}{=} H(\tau_1 \otimes \tau_2) - H(\tau). \tag{2.81}$$

5 *For a joint distribution* τ *as before, non-entwinedness of* τ *can also be expressed via entropy, namely as:*

$$\tau = \tau_1 \otimes \tau_2 \iff H(\tau) = H(\tau_1) + H(\tau_2).$$

6 *Let the set* X *now be non-empty and finite, say with* $n = |X| > 0$ *elements. The maximal entropy for distributions on* X *is* $\ln(n)$ *and is reached by the uniform distribution:*

$$\bigvee \{H(\omega) \mid \omega \in \mathcal{D}(X)\} = \ln(n) = H(unif_X).$$

Proof. 1 We use the following line of reasoning:

$$\begin{aligned} H(\omega) = 0 &\overset{(2.80)}{\iff} D_{KL}(\mathcal{D}(\Delta)(\omega),\, \omega \otimes \omega) = 0 \\ &\iff \mathcal{D}(\Delta)(\omega) = \omega \otimes \omega && \text{by Proposition 2.10.6 (2)} \\ &\iff \omega \text{ is a point distribution} && \text{by Exercise 2.3.13.} \end{aligned}$$

2 Since:

$$\begin{aligned} H(\omega \otimes \rho) &= -\sum_{x \in X,\, y \in Y} (\omega \otimes \rho)(x,y) \cdot \ln\Big((\omega \otimes \rho)(x,y)\Big) \\ &= -\sum_{x \in X,\, y \in Y} \omega(x) \cdot \rho(y) \cdot \Big(\ln(\omega(x)) + \ln(\rho(y))\Big) \\ &= -\sum_{x \in X,\, y \in Y} \omega(x) \cdot \rho(y) \cdot \ln(\omega(x)) - \sum_{x \in X,\, y \in Y} \omega(x) \cdot \rho(y) \cdot \ln(\rho(y)) \\ &= -\sum_{x \in X} \omega(x) \cdot \ln(\omega(x)) - \sum_{y \in Y} \rho(y) \cdot \ln(\rho(y)) = H(\omega) + H(\rho). \end{aligned}$$

3 We use Jensen's inequality to show that the difference non-positive:

$$\begin{aligned} H(\tau) - H(\tau_1) - H(\tau_2) &= -\sum_{x \in X,\, y \in Y} \tau(x,y) \cdot \ln(\tau(x,y)) \\ &\quad + \sum_{x \in X} \left(\sum_{y \in Y} \tau(x,y)\right) \cdot \ln(\tau_1(x)) + \sum_{y \in Y} \left(\sum_{x \in X} \tau(x,y)\right) \cdot \ln(\tau_2(y)) \\ &= \sum_{x \in X,\, y \in Y} \tau(x,y) \cdot \Big[\ln(\tau_1(x)) + \ln(\tau_2(y)) - \ln(\tau(x,y))\Big] \\ &= \sum_{x \in X,\, y \in Y} \tau(x,y) \cdot \ln\left(\frac{\tau_1(x) \cdot \tau_2(y)}{\tau(x,y)}\right) \\ &\leq \ln\left(\sum_{x \in X,\, y \in Y} \tau(x,y) \cdot \frac{\tau_1(x) \cdot \tau_2(y)}{\tau(x,y)}\right) = \ln(1) = 0. \end{aligned}$$

4 We only have to prove the first equation in (2.81) since the second one follows directly from item (2).

$$\begin{aligned}
D_{KL}(\tau,\, \tau_1 \otimes \tau_2) &= \sum_{x\in X,\, y\in Y} \tau(x,y) \cdot \ln\left(\frac{\tau(x,y)}{(\tau_1 \otimes \tau_2)(x,y)}\right) \\
&= \sum_{x\in X,\, y\in Y} \tau(x,y) \cdot \ln\big(\tau(x,y)\big) - \sum_{x\in X,\, y\in Y} \tau(x,y) \cdot \ln\big(\tau_1(x) \cdot \tau_2(y)\big) \\
&= -H(\tau) - \sum_{x\in X,\, y\in Y} \tau(x,y) \cdot \ln\big(\tau_1(x)\big) - \sum_{x\in X,\, y\in Y} \tau(x,y) \cdot \ln\big(\tau_2(y)\big) \\
&= -H(\tau) - \sum_{x\in X}\left(\sum_{y\in Y} \tau(x,y)\right) \cdot \ln\big(\tau_1(x)\big) - \sum_{y\in Y}\left(\sum_{x\in X} \tau(x,y)\right) \cdot \ln\big(\tau_2(y)\big) \\
&= -H(\tau) - \sum_{x\in X} \tau_1(x) \cdot \ln\big(\tau_1(x)\big) - \sum_{y\in Y} \tau_2(y) \cdot \ln\big(\tau_2(y)\big) \\
&= -H(\tau) + H(\tau_1) + H(\tau_2).
\end{aligned}$$

5 The direction ($\Rightarrow$) follows from item (2). For ($\Leftarrow$) we see by item (4) that the divergence $D_{KL}(\tau,\, \tau_1 \otimes \tau_2)$ is zero. Hence $\tau = \tau_1 \otimes \tau_2$ by Proposition 2.10.6 (2).

6 First, we compute the entropy of the uniform distribution:

$$H(unif_X) \;=\; -\sum_{x\in X} unif_X(x) \cdot \ln\big(unif_X(x)\big) \;=\; -\sum_{x\in X} \tfrac{1}{n} \cdot \ln\big(\tfrac{1}{n}\big) \;=\; \sum_{x\in X} \tfrac{1}{n} \cdot \ln(n) \;=\; \ln(n).$$

Next, we show that it is maximal. For arbitrary $\omega \in \mathcal{D}(X)$,

$$\begin{aligned}
0 \;\le\; & D_{KL}(\omega, unif_X) && \text{by Proposition 2.10.6 (1)} \\
&= \sum_{x\in X} \omega(x) \cdot \ln\left(\frac{\omega(x)}{1/n}\right) \\
&= \sum_{x\in X} \omega(x) \cdot \ln\big(\omega(x)\big) - \sum_{x\in X} \omega(x) \cdot \ln\big(1/n\big) \\
&= \sum_{x\in X} \omega(x) \cdot \ln\big(\omega(x)\big) \;+\; \ln(n) \;=\; H(unif_X) - H(\omega).
\end{aligned}$$

❑

For mutual information we observe the following fact.

Proposition 2.10.9 *Let $\tau \in \mathcal{D}(X \times Y)$ be joint distribution with marginals $\tau_1 = \tau[1,0] \in \mathcal{D}(X)$ and $\tau_2 = \tau[0,1] \in \mathcal{D}(Y)$.*

1 For arbitrary distributions $\omega \in \mathcal{D}(X)$, $\rho \in \mathcal{D}(Y)$

$$\begin{aligned}
D_{KL}(\tau,\, \omega \otimes \rho) \;&=\; MI(\tau) \;+\; D_{KL}(\tau_1,\, \omega) \;+\; D_{KL}(\tau_2,\, \rho) \\
&=\; D_{KL}(\tau,\, \tau_1 \otimes \tau_2) + D_{KL}(\tau_1 \otimes \tau_2,\, \omega \otimes \rho).
\end{aligned}$$

This last equation follows from Proposition 2.10.2 (1), and the definition of mutual information.

2 As a result, for each non-entwined distribution $\sigma \in \mathcal{D}(X \times Y)$,

$$D_{KL}(\tau, \tau_1 \otimes \tau_2) \;\le\; D_{KL}(\tau, \sigma).$$

This means that the divergence between τ and the (non-entwined) product of its marginals is least among the divergences between τ and any non-entwined distributions.

Proof. 1 We start from the difference:

$$\begin{aligned}
&D_{KL}(\tau,\, \omega\otimes\rho) - D_{KL}(\tau_1,\, \omega) - D_{KL}(\tau_2,\, \rho)\\
&= \sum_{x\in X,\, y\in Y} \tau(x,y)\cdot \ln\left(\frac{\tau(x,y)}{\omega(x)\cdot\rho(y)}\right)\\
&\qquad - \sum_{x\in X} \tau_1(x)\cdot\ln\left(\frac{\tau_1(x)}{\omega(x)}\right) - \sum_{y\in Y}\tau_2(y)\cdot\ln\left(\frac{\tau_2(y)}{\rho(y)}\right)\\
&= \sum_{x\in X,\, y\in Y} \tau(x,y)\cdot\left[\ln\left(\frac{\tau(x,y)}{\omega(x)\cdot\rho(y)}\right) - \ln\left(\frac{\tau_1(x)}{\omega(x)}\right) - \ln\left(\frac{\tau_2(y)}{\rho(y)}\right)\right]\\
&= \sum_{x\in X,\, y\in Y} \tau(x,y)\cdot\ln\left(\frac{\tau(x,y)}{\tau_1(x)\cdot\tau_2(y)}\right) \;=\; MI(\tau).
\end{aligned}$$

2 Since σ is non-entwined, it is of the form $\sigma = \sigma_1\otimes\sigma_2$ for its marginals $\sigma_1 = \sigma[1,0]$ and $\sigma_2 = \sigma[0,1]$. Then, by the previous item:

$$\begin{aligned}
&D_{KL}(\tau,\sigma) - D_{KL}(\tau,\tau_1\otimes\tau_2)\\
&= D_{KL}(\tau,\, \tau_1\otimes\tau_2) + D_{KL}(\tau_1,\, \sigma_1) + D_{KL}(\tau_2,\, \sigma_2) - D_{KL}(\tau,\, \tau_1\otimes\tau_2)\\
&= D_{KL}(\tau_1,\, \sigma_1) + D_{KL}(\tau_2,\, \sigma_2) \;\geq\; 0.
\end{aligned}$$

❑

Exercises

2.10.1 Take $\omega = \frac{1}{4}|a\rangle + \frac{3}{4}|b\rangle$ and $\rho = \frac{1}{2}|a\rangle + \frac{1}{2}|b\rangle$. Check that:

1 $D_{KL}(\omega,\rho) = \frac{3}{4}\cdot\ln(3) - \cdot\ln(2) \approx 0.13$.
2 $D_{KL}(\rho,\omega) = \ln(2) - \frac{1}{2}\cdot\ln(3). \approx 0.14$.

2.10.2 Check that for $y\in supp(\rho)$ one has:

$$D_{KL}(1|y\rangle,\rho) \;=\; -\ln(\rho(y)).$$

The swapped version $D_{KL}(\rho, 1|y\rangle)$ is defined only when $supp(\rho)\subseteq supp(1|y\rangle) = \{y\}$, and thus when $\rho = 1|y\rangle$. Its outcome is then 0.

2.10.3 Let $\omega,\rho\in\mathcal{D}(X)$ be distributions on the domain of an injective function $f\colon X\to Y$. Show that $D_{KL}(\mathcal{D}(f)(\omega),\mathcal{D}(f)(\rho)) = D_{KL}(\omega,\rho)$.

2.10.4 Prove Gibb's inequality (2.78) via the inequaltiy $\ln(a)\leq a-1$, for $a\in\mathbb{R}_{>0}$.

2.10.5 Give a proof of the familiar inequality:

$$\tfrac{1}{2}(x+y) \;\geq\; (x\cdot y)^{1/2}$$

for $x,y\in\mathbb{R}_{\geq 0}$ via Jensen's inequality.

Hint: Use that the logarithm function is a homomorphism of monoids that reflects the order, see Exercise 1.4.2.

2.10.6 Let M be a commutative monoid, with distributions $\omega,\omega',\rho,\rho'\in\mathcal{D}(M)$. Show that their convolution sums satisfy:

$$D_{KL}(\omega+\rho,\, \omega'+\rho') \;\leq\; D_{KL}(\omega,\omega') + D_{KL}(\rho,\rho').$$

2.10.7 Let $\omega_1, \ldots, \omega_n, \rho \in \mathcal{D}(X)$ be given with probabilities $r_1, \ldots, r_n \in [0,1]$ satisfying $\sum_i r_i = 1$. Write $\omega = \sum_i r_i \cdot \omega_i$ and show that this convex sum of distributions satisfies:

$$\textstyle\sum_i r_i \cdot D_{KL}(\omega_i, \rho) \;=\; D_{KL}(\omega, \rho) \;+\; \sum_i r_i \cdot D_{KL}(\omega_i, \omega)$$

As special case of this equation for $r_i = \frac{1}{n}$ one obtains:

$$\textstyle\sum_i D_{KL}(\omega_i, \rho) \;=\; n \cdot D_{KL}(\omega, \rho) \;+\; \sum_i D_{KL}(\omega_i, \omega).$$

In this form, for $n = 2$, it occurs in (Csiszár, 1975, proof of Thm. 2.1).

2.10.8 Let $\omega, \rho_1, \ldots \rho_n \in \mathcal{D}(X)$ now be given with $r_i \in [0,1]$ satisfying $\sum_i r_i = 1$.

1 Show that the function $D_{KL}(\omega, -)$ yields an inequality on convex sums:

$$D_{KL}(\omega,\, r_1 \cdot \rho_1 + \cdots + r_n \cdot \rho_n) \;\leq\; r_1 \cdot D_{KL}(\omega, \rho_1) + \cdots + r_n \cdot D_{KL}(\omega, \rho_n).$$

Prove the following 'log-sum inequality' from (Cover and Thomas, 2006, Thm. 2.7.1): for finite collections $a_i, b_i \in \mathbb{R}_{>0}$,

$$\Big(\sum\nolimits_i a_i\Big) \cdot \ln\left(\frac{\sum_i a_i}{\sum_i b_i}\right) \;\leq\; \sum\nolimits_i a_i \cdot \ln\left(\frac{a_i}{b_i}\right).$$

Hint: Use Lemma 2.10.3 (3) for $f(x) = -x \cdot \ln(x)$.

2 Use this log sum inequality to show that D_{KL} also yields an inequality on (the same) convex sums in both arguments (as in (Cover and Thomas, 2006, Thm. 2.7.2)):

$$\begin{aligned} &D_{KL}(r_1 \cdot \omega_1 + \cdots + r_n \cdot \omega_n,\, r_1 \cdot \rho_1 + \cdots + r_n \cdot \rho_n) \\ &\leq\; r_1 \cdot D_{KL}(\omega_1, \rho_1) + \cdots + r_n \cdot D_{KL}(\omega_n, \rho_n). \end{aligned}$$

2.10.9 Let X be a finite set with $N := |X| > 0$ elements, and let υ be the uniform distribution on X.

1 Show that for $\omega \in \mathcal{D}(X)$, entropy can be expressed in terms of KL-divergence:

$$H(\omega) \;=\; \ln(N) - D_{KL}(\omega, \upsilon).$$

2 Use this equation, and the previous exercise, to show that entropy is concave:

$$H(r_1 \cdot \omega_1 + \cdots + r_n \cdot \omega_n) \;\geq\; r_1 \cdot H(\omega_1) + \cdots + r_n \cdot H(\omega_n).$$

2.10.10 Recall the entwined distribution $\omega = \frac{1}{8}|u,a\rangle + \frac{1}{4}|u,b\rangle + \frac{3}{8}|v,a\rangle + \frac{1}{4}|v,b\rangle$ from Example 2.3.9 (2). Compute the mutual information and show that it satisfies:

$$MI(\omega) \;=\; D_{KL}(\omega,\, \omega[1,0] \otimes \omega[0,1]) \;\approx\; 0.05.$$

2.10.11 This exercise demonstrates that Kullback-Leibler divergence is unbounded: it can become arbitrarily large. Consider the following two distributions on the set $\{0, 1, \ldots, 2^n - 1\}$ with 2^n elements, for $n \geq 1$.

$$\omega \;=\; \tfrac{1}{2}\big|0\big\rangle + \tfrac{1}{2}\big|1\big\rangle \qquad\qquad \rho_n \;=\; \tfrac{1}{2^n}\big|0\big\rangle + \cdots + \tfrac{1}{2^n}\big|2^n - 1\big\rangle.$$

Check that $D_{KL}(\omega, \rho_n) = (n-1) \cdot \ln(2)$.

2.10.12 In a similar way one can show that mutual information is unbounded (with multiple products). Let $\upsilon = \mathit{flip}(\frac{1}{2}) = \frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle \in \mathcal{D}(\mathbf{2})$ and let $\rho_n \in \mathcal{D}(\mathbf{2}^n)$, for $n \geq 1$, be defined by:

$$\rho_n \;=\; \tfrac{1}{2}\big|\underbrace{0, \ldots, 0}_{n \text{ times}}\big\rangle + \tfrac{1}{2}\big|\underbrace{1, \ldots, 1}_{n \text{ times}}\big\rangle.$$

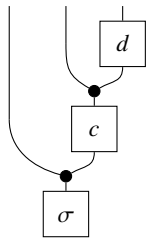
1 Show that each i-th marginal $\mathcal{D}(\pi_i)(\rho_n) = \rho_n[0, \ldots, 0, 1, 0, \ldots, 0]$ is equal to υ.

2 Deduce from Exercise 2.3.4 that:

$$\upsilon^n \;=\; \upsilon \otimes \cdots \otimes \upsilon \;=\; unif_{\mathbf{2}^n} \;=\; \tfrac{1}{2^n}\big|0, \ldots, 0\rangle + \cdots + \tfrac{1}{2^n}\big|1, \ldots, 1\rangle.$$

3 Conclude, like in the previous exercise, that the n-fold mutual information $D_{KL}(\rho_n, \upsilon^n)$ is equal to $(n-1) \cdot \ln(2)$.

2.10.13 Now consider the distribution $\tau \in \mathcal{D}(X \times Y \times Z)$ given by the string diagram:

$\tau \;=$ 

where $\sigma \in \mathcal{D}(X)$, $c \colon X \rightarrowtail Y$, $d \colon Y \rightarrowtail Z$

Prove for the distribution $\tau \in \mathcal{D}(X \times Y \times Z)$,

$$MI\big(\tau[1, 1, 0]\big) \;\geq\; MI\big(\tau[1, 0, 1]\big).$$

This is commonly called the data processing inequality, see *e.g.* (Cover and Thomas, 2006, Thm. 2.8.1). It follows from Proposition 2.10.6 (3).

2.11 Exchangeability for positions and elements

When we have a list of items in $X^K = X \times \cdots \times X$, for a set X and number K, there are two ways to shuffle them: (1) we can keep the elements from X unchanged, but change their positions in the list, and (2) we can change the elements in the list to other elements (also from X). In cryptography, these two methods are called *transposition* and *substitution*. They are typically combined in multiple rounds to form non-trivial encipherment mechanisms. This section collects some basic results about transposition and substitution that set the scene for later. It is shown that transposition and substitution are closely related to multisets and quotients, via the accumulation and matching maps $acc \colon X^K \to \mathcal{M}[K](X)$ and $mat \colon X^K \to SP(K)$ that we saw before. The story has a probabilistic flavour via the probabilistic inverses $acc^{\sim 1}$ and $mat^{\sim 1}$. They turn out to be equalisers in the category of channels.

Definition 2.11.1 Let $\vec{x} = (x_0, \ldots, x_{K-1}) \in X^K$ be a list of length K, with elements from an arbitrary set X.

1 A *transposition* is a bijective endofunction $t \colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}$ on the set of indices $\boldsymbol{K} = \{0, \ldots, K-1\}$. It gives rise to a bijection $\underline{t} := t^K \colon X^K \xrightarrow{\cong} X^K$ that reorders the elements in the list $\vec{x}$ as:

$$\underline{t}(\vec{x}) \;=\; (x_{t(0)}, \ldots, x_{t(K-1)}).$$

The underlining of t in $\underline{t}$ suggests it action on the indices. The result $\underline{t}(\vec{x})$ is also called a transposition (of $\vec{x}$, via t).

2 A *substitution* is a bijective endofunction $s\colon X \xrightarrow{\cong} X$ on the set of elements in the list. It induces a bijection $s^\star\colon X^K \xrightarrow{\cong} X^K$ given by:

$$s^\star(\vec{x}) = \mathcal{L}(s)(\vec{x}) = (s(x_0), \ldots, s(x_{K-1})).$$

The notation $s^\star$ emphasises an analogy with the Kleene star — also used for lists, as $A^\star = \mathcal{L}(A)$. Indeed, this substitution operation $s^\star$ uses the functoriality of the list operation $\mathcal{L}$. The result $s^\star(\vec{x})$ will also be called a substitution.

3 A function $f\colon X^K \to Y$ is called:

- *stable under transposition* when $f(\underline{t}(\vec{x})) = f(\vec{x})$, for all transpositions $t\colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}$ and sequences $\vec{x} \in X^K$; this is equivalent to: $f(\underline{t_1}(\vec{x})) = f(\underline{t_2}(\vec{x}))$, for all transpositions t_1, t_2, since the identity is a transposition itself;
- *stable under substitution* when f satisfies $f(s^\star(\vec{x})) = f(\vec{x})$, for all substitutions $s\colon X \xrightarrow{\cong} X$ and sequences $\vec{x} \in X^K$; this is again equivalent to $f(s_1^\star(\vec{x})) = f(s_2^\star(\vec{x}))$, for all substitutions s_1, s_2.

4 A joint distribution $\omega \in \mathcal{D}(X^K)$ is called *transposition-exhangeable* (resp. *substitution-exchangeable*) if, considered as a function $\omega\colon X^K \to [0,1]$, it is stable under transposition (resp. substitution).

We call a channel $c\colon Y \rightsquigarrow X^K$ transposition- / substitution-exchangeable if each distribution $c(y) \in \mathcal{D}(X^K)$, for $y \in Y$, is transposition- / substitution-exchangeable.

For a finite set X, any uniform distribution on X^K is both transposition and substitution-exchangeble. We shall see more interesting examples later. The accumulation function $acc\colon X^K \to \mathcal{N}[K](X)$ from (1.38) is the archetypical function that is stable under transposition; similarly, the matching function $mat\colon X^K \to SP(X)$ from (1.23) is archetypically stable under substitution. Indeed, transposing the elements does not change the accumulated multiset, which only considers numbers of occurrences, not positions. Matching on the other hand only registers the positions where elements are equal (match), and not what these elements are. Hence the associated set partition does not change if we replace these elements by others — bijectively.

In order to express what 'archetypical' means we briefly move to a categorical perspective and borrow the notion of (co)equaliser.

Definition 2.11.2 Let A be an object in an arbitrary category, with parallel maps $f_1, \ldots, f_n\colon A \to B$, for $n \geq 2$.

1 An *equaliser* of these parallel maps f_i is a map $e\colon E \to A$ in the category satisfying the following two properties.

- $f_i \circ e = f_j \circ e$, for all $i, j \in \{1, \ldots, n\}$;
- if $g\colon D \to A$ satisfies $f_i \circ g = f_j \circ g$ for all i, j, then there is a unique map $h\colon D \to E$ with $e \circ h = g$. This is expressed in the following diagram, where the dashed map expresses uniqueness.

$$\begin{tikzcd} E \arrow[r, "e"] & A \arrow[r, bend left, "f_1"] \arrow[r, bend right, "f_n"'] \arrow[r, phantom, "\vdots"] & B \\ D \arrow[u, dashed, "h"] \arrow[ur, "g"'] & & \end{tikzcd} \tag{2.82}$$

2 Dually, a *coequaliser* of these parallel maps $f_i\colon A \to B$ is a morphism $c\colon B \to C$ with:

- $c \circ f_i = c \circ f_j$ for all i, j;
- if $g\colon B \to D$ satisfies $g \circ f_i = g \circ f_j$ for all i, j, then there is a unique $h\colon C \to D$ with $h \circ c = g$.

$$\begin{tikzcd} A \arrow[r, bend left, "f_1"] \arrow[r, bend right, "f_n"'] \arrow[r, phantom, "\vdots"] & B \arrow[r, "c"] \arrow[dr, "g"'] & C \arrow[d, dashed, "h"] \\ & & D \end{tikzcd} \tag{2.83}$$

The second bullets, both in item (1) and (2), describe what is called a *universal property*. The maps e and c do not only (co)equalise the maps f_i, in the first bullets, but do so in a 'maximal' or 'minimal' manner. The unique map, written as h in both situations, is often called a 'mediating' map.

In the category of sets the above equaliser is an inclusion $E \hookrightarrow A$, where $E = \{a \in A \mid \forall i, j.\ f_i(a) = f_j(a)\}$. The coequaliser $B \to C$ is more difficult to describe and is best understood as the quotient (or collapse) of B that forces all elements $f_i(a) \in B$ to be equal, for each $a \in A$, see Exercise 2.11.2.

Equalisers and coequalisers are usually defined for two parallel maps f_1, f_2 only. One can then obtain a (co)equaliser for multiple maps $f_1, \ldots, f_n$ by iteration. Since our usage typically involves multiple maps, we have introduced the notion of (co)equaliser immediately in n-ary form. Occasinaly, we do use the binary form as well, like in Theorem 2.11.6 below.

The next result give some key examples in the current context. In short, it says that accumulation is the coequaliser of transpositions and that matching is the coequaliser of substitutions. This is the case in the category **Sets** of sets and functions.

Proposition 2.11.3 *Let X be an arbitrary set and let $K \in \mathbb{N}$.*

1 *The accumulation function* $\mathrm{acc}\colon X^K \to \mathcal{N}[K](X)$ *is the coequaliser of all functions* $\underline{t}\colon X^K \to X^K$ *induced by the* $K!$ *transpositions* $t\colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}$, *in:*

$$\begin{tikzcd} X^K \arrow[r, bend left, "\textit{transpositions } \underline{t}"] \arrow[r, bend right] & X^K \arrow[r, "\mathrm{acc}"] & \mathcal{N}[K](X) \end{tikzcd} \tag{2.84}$$

2 *When the set X is finite with at least K elements, the match function* $\mathrm{mat}\colon X^K \to SP(K)$ *is the coequaliser of the functions* $s^\star\colon X^K \to X^K$ *obtained from all the* $|X|!$ *substitutions* $s\colon X \xrightarrow{\cong} X$, *in:*

$$\begin{tikzcd} X^K \arrow[r, bend left, "\textit{substitutions } s^\star"] \arrow[r, bend right] & X^K \arrow[r, "\mathrm{mat}"] & SP(K) \end{tikzcd} \tag{2.85}$$

Proof. 1 If ℓ' is a transposition of $\ell \in X^K$, then $acc(\ell') = acc(\ell)$. Hence acc is stable under transposition. Let the function $g\colon X^K \to Y$ be stable under transposition as well. For a multiset $\varphi = \sum_{1 \le i \le L} n_i | x_i \rangle \in \mathcal{N}[K](X)$, one can choose a list of elements, say $\vec{x} = \langle x_1, \ldots, x_1, \ldots, x_L, \ldots x_L \rangle \in X^K$, where $x_i \in X$ occurs n_i many times. Then $acc(\vec{x}) = \varphi$, by construction. We can now define $h(\varphi) := g(\vec{x})$. Since g is stable under transposition, any list that accumulates to φ gives the same outcome. Clearly, $h \circ acc = g$. Moreover, h is unique with this property.

2 Let ℓ' be a substitution of $\ell \in X^K$. If elements at positions i, j in ℓ are equal, then the elements at positions i, j in ℓ' are also equal. Thus $mat(\ell) = mat(\ell')$, making the match function stable under substitution. Next, let $g\colon X^K \to Y$ be stable under substitution. Given a set partition $P \in SP(K)$, say with $|P| = n \le K$ blocks, we can choose pairwise different elements $x_1, \ldots, x_n \in X$

and put them in such a way in a list $\vec{x} \in X^K$ that $mat(\vec{x}) = P$. Then we define $h(P) := g(\vec{x})$. This definition does not depend on the choice of the list $\vec{x}$, as long as it matches to P, since g is stable under substitution. Thus, $h \circ mat = g$. Clearly, h is the unique function satisfying this equation. ❑

Recall that for accumulation we have identified a probabilistic inverse channel $acc^{\sim 1} = arr$, called arrangement, see (2.38). Also for matching there is a probabilistic inverse $mat^{\sim 1}$, see Exercise 2.4.16. We have not given this inverse a separate name. In general, $f \circledcirc f^{\sim 1} = id$ holds for probabilistic inverses $f^{\sim 1}$, see Definition 2.4.7. As a result, $f^{\sim 1} \circledcirc f$ is a split idempotent with respect to channel composition $\circledcirc$, that is, in the category $\mathbf{Chan} = \mathbf{Chan}(\mathcal{D})$ of probabilistic channels. We make explicit what this idempotent does for accumulation and for matching. This gives alternative descriptions of exchangeability for joint distributions.

Lemma 2.11.4 *Fix a set X and a number K.*

1 *Abbreviate* $transp := acc^{\sim 1} \circledcirc acc \colon X^K \rightsquigarrow X^K$. *This is the transposition idempotent sending a list $\vec{x} \in X^K$ to all its transpositions:*

$$transp(\vec{x}) := \sum_{t \colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}} \frac{1}{K!} \left| \underline{t}(\vec{x}) \right\rangle = \sum_{t \colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}} \frac{1}{K!} \left| x_{t(0)}, \ldots, x_{t(K-1)} \right\rangle. \tag{2.86}$$

2 *Now assume that X is finite; write* $subst := mat^{\sim 1} \circledcirc mat \colon X^K \rightsquigarrow X^K$. *Then:*

$$subst(\vec{x}) := \sum_{s \colon X \xrightarrow{\cong} X} \frac{1}{|X|!} \left| s^{\star}(\vec{x}) \right\rangle = \sum_{s \colon X \xrightarrow{\cong} X} \frac{1}{|X|!} \left| s(x_0), \ldots, s(x_{K-1}) \right\rangle. \tag{2.87}$$

3 *Let $\omega \in \mathcal{D}(X^K)$ be a joint distribution. The following points are equivalent:*

(a) *ω is transposition-exchangeble, see Definition 2.11.1 (4);*
(b) *$\mathcal{D}(\underline{t})(\omega) = \omega$, for all $t \colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}$;*
(c) *$transp_*(\omega) = \omega$.*

4 *Similarly, the next three points are equivalent, for $\omega \in \mathcal{D}(X^K)$, when X is finite:*

(a) *ω is substitution-exchangeble, see Definition 2.11.1 (4);*
(b) *$\mathcal{D}(s^{\star})(\omega) = \omega$, for all $s \colon X \xrightarrow{\cong} X$;*
(c) *$subst_*(\omega) = \omega$.*

Proof. 1 We have:

$$\begin{aligned} transp(\vec{x}) = (acc^{\sim 1} \circledcirc acc)(\vec{x}) &= \sum_{\vec{y} \in acc^{-1}(acc(\vec{x}))} \frac{1}{\big(acc(\vec{x})\big)} \left| \vec{y} \right\rangle = \sum_{\vec{y} \text{ is a permutation of } \vec{x}} \frac{1}{K!} \left| \vec{y} \right\rangle \\ &= \sum_{t \colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}} \frac{1}{K!} \left| x_{t(0)}, \ldots, x_{t(K-1)} \right\rangle. \end{aligned}$$

2 Similarly.

3 The equation $\mathcal{D}(\underline{t})(\omega)(\vec{x}) = \omega(\underline{t^{-1}}(\vec{x}))$ gives the equivalence (a) $\Leftrightarrow$ (b) in item (3). Further, the pushforward $transp_*(\omega)$ in (c) boils down to:

$$transp_*(\omega)(\vec{y}) = \sum_{\vec{x} \in X^K} transp(\vec{x})(\vec{y}) \cdot \omega(\vec{x}) = \sum_{t \colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}} \frac{1}{K!} \cdot \omega(\underline{t^{-1}}(\vec{y})) = \sum_{t \colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}} \frac{1}{K!} \cdot \mathcal{D}(\underline{t})(\omega)(\vec{y}).$$

Clearly, the latter equals $\omega(\vec{y})$ if (b) holds. This gives (b) $\Rightarrow$ (c). For the reverse, take an arbitrary transposition $r\colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}$. Then:

$$\begin{aligned}
\mathcal{D}(\underline{r})(\omega)(\vec{x}) \;=\; \omega(\underline{r^{-1}}(\vec{x})) &\overset{(c)}{=} \mathit{transp}_*(\omega)(\underline{r^{-1}}(\vec{x})) \\
&= \sum_{t\colon \boldsymbol{K}\xrightarrow{\cong}\boldsymbol{K}} \frac{1}{K!}\cdot \mathcal{D}(\underline{t})(\omega)(\underline{r^{-1}}(\vec{x})) \\
&= \sum_{t\colon \boldsymbol{K}\xrightarrow{\cong}\boldsymbol{K}} \frac{1}{K!}\cdot \mathcal{D}(\underline{r})\big(\mathcal{D}(\underline{t})(\omega)\big)(\vec{x}) \\
&= \sum_{t\colon \boldsymbol{K}\xrightarrow{\cong}\boldsymbol{K}} \frac{1}{K!}\cdot \mathcal{D}(\underline{r}\circ\underline{t})(\omega)(\vec{x}) \\
&= \sum_{t\colon \boldsymbol{K}\xrightarrow{\cong}\boldsymbol{K}} \frac{1}{K!}\cdot \mathcal{D}(\underline{t})(\omega)(\vec{x}) \;=\; \mathit{transp}_*(\omega)(\vec{x}) \overset{(c)}{=} \omega(\vec{x}).
\end{aligned}$$

4 By the same reasoning. ❑

The situation gets more interesting if we move to the category **Chan** = **Chan**($\mathcal{D}$) of probabilistic channels. We are now going to instantiate the notion of (co)equaliser from Definition 2.11.2 in **Chan**, where arrows are channels. The coequalisers that we saw in **Sets** turn out to be coequalisers in **Chan** as well. The probabilistic inverses of *acc* and *mat* turn out to be equalisers in **Chan**. This is a useful observation, since it allows us to transform exchangeable distributions on sequences into distributions on multisets or set partitions. This forms the basis of the description of a famous result of De Finetti in terms of multisets, see Proposition 3.2.10.

Proposition 2.11.5 *1 When we view all the maps in Diagrams (2.84) and (2.85) as deterministic channels, via* $\langle - \rangle$*, then accumulation and matching are still coequalisers of all transposition / substitution maps, but now in the category* **Chan**.

2 The probabilistic inverse accumulation channel $\mathrm{acc}^{\sim 1} = \mathrm{arr}$ *is the equaliser of all transpositions (as deterministic channels), in:*

$$\mathcal{N}[K](X) \xrightarrow{\mathrm{acc}^{\sim 1}} X^K \underset{\text{transpositions } \underline{t}}{\rightrightarrows} X^K \tag{2.88}$$

3 If the set X is finite and has at least K elements, then the probabilistic inverse match channel $\mathrm{mat}^{\sim 1}$ *is the equaliser of all substitutions (as deterministic channels), in:*

$$SP(K) \xrightarrow{\mathrm{mat}^{\sim 1}} X^K \underset{\text{substitutions } s^\star}{\rightrightarrows} X^K \tag{2.89}$$

Proof. 1 The proofs of Proposition 2.11.3 also work when the map g is a probabilistic function, instead of an ordinary function.

(The abstract argument that category theorists like in this situation is: the functor **Sets** $\to$ **Chan** is a left adjoint and thus preserves colimits, including coequalisers.)

2 For a transposition $t\colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}$ we have, for an arbitrary multiset $\varphi \in \mathcal{N}[K](X)$,

$$\begin{aligned}
(\underline{t} \circ \mathrm{acc}^{\sim 1})(\varphi) \;=\; \mathcal{D}(\underline{t})(\mathrm{acc}^{\sim 1}(\varphi)) &\overset{(2.38)}{=} \sum_{\vec{x}\in \mathrm{acc}^{-1}(\varphi)} \frac{1}{(\varphi)} \big|\underline{t}(\vec{x})\rangle \qquad \text{since } \mathrm{acc}^{\sim 1} = \mathrm{arr} \\
&= \sum_{\vec{x}\in \mathrm{acc}^{-1}(\varphi)} \frac{1}{(\varphi)} \big|\vec{x}\rangle \;=\; \mathrm{acc}^{\sim 1}(\varphi).
\end{aligned}$$

Next, let $g\colon Y \rightsquigarrow X^K$ be a channel with $\underline{t} \circ g = g$ for each $t\colon \boldsymbol{K} \xrightarrow{\cong} \boldsymbol{K}$. This means $\mathcal{D}(\underline{t})(g(y)) =$

$g(y)$, for each t, and thus $transp_*(g(y)) = g(y)$ by Lemma 2.11.4 (3), so that $transp \circ g = g$. We take $h = acc \circ g \colon Y \rightsquigarrow \mathcal{N}[K](X)$. Then:

$$acc^{\sim 1} \circ h \;=\; acc^{\sim 1} \circ acc \circ g \;=\; transp \circ g \;=\; g.$$

If also $k \colon Y \rightsquigarrow \mathcal{N}[K](X)$ satisfies $acc^{\sim 1} \circ k = g$, then $k = h$ since:

$$h \;=\; acc \circ g \;=\; acc \circ acc^{\sim 1} \circ k \;=\; id \circ k \;=\; k.$$

3 Along the same lines, with *mat* in place of *acc*, and *subst* in place of *transp*. ❑

In the above coequaliser and equaliser diagrams we have been using all transpositions and substitutions, as deterministic channels. We can reduce to only two channels, where one of them is the identity.

Theorem 2.11.6 *Consider the category* **Chan** *of probabilistic channels.*

1 *Accumulation and its probabilistic inverse are coequaliser and equaliser of the transposition-idempotent* transp *and the identity:*

$$\begin{tikzcd} \mathcal{N}[K](X) \arrow[r, "acc^{\sim 1}", "= arr"'] & X^K \arrow[r, bend left, "transp"] \arrow[r, bend right, "id"'] & X^K \arrow[r, "acc"] & \mathcal{N}[K](X) \end{tikzcd} \tag{2.90}$$

Explicitly:

- *For a channel $f \colon Y \rightsquigarrow X^K$ with* transp $\circ f = f$*, the unique mediating channel $\overline{f} \colon Y \rightsquigarrow \mathcal{N}[K](X)$ with* arr $\circ \overline{f} = f$ *is* $\overline{f} := acc \circ f$.
- *For a channel $g \colon X^K \rightsquigarrow Y$ with $g \circ$* transp $= g$*, the unique mediating channel $\overline{g} \colon \mathcal{N}[K](X) \rightsquigarrow Y$ with $\overline{g} \circ acc = g$ is $\overline{g} := g \circ$* arr.

2 *Matching and its probabilistic inverse are coequaliser and equaliser of the substitution-idempotent* subst *and the identity:*

$$\begin{tikzcd} SP(K) \arrow[r, "mat^{\sim 1}"] & X^K \arrow[r, bend left, "subst"] \arrow[r, bend right, "id"'] & X^K \arrow[r, "mat"] & SP(K) \end{tikzcd} \tag{2.91}$$

For this we assume that the set X has at least K elements. In this case:

- *For $f \colon Y \rightsquigarrow X^K$ with* subst $\circ f = f$*, one has $\overline{f} :=$* mat $\circ f$ *uniquely satisfying* mat$^{\sim 1} \circ \overline{f} = f$*;*
- *And for $g \colon X^K \rightsquigarrow Y$ with $g \circ$* subst $= g$*, the unique mediating map $\overline{g}$ with $\overline{g} \circ$* mat $= g$ *is $\overline{g} := g \circ$* mat$^{\sim 1}$.

Proof. The two statements can be proven with what we have described earlier in this section. However, both results follow via very general abstract reasoning, in the lemma below. ❑

These results are instances of a more general, purely categorical fact about split idempotents.

Lemma 2.11.7 *Let $f \colon A \to A$ be a split idempotent in an arbitrary category. This means that f can be written as $f = s \circ r$ where $r \circ s = id$. Then:*

1 *s is the equaliser of $f, id \colon A \rightrightarrows A$;*
2 *r is their coequaliser.*

In this situation f is called a split idempotent, with (monic) section s and (epic) retraction r, in a situation:

$$\begin{tikzcd} B \arrow[r, tail, "s"] & A \arrow[r, bend left, "f=s\circ r"] \arrow[r, bend right, "id"'] & A \arrow[r, two heads, "r"] & B \end{tikzcd} \tag{2.92}$$

It may be clear that the two items in the above theorem are both an instance of this general result, interpreted in the category **Chan**, for the split idempotents $transp = acc^{\sim 1} \circledcirc acc$ and $subst = mat^{\sim 1} \circledcirc mat$.

Proof. Let $r\colon A \to B$ and $s\colon B \to A$ be the retraction and section maps, where $f = s \circ r$ and $r \circ s = id$.

1 First, we show that the section is an equaliser: $f \circ s = s \circ r \circ s = s = id \circ s$. Moreover, if $f \circ g = g = id \circ g\colon C \to A$, then there is $\overline{g} := r \circ g\colon C \to B$ satisfying:

- $s \circ \overline{g} = s \circ r \circ g = f \circ g = g$.
- if also $s \circ h = g$, then $h = r \circ s \circ h = r \circ g = \overline{g}$.

2 We need to show that the retraction r is a coequaliser: $r \circ f = r \circ s \circ r = r = r \circ id$. If $g \circ f = g = g \circ id : A \to C$, then $\overline{g} := g \circ s\colon B \to C$ satisfies:

- $\overline{g} \circ r = g \circ s \circ r = g \circ f = g$.
- if also $h \circ r = g$, then $h = h \circ r \circ s = g \circ s = \overline{g}$. ❑

So far this section has concentrated on accumulation and matching, and on their special roles for transposition and substitution. The accumulation and matching maps *acc* and *mat* form a commuting diamond at the top of the triangular prism (1.52), together with the muliplicity count and size count maps *mc* and *sc*, see Definition 1.9.2. The above account in terms of (co)equalisers can be extended to multiplicity and size count. We concentrate on the essentials.

Theorem 2.11.8 *Fix a finite set X and a number $K \geq 1$.*

1 In the category **Chan** *of probabilistic channels, the multiplicity count function* $\mathrm{mc}\colon \mathcal{N}[K](X) \to MP(K)$ *is a coequaliser and its probabilistic inverse* $\mathrm{mc}^{\sim 1}$ *is an equaliser:*

$$\begin{tikzcd} MP(K) \arrow[r, "mc^{\sim 1}", "\circ"'] & \mathcal{N}[K](X) \arrow[r, bend left, "\circ"] \arrow[r, bend right, "\circ"'] \arrow[r, phantom, "substitutions"] & \mathcal{N}[K](X) \arrow[r, "mc", "\circ"'] & MP(K) \end{tikzcd} \tag{2.93}$$

These substitutions are of the form $\mathcal{M}(s)\colon \mathcal{N}[K](X) \xrightarrow{\cong} \mathcal{N}[K](X)$, for $s\colon X \xrightarrow{\cong} X$. The probabilistic inverse $\mathrm{mc}^{\sim 1}\colon MP(K) \rightsquigarrow \mathcal{N}[K](X)$ takes the following form, via Lemma 1.9.6.

$$\mathrm{mc}^{\sim 1}(\alpha) := \sum_{\varphi \in mc^{-1}(\alpha)} \frac{1}{\binom{|X|}{\alpha}} \big|\varphi\rangle = \sum_{\varphi \in mc^{-1}(\alpha)} \frac{\alpha_{\circ}^{!} \cdot (|X| - \|\alpha\|)!}{N!} \big|\varphi\rangle. \tag{2.94}$$

2 Similarly, the size count function $\mathrm{sc}\colon SP(K) \to MP(K)$ *forms a coequaliser, and its probabilistic inverse* $\mathrm{sc}^{\sim 1}$ *an equaliser, in the diagram of channels:*

$$\begin{tikzcd} MP(K) \arrow[r, "sc^{\sim 1}", "\circ"'] & SP(K) \arrow[r, bend left, "\circ"] \arrow[r, bend right, "\circ"'] \arrow[r, phantom, "transpositions"] & SP(K) \arrow[r, "sc", "\circ"'] & MP(K) \end{tikzcd} \tag{2.95}$$

These transpositions in the middle are of the form $\mathcal{P}(\mathcal{P}(t))\colon SP(K) \xrightarrow{\cong} SP(K)$, for $t\colon \{1, \ldots, K\} \xrightarrow{\cong}$

$\{1, \ldots, K\}$, *see Exercise 1.5.4. The probabilistic inverse* $sc^{\sim 1} \colon MP(K) \rightsquigarrow SP(K)$ *is of the following form.*

$$sc^{\sim 1}(\alpha) := \sum_{P \in sc^{-1}(\alpha)} \frac{\alpha!}{(\alpha)_p} \big| P \rangle. \tag{2.96}$$

This is justified by Lemma 1.9.7. ❑

The multiple substitution and transposition maps in (2.93) and (2.95) may be replaced by just two maps, like in (2.92), one of which is the identity.

The equaliser in (2.95) is used implicitly for what Pitman Pitman (1995, 2006) calls an *exchangeable partition probability function* (EPPF) of a distribution $\sigma \in \mathcal{D}(SP(K))$ on set partitions. When such a distribution σ is exchangeable, the EPPF arises via the equaliser $sc^{\sim 1}$ in:

$$\begin{tikzcd}
MP(K) \arrow[r, "sc^{\sim 1}"] & SP(K) \arrow[r, bend left, "\text{transpositions}"] \arrow[r, bend right] & SP(K) \\
\mathbf{1} \arrow[u, dashed, "\text{EPPF}"] \arrow[ur, "\sigma"'] & &
\end{tikzcd}$$

Thus, the EPPF is the distribution on multiset partitions uniquely corresponding to a (transposition) exchangeable distribution on set partitions.

We now have for each of the four basic maps — accumulation, matching, multiplicity count and size count — a probabilistic inverse. There is a bit more to say about their interaction.

Proposition 2.11.9 *1 The following two rectangles commute.*

$$\begin{tikzcd}
\mathcal{N}[K](X) \arrow[r, "acc^{\sim 1}"] \arrow[d, "mc"'] & X^K \arrow[d, "mat"] \\
MP(K) \arrow[r, "sc^{\sim 1}"] & SP(K)
\end{tikzcd}
\qquad
\begin{tikzcd}
SP(K) \arrow[r, "mat^{\sim 1}"] \arrow[d, "sc"'] & X^K \arrow[d, "acc"] \\
MP(K) \arrow[r, "mc^{\sim 1}"] & \mathcal{N}[K](X)
\end{tikzcd}$$

For the diagram on the right we have to assume $|X| \geq K$.

2 The probabilistic inverses commute:

$$\begin{tikzcd}
MP(K) \arrow[r, "sc^{\sim 1}"] \arrow[d, "mc^{\sim 1}"'] & SP(K) \arrow[d, "mat^{\sim 1}"] \\
\mathcal{N}[K](X) \arrow[r, "acc^{\sim 1}"] & X^K
\end{tikzcd}$$

Proof. 1 We use Lemma 1.9.8 for both rectangles. First, for $\varphi \in \mathcal{N}[K](X)$,

$$\begin{aligned}
(sc^{\sim 1} \circ mc)(\varphi) &\overset{(2.96)}{=} \sum_{P \in sc^{-1}(mc(\varphi))} \frac{mc(\varphi)!}{(mc(\varphi))_p} \big| P \rangle \\
&\overset{(1.53)}{=} \sum_{P \in sc^{-1}(mc(\varphi))} \sum_{\vec{x} \in mat^{-1}(P) \cap acc^{-1}(\varphi)} \frac{1}{mc(\varphi)!} \cdot \frac{mc(\varphi)!}{(\varphi)} \big| P \rangle \\
&= \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big| mat(\vec{x}) \rangle \;=\; (mat \circ acc^{\sim 1})(\varphi).
\end{aligned}$$

Similary, when $N = |X| \geq K$,

$$\begin{aligned}
(mc^{\smile 1} \circ sc)(P) &\overset{(2.94)}{=} \sum_{\varphi \in mc^{-1}(sc(P))} \frac{sc(P)\llbracket \cdot (N - \|sc(P)\|)!}{N!} \big|\varphi\rangle \\
&= \sum_{\varphi \in mc^{-1}(sc(P))} \sum_{\vec{x} \in mat^{-1}(P) \cap acc^{-1}(\varphi)} \frac{1}{sc(P)\llbracket} \cdot \frac{sc(P)\llbracket \cdot (N - |P|)!}{N!} \big|\varphi\rangle \\
&= \sum_{\vec{x} \in mat^{-1}(P)} \frac{(N - |P|)!}{N!} \big|acc(\vec{x})\rangle \;=\; (acc \circ mat^{\smile 1})(P).
\end{aligned}$$

2 By unpacking the relevant definitions, for $\alpha \in MP(K)$,

$$\begin{aligned}
(acc^{\smile 1} \circ mc^{\smile 1})(\alpha) &= \sum_{\varphi \in mc^{-1}(\alpha)} \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{\alpha\llbracket \cdot (N - \|\alpha\|)!}{N!} \cdot \frac{1}{(\varphi)} \big|\vec{x}\rangle \\
&\overset{(1.53)}{=} \sum_{\varphi \in mc^{-1}(\alpha)} \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{\alpha\llbracket \cdot (N - \|\alpha\|)!}{N!} \cdot \frac{1}{(\alpha)_p} \big|\vec{x}\rangle \\
&= \sum_{\vec{x} \in acc^{-1}(mc^{-1}(\alpha))} \frac{(N - \|\alpha\|)!}{N!} \cdot \frac{\alpha\llbracket}{(\alpha)_p} \big|\vec{x}\rangle \\
&= \sum_{\vec{x} \in mat^{-1}(sc^{-1}(\alpha))} \frac{(N - \|\alpha\|)!}{N!} \cdot \frac{\alpha\llbracket}{(\alpha)_p} \big|\vec{x}\rangle \\
&= \sum_{P \in sc^{-1}(\alpha)} \sum_{\vec{x} \in mat^{-1}(P)} \frac{(N - |P|)!}{N!} \cdot \frac{\alpha\llbracket}{(\alpha)_p} \big|\vec{x}\rangle \;=\; (mat^{\smile 1} \circ sc^{\smile 1})(\alpha). \qquad \square
\end{aligned}$$

Exercises

2.11.1 We have seen that the match function is stable under substitution. Prove the following slightly more general statement. Let X, Y be isomorphic sets, say via the isomorphism $f \colon X \xrightarrow{\cong} Y$. Show that for a list $\ell \in X^K$,

$$mat(\ell) \;=\; mat(f^K(\ell)).$$

2.11.2 This exercise sketches how coequalisers can be obtained in the category **Sets**. Let $f, g \colon A \to B$ two arbitrary functions. Define the relation $R \subseteq B \times B$ as:

$$R \;:=\; \{(f(a), g(a)) \mid a \in A\}.$$

Now let $\overline{R} \supseteq R$ be the least equivalent relation containing R. Show that the quotient map $q \colon B \twoheadrightarrow B/\overline{R}$ is a coequaliser of f, g.

2.11.3 Recall the uniform projection channel $unpr[K] \colon X^K \rightsquigarrow X$ from Exercise 2.4.9. Check that it can be used to give an axiomatic definition of the frequentist learning map, as in Jacobs (2021c), via the coequaliser property of accumulation:

$$\begin{tikzcd}
X^K \arrow[r, bend right, "\text{transpositions } \underline{t}"'] \arrow[r, bend left] & X^K \arrow[r, "acc"] \arrow[dr, "unpr[K]"'] & \mathcal{N}[K](X) \arrow[d, dashed, "Flrn"] \\
& & \mathcal{D}(X)
\end{tikzcd}$$

2.11.4 Fix a set X and a number $K \geq 1$.

1 Show that the composite $\mathcal{D}(acc) \circ \bigotimes \colon \mathcal{D}(X)^K \to \mathcal{D}(\mathcal{N}[K](X))$ is stable under transposition.

2 Deduce that there is a map from multisets over distributions to distributions over multisets, in:

$$\begin{tikzcd}
\mathcal{D}(X)^K \arrow[rr, bend left, "\text{transpositions } \underline{t}"] \arrow[rr, bend right] & & \mathcal{D}(X)^K \arrow[r, "acc"] \arrow[dr, "\mathcal{D}(acc)\circ\bigotimes"'] & \mathcal{N}[K](\mathcal{D}(X)) \arrow[d, dashed] \\
& & & \mathcal{D}(\mathcal{N}[K](X))
\end{tikzcd}$$

In the next chapter, Sections 3.7 – 3.9 are devoted to this map, where it is called 'parallel multinomial'.

2.11.5 Show that the transposition idempotent $transp = arr \circ acc$ commutes with (big) tensors, as in the following diagram of channels.

$$\begin{tikzcd}
\mathcal{D}(X)^K \arrow[r, "transp"] \arrow[d, "\bigotimes"'] & \mathcal{D}(X)^K \arrow[d, "\bigotimes"] \\
X^K \arrow[r, "transp"] & X^K
\end{tikzcd}$$

2.11.6 Use Proposition 2.11.9 to show that the following diagram of split idempotents commutes.

$$\begin{tikzcd}
& X^K \arrow[dr, "subst"] & \\
X^K \arrow[ur, "transp"] \arrow[dr, "subst"'] & & X^K \\
& X^K \arrow[ur, "transp"'] &
\end{tikzcd}$$

2.11.7 Show also that the transposition idempotent *transp* interacts with zip as in the following diagram, in which *zip* occurs as a deterministic channel.

$$\begin{tikzcd}
\mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[r, "arr\otimes arr"] \arrow[dr, bend right, "arr\otimes arr"'] & X^K \times Y^K \arrow[r, "zip"] & (X \times Y)^K \\
& X^K \times Y^K \arrow[r, "zip"] & (X \times Y)^K \arrow[u, bend right, "transp"']
\end{tikzcd}$$

Informally, this says that transposing a zip of already transposed inputs does not add anything.

2.11.8 Recall the triangular prism diagram (1.52) and define the equality count map $ec \colon X^K \to MP(K)$ as:

$$ec \coloneqq mc \circ acc = sc \circ mat.$$

Develop a theory for *ec*, as we have done for *acc* and *mat* in this section, involving *both* transposition and substitution.

3

Drawing from an urn

Drawing from an urn, filled with thoroughly shuffled balls of different colours, is a basic model of probability, see *e.g.* David (1962); Johnson and Kotz (1977); Pishro-Nik (2014); Ross (2018); Mahmoud (2008) and many other references. In this book such an urn is identified with a multiset over the set of colours. The resulting probabilities arise 'by counting', see the picture in (0.1). The frequentist learning operation gives the probability distribution $Flrn(\upsilon)$ associated with an urn υ and determines the probability of drawing a ball of colour x, see Remark 2.2.1. This probability is the ratio $Flrn(\upsilon)(x) = \frac{\upsilon(x)}{\|\upsilon\|}$ of the number of balls $\upsilon(x)$ of colour x and the total number of balls $\|\upsilon\|$ in the urn (also called the size of the urn / multiset). In the previous chapter, in Section 2.6, we have briefly seen the three basic modes of drawing balls from an urn:

- multinomial, that is, with replacement, or draw-and-replace;
- hypergeometric, that is, without replacement, or draw-and-delete;
- Pólya, that is draw-and-duplicate.

We have seen that all these modes give rise to probabilistic channels.

In this chapter we continue our investigations of drawing from urns, especially from this channel-based perspective. We shall see that several basic properties of drawing are expressed properly in the calculus of channels, involving especially sequential and parallel composition. For instance, applying frequentist learning to the multisets that are obtained from multinomial draws, returns the original distribution if we take the probabilities of these draws into account. We shall express this as a pushforward $Flrn_*(mn[K](\omega)) = \omega$, see Theorem 3.3.4, or equivalently as a commuting diagram of channels, in the style of category theory. Indeed, as we proceed in this chapter, the level of categorical sophistication increases. Hence, at first reading, the later Sections 3.7 – 3.10 may be skipped.

Multisets play a basic role in this chapter, both as urns and as draws from such urns. Hence we start in Section 3.1 with a closer look at multisets, in particular at a probabilistic 'multizip' operation for combining two multisets of the same size. This multizip is the multiset-analogue of the zipping of two lists of equal length, except that it is not deterministic but probabilistic. Multizip interacts nicely with frequentist learning. It can be described in different ways, via accumulation and arrangement, and via couplings of natural multisets. At the end of the chapter, in a categorical setting, we recognise multizip as a monoidal transformation. In Section 3.2 we make a cautious start with the topic of drawing, namely by considering draws of single balls only. It makes sense to consider them first, and to see how they are related to other operations such as frequentist learning and multizip, before we consider draws of arbitrary size. The single draws are in particular relevant in what we call draw-delete cones, which correspond to consistent sequences of (transposition) exchangeable distributions.

In Sections 3.3 – 3.6 we describe many of the basic properties of multinomial, hypergeometric and Pólya distributions. Here our channel-perspective is fully exploited, for instance in properties like: $hg[K] \circ mn[K+L] = mn[K]$. This says that if you first take $K+L$-size multinomial draws, then take K-sized hypergeometric draws, using these multinomial draws as urns, then you may as well take multinomial K-sized draws in the first place. The channel composition $\circ$ in this equation ensures that all probabilities in this statement interact in the appropriate manner. In fact, such equations can be expressed only in a channel-based formalism. And indeed, they have not emerged in the traditional formalism of probability theory.

The subsequent three sections 3.7 – 3.9 deal with a non-trival operation that turns multisets of distributions into distributions of multisets. Categorically, this is a so-called distributive law. It embodies a fundamental relationship between multisets and distributions and allows us to express various basic facts. We spend ample time introducing this 'parallel multinomial' law: Section 3.7 contains no less than four different definitions — all equivalent. Subsequently, various properties are demonstrated of this parallel multinomial law, including commutation with hypergeometric channels, with frequentist learning and with multizip. This parallel multinomial law is a distributive law of monads, of $\mathcal{N}$ over $\mathcal{D}$, giving a composite monad $\mathcal{DN}$. Similarly, there is a distributive law of lists over distributions, giving a monad $\mathcal{DL}$. The final section of this chapter uses such distributions over multisets / lists in discrete Poisson processes, as infinite mixtures of iid and multinomial distributions, where the sizes involved are determined by a Poisson rate parameter.

This chapter demonstrates that the formalism of category theory helps to navigate the intricacies of drawing from an urn — and of other probabilistic operations in other chapters too. More boldly, one can say that drawing from urns is one of the most convincing applications of category theory, where its language is crucial for expressing the most elementary properties — many of which are missing in the literature.

In the beginning of Chapter 1 we have seen a table describing elementary properties of lists, multisets and subsets. At the end of this chapter we can extend this table in the following manner.

	lists	**subsets**	**multisets**
order matters	+	-	-
multiplicity matters	+	-	+
parallel products $\otimes$ exist	-	+	+
distributive law with $\mathcal{D}$ exists	+	-	+

This last property, involving a distributive law, is unexplained at this stage and is mathematically most sophisticated. It will appear in Section 3.9.

3.1 Zipping multisets

Multisets play an important role in drawing from an urn. The draws themselves are described as multisets, but also the urns, in the hypergeometric and Pólya cases. This first section is preparatory and looks at a zip operation for multisets, as a way of probabilistically combining two multisets of the same size. This operation is called multizip and is written as *mzip*. It is of a fundamental nature.

For two (natural) multisets $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[L](Y)$ we can form their tensor product

$\varphi \otimes \psi \in \mathcal{N}[K \cdot L](X \times Y)$. The fact that it has size $K \cdot L$ follows from Lemma 2.3.2 (1). For sequences there is the familiar *zip* combination map $X^K \times Y^K \to (X \times Y)^K$ that does maintain size, see Exercise 1.3.7. Interestingly, there is also a zip-like operation for natural multisets of the same size, with outcomes of this same size. This multizip operation *mzip* makes systematic use of accumulation *acc* and its probabilistic inverse arrangement $arr = acc^{\sim 1}$. The multizip was first described in Jacobs (2021a). As will be shown, it can also be formulated in terms of couplings.

The idea is the following. In order to combine two natural multisets $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[K](Y)$ of the same size K but on different sets X and Y, we first look at all lists $\vec{x} \in X^K$ and $\vec{y} \in Y^K$ that accumulate to φ and ψ, that is, with $acc(\vec{x}) = \varphi$ and $acc(\vec{y}) = \psi$. We then zip the two lists $\vec{x} \in X^K$ and $\vec{y} \in Y^K$ into a single list $zip(\vec{x}, \vec{y}) \in (X \times Y)^K$. Finally we apply accumulation $acc\colon (X \times Y)^K \to \mathcal{M}[K](X \times Y)$, giving a combined multiset $mzip(\varphi, \psi) \in \mathcal{M}[K](X \times Y)$, still of size K. Diagrammatically we may describe *mzip* thus as the following composite.

$$\mathcal{N}[K](X) \times \mathcal{N}[K](Y) \xrightarrow{arr \otimes arr} \mathcal{D}(X^K \times Y^K) \underset{\cong}{\xrightarrow{\mathcal{D}(zip)}} \mathcal{D}((X \times Y)^K) \xrightarrow{\mathcal{D}(acc)} \mathcal{D}(\mathcal{N}[K](X \times Y)) \tag{3.1}$$

Definition 3.1.1 For two sets X, Y and a number $K \in \mathbb{N}$ we define the multizip channel $mzip[K]\colon \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \rightsquigarrow \mathcal{N}[K](X \times Y)$ as the composite:

$$mzip[K] \coloneqq acc \circ zip \circ (arr \otimes arr). \tag{3.2}$$

By construction, $mzip[K]$ makes the following rectangle commute, corresponding to (3.1).

$$\begin{tikzcd}
\mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[r, "mzip[K]"] \arrow[d, "arr \otimes arr"'] & \mathcal{N}[K](X \times Y) \\
X^K \times Y^K \arrow[r, "zip"] & (X \times Y)^K \arrow[u, "acc"']
\end{tikzcd}$$

Explicitly, for $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[K](Y)$,

$$mzip[K](\varphi, \psi) \coloneqq \sum_{\vec{x} \in acc^{-1}(\varphi)} \sum_{\vec{y} \in acc^{-1}(\psi)} \frac{1}{(\varphi) \cdot (\psi)} \Big| acc(zip(\vec{x}, \vec{y}))\Big\rangle. \tag{3.3}$$

We drop the parameter K in $mzip[K]$ when it is clear from the context.

An illustration may help to see what happens here.

Example 3.1.2 Let's use two set $X = \{a, b\}$ and $Y = \{0, 1\}$ with two multisets of size three:

$$\varphi = 1|a\rangle + 2|b\rangle \qquad \text{and} \qquad \psi = 2|0\rangle + 1|1\rangle.$$

Then:

$$(\varphi) = \binom{3}{1,2} = 3 \qquad\qquad (\psi) = \binom{3}{2,1} = 3.$$

The sequences in X^3 and Y^3 that accumulate to φ and ψ are:

$$\begin{cases} a, b, b \\ b, a, b \\ b, b, a \end{cases} \qquad \text{and} \qquad \begin{cases} 0, 0, 1 \\ 0, 1, 0 \\ 1, 0, 0. \end{cases}$$

Zipping them together gives the following nine sequences in $(X \times Y)^3$.

$$\begin{array}{lll}
(a,0)\ (b,0)\ (b,1) & (b,0)\ (a,0)\ (b,1) & (b,0)\ (b,0)\ (a,1) \\
(a,0)\ (b,1)\ (b,0) & (b,0)\ (a,1)\ (b,0) & (b,0)\ (b,1)\ (a,0) \\
(a,1)\ (b,0)\ (b,0) & (b,1)\ (a,0)\ (b,0) & (b,1)\ (b,0)\ (a,0).
\end{array}$$

By applying the accumulation function *acc* to each of these we get multisets:

$$\begin{array}{ccc}
1|a,0\rangle + 1|b,0\rangle + 1|b,1\rangle & 1|b,0\rangle + 1|a,0\rangle + 1|b,1\rangle & 2|b,0\rangle + 1|a,1\rangle \\
1|a,0\rangle + 1|b,1\rangle + 1|b,0\rangle & 2|b,0\rangle + 1|a,1\rangle & 1|b,0\rangle + 1|b,1\rangle + 1|a,0\rangle \\
1|a,1\rangle + 2|b,0\rangle & 1|b,1\rangle + 1|a,0\rangle + 1|b,0\rangle & 1|b,1\rangle + 1|b,0\rangle + 1|a,0\rangle.
\end{array}$$

We see that are only two different multisets involved. Counting them and multiplying with $\frac{1}{(\varphi)(\psi)} = \frac{1}{9}$ gives:

$$mzip[3]\Big(1|a\rangle + 2|b\rangle,\ 2|0\rangle + 1|1\rangle\Big) \;=\; \tfrac{1}{3}\Big|\,1|a,1\rangle + 2|b,0\rangle\Big\rangle \;+\; \tfrac{2}{3}\Big|\,1|a,0\rangle + 1|b,0\rangle + 1|b,1\rangle\Big\rangle.$$

This shows that calculating *mzip* is laborious. But it is quite mechanical and easy to implement.

The next result contains some burocracy on the *mzip* operation showing that it satisfies several reasonable properties.

Proposition 3.1.3 *Consider* mzip$\colon \mathcal{N}[K](X)\times\mathcal{N}[K](Y) \rightsquigarrow \mathcal{N}[K](X\times Y)$*, either in in diagrammatic form* (3.1) *or in its equational formulation* (3.3)*.*

1 *The* mzip *map is natural in X and Y: for functions $f\colon X \to U$ and $g\colon Y \to V$ the following diagram commutes.*

$$\begin{tikzcd}
\mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[r, "mzip"] \arrow[d, "\mathcal{N}(f)\times\mathcal{N}(g)"'] & \mathcal{D}(\mathcal{N}[K](X \times Y)) \arrow[d, "\mathcal{D}(\mathcal{N}(f\times g))"] \\
\mathcal{N}[K](U) \times \mathcal{N}[K](V) \arrow[r, "mzip"] & \mathcal{D}(\mathcal{N}[K](U \times V))
\end{tikzcd}$$

2 *For $\varphi \in \mathcal{N}[K](X)$ and $y \in Y$,*

$$mzip(\varphi, K|y\rangle) \;=\; 1\Big|\varphi \otimes 1|y\rangle\Big\rangle.$$

And similarly in symmetric form.

3 *Multizip commutes with projections in the following sense.*

$$\begin{tikzcd}
\mathcal{N}[K](X) \arrow[d, "unit"'] & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[l, "\pi_1"'] \arrow[r, "\pi_2"] \arrow[d, "mzip"] & \mathcal{N}[K](Y) \arrow[d, "unit"] \\
\mathcal{D}(\mathcal{N}[K](X)) & \mathcal{D}(\mathcal{N}[K](X \times Y)) \arrow[l, "\mathcal{D}(\mathcal{N}(\pi_1))"'] \arrow[r, "\mathcal{D}(\mathcal{N}(\pi_2))"] & \mathcal{D}(\mathcal{N}[K](Y))
\end{tikzcd}$$

This means that $\langle \mathcal{N}(\pi_1), \mathcal{N}(\pi_2)\rangle \circ$ mzip $=$ id, and thus that in the other order mzip$\circ\langle \mathcal{N}(\pi_1), \mathcal{N}(\pi_2)\rangle\colon \mathcal{N}[K](X\times Y) \rightsquigarrow \mathcal{N}[K](X \times Y)$ *is a split idempotent channel.*

4 *Arrangement* arr *relates* zip *and* mzip *as in:*

$$\begin{tikzcd}
\mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[r, "arr\otimes arr", "\circ"'] \arrow[d, "mzip"', "\circ"] & X^X \times Y^K \arrow[d, "zip", "\circ"'] \\
\mathcal{N}[K](X \times Y) \arrow[r, "arr", "\circ"'] & (X \times Y)^K
\end{tikzcd}$$

5 *Multizip is associative, as given by:*

$$\begin{tikzcd}
\mathcal{N}[K](X) \times \mathcal{N}[K](Y) \times \mathcal{N}[K](Z) \arrow[r, "id\otimes mzip", "\circ"'] \arrow[d, "mzip\otimes id"', "\circ"] & \mathcal{N}[K](X) \times \mathcal{N}[K](Y \times Y) \arrow[d, "mzip", "\circ"'] \\
\mathcal{N}[K](X \times Y) \times \mathcal{N}[K](Z) \arrow[r, "mzip", "\circ"'] & \mathcal{N}[K](X \times Y \times Z)
\end{tikzcd}$$

Here we take associativity of $\times$ for granted.

Proof. 1 Easy, via the diagrammatic formulation (3.1), using naturality of *arr* (Exercise 2.4.15), of *zip* (Exercise 1.12.4), and of *acc* (Exercise 1.7.15).

2 Since:

$$\begin{aligned} mzip(\varphi, K|y\rangle) &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \Big| acc(zip(\vec{x}, \langle y, \ldots, y\rangle))\Big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \Big| acc(\overrightarrow{x_i, y})\Big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \Big| acc(\vec{x}) \otimes 1|y\rangle\Big\rangle \;=\; \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \Big|\varphi \otimes 1|y\rangle\Big\rangle \;=\; 1\Big|\varphi \otimes 1|y\rangle\Big\rangle. \end{aligned}$$

3 By naturality of *acc* and *zip*:

$$\begin{aligned} \mathcal{D}(\mathcal{N}(\pi_1))\Big(mzip(\varphi, \psi)\Big) &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \sum_{\vec{y}\in acc^{-1}(\psi)} \frac{1}{(\varphi)\cdot(\psi)} \Big| \mathcal{N}(\pi_1)(acc(zip(\vec{x}, \vec{y})))\Big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \sum_{\vec{y}\in acc^{-1}(\psi)} \frac{1}{(\varphi)\cdot(\psi)} \Big| acc((\pi_1)^K(zip(\vec{x}, \vec{y})))\Big\rangle \\ &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \sum_{\vec{y}\in acc^{-1}(\psi)} \frac{1}{(\varphi)\cdot(\psi)} \Big| acc(\vec{x})\Big\rangle \;=\; \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \big|\varphi\rangle \;=\; 1\big|\varphi\rangle. \end{aligned}$$

4 Directly from Exercise 2.11.7.

5 Consider the following diagram chase, obtained by unpacking the (diagrammatic) definition of *mzip* on the right-hand side.

$$\begin{tikzcd}
\mathcal{N}[K](X)\times\mathcal{N}[K](Y)\times\mathcal{N}[K](Z) \arrow[rr, "id\otimes mzip"] \arrow[dr, "arr\otimes arr\otimes arr"] \arrow[dd, "mzip\otimes id"'] & & \mathcal{N}[K](X)\times\mathcal{N}[K](Y\times Z) \arrow[d, "arr\otimes arr"] \arrow[ddd, bend left, "mzip"] \\
& X^K\times Y^K\times Z^K \arrow[r, "id\otimes zip"] \arrow[d, "zip\otimes id"'] & X^K\times(Y\times Z)^K \arrow[d, "zip"] \\
\mathcal{N}[K](X)\times\mathcal{N}[K](Y\times Z) \arrow[r, "arr\otimes arr"] \arrow[d, "mzip"'] & (X\times Y)^K\times Z^K \arrow[r, "zip"] & (X\times Y\times Z)^K \arrow[d, "acc"] \\
\mathcal{N}[K](X\times Y\times Z) \arrow[urr, "arr"] \arrow[rr, equal] & & \mathcal{N}[K](X\times Y\times Z)
\end{tikzcd}$$

We can see that the outer diagram commutes by going through the internal subdiagrams. In the middle we use associativity of the (ordinary) *zip* function, formulated in terms of (deterministic) channels. Three of the (other) internal subdiagrams commute by item (4). The *acc*-*arr* triangle at the bottom commutes by (2.39). ❑

The following result deserves a separate status. It tells that what we learn from a multiset zip is the same as what we learn from a parallel product (of multisets).

Theorem 3.1.4 *Multiset zip and frequentist learning interact well, namely as:*

$$Flrn_*(mzip(\varphi, \psi)) \;=\; Flrn(\varphi \otimes \psi) \;=\; Flrn(\varphi) \otimes Flrn(\psi).$$

The second equation is known from Lemma 2.3.2 (5) but is added for completeness. In diagram-

matic form the first equation reads:

$$\begin{tikzcd} \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[r, "mzip"] \arrow[d, "\otimes"'] & \mathcal{N}[K](X \times Y) \arrow[d, "Flrn"] \\ \mathcal{N}[K^2](X \times Y) \arrow[r, "Flrn"] & X \times Y \end{tikzcd}$$

Proof. Let multisets $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[K](Y)$ be given and let $a \in X$ and $b \in Y$ be arbitrary elements. We need to show that the probability:

$$Flrn_*(mzip(\varphi, \psi))(a, b) \;=\; \sum_{\vec{x} \in acc^{-1}(\varphi)} \; \sum_{\vec{y} \in acc^{-1}(\psi)} \frac{acc(zip(\vec{x}, \vec{y}))(a, b)}{K \cdot \mathbf{(}\varphi\mathbf{)} \cdot \mathbf{(}\psi\mathbf{)}}$$

is the same as the probability:

$$Flrn(\varphi \otimes \psi)(a, b) \;=\; \frac{\varphi(a) \cdot \psi(a)}{K \cdot K}.$$

We reason informally, as follows. For arbitrary $\vec{x} \in acc^{-1}(\varphi)$ and $\vec{y} \in acc^{-1}(\psi)$ we need to find the fraction of occurrences (a, b) in $zip(\vec{x}, \vec{y})$. The fraction of occurrences of a in $\vec{x}$ is $\frac{\varphi(a)}{K} = Flrn(\varphi)(a)$, and the fraction of occurrences of b in $\vec{y}$ is $\frac{\psi(b)}{K} = Flrn(\psi)(b)$. Hence the fraction of occurrences of (a, b) in $zip(\vec{x}, \vec{y})$ is $Flrn(\varphi)(a) \cdot Flrn(\psi)(b) = Flrn(\varphi \otimes \psi)(a, b)$. ❑

We continue with an alternative way to describe multizip[1]. It uses the concept of a coupling for natural multisets. We have described couplings of distributions σ, τ in Definition 2.3.8 (3) as a joint distribution that marginalises to σ and to τ. We first give a general formulation of coupling with respect a functor F. We shall use it below with $F = \mathcal{N}[K]$. In the probabilistic case one uses $F = \mathcal{D}$.

Definition 3.1.5 Let $F\colon \mathbf{Sets} \to \mathbf{Sets}$ be a functor. A *coupling* of two elements $a \in F(X)$ and $b \in F(Y)$ is an element $c \in F(X \times Y)$ with $F(\pi_1)(c) = a$ and $F(\pi_2)(c) = b$.

Let's introduce the *decouple* map *dcpl* as:

$$dcpl \;:=\; \Big(F(X \times Y) \xrightarrow{\langle F(\pi_1), F(\pi_2) \rangle} F(X) \times F(Y)\Big). \tag{3.4}$$

A coupling of (a, b) is then an element c with $dcpl(c) = (a, b)$, that is, with $c \in dcpl^{-1}(a, b)$.

A mentioned we use this general definition with $F = \mathcal{N}[K]$, so that the decouple map in (3.4) is the pair of marginalisation functions. In Proposition 3.1.3 (3) we have seen that $mzip(\varphi, \psi)$ is a distribution of couplings of φ, ψ. In Proposition 3.1.7 it will be shown that $mzip(\varphi, \psi)$ is a distribution of all such couplings, with probabilities that can be described in terms of multiset coefficients.

The next result uses the zip isomorphism $zip\colon X^K \times Y^K \xrightarrow{\cong} (X \times Y)^K$ from Exercise 1.3.7. As shown there, its inverse is the tuple $\langle (\pi_1)^K, (\pi_2)^K \rangle\colon (X \times Y)^K \xrightarrow{\cong} X^K \times Y^K$.

Lemma 3.1.6 *Consider the setting of Definition 3.1.5.*

[1] Developed together with Dario Stein.

1 *There is a commuting triangle of the form:*

$$\begin{tikzcd} X^K \times Y^K \arrow[r, "zip", "\cong"'] \arrow[dr, two heads, "acc\times acc"'] & (X \times Y)^K \arrow[r, two heads, "acc"] & \mathcal{N}[K](X \times Y) \arrow[dl, two heads, "dcpl"] \\ & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) & \end{tikzcd} \tag{3.5}$$

As a result, the decouple function $dcpl\colon \mathcal{N}[K](X \times Y) \to \mathcal{N}[K](X) \times \mathcal{N}[K](Y)$ *is surjective.*

2 *There is the following equality of subsets of* $\mathcal{N}[K](X \times Y)$.

$$dcpl^{-1}(\varphi, \psi) \;=\; \{\, acc(zip(\vec{x}, \vec{y})) \;\big|\; \vec{x} \in acc^{-1}(\varphi), \vec{y} \in acc^{-1}(\psi) \,\}.$$

In particular, the subset $dcpl^{-1}(\varphi, \psi) \subseteq \mathcal{N}[K](X \times Y)$ *is finite.*

3 *Further:*

$$\sum_{\chi \in dcpl^{-1}(\varphi,\psi)} (\chi) \;=\; (\varphi) \cdot (\psi).$$

Proof. 1 Via the naturality of *acc*, see Exercise 1.7.15: for $\vec{x} \in acc^{-1}(\varphi)$ and $\vec{y} \in acc^{-1}(\psi)$,

$$\begin{aligned} dcpl\Big(acc(zip(\vec{x}, \vec{y}))\Big) &\overset{(3.4)}{=} \big(\, \mathcal{N}(\pi_1)\Big(acc(zip(\vec{x}, \vec{y}))\Big),\; \mathcal{N}(\pi_2)\Big(acc(zip(\vec{x}, \vec{y}))\Big) \,\big) \\ &= \big(\, acc\Big((\pi_1)^K(zip(\vec{x}, \vec{y}))\Big),\; acc\Big((\pi_2)^K(zip(\vec{x}, \vec{y}))\Big) \,\big) \\ &= \big(\, acc(\pi_1(\vec{x}, \vec{y})),\; acc(\pi_2(\vec{x}, \vec{y})) \,\big) \;=\; \big(\, acc(\vec{x}),\; acc(\vec{y}) \,\big) \;=\; (\,\varphi,\; \psi\,). \end{aligned}$$

This allows us to show that $dcpl\colon \mathcal{N}[K](X \times Y) \to \mathcal{N}[K](X) \times \mathcal{N}[K](Y)$ is a surjective function: let a pair $(\varphi, \psi) \in \mathcal{N}[K](X) \times \mathcal{N}[K](Y)$ be given. Using the surjectivity of accumulation we can find $\vec{x} \in X^K$ and $\vec{y} \in Y^K$ with $acc(\vec{x}) = \varphi$ and $acc(\vec{y}) = \psi$. The above argument shows that $\chi := acc(zip(\vec{x}, \vec{y})) \in \mathcal{N}[K](X \times Y)$ satisfies $dcpl(\chi) = (acc(\vec{x}), acc(\vec{y})) = (\varphi, \psi)$.

2 The previous point gives the inclusion ($\supseteq$). For ($\subseteq$), let $\chi \in dcpl^{-1}(\varphi, \psi)$ be given. There is a sequence $\vec{z} \in (X \times Y)^K$ with $acc(\vec{z}) = \chi$. Write $\vec{z} = zip(\vec{x}, \vec{y})$, for $\vec{x} = (\pi_1)^K(\vec{z})$ and $\vec{y} = (\pi_2)^K(\vec{z})$. We claim that $acc(\vec{x}) = \varphi$ and $acc(\vec{y}) = \psi$. We elaborate only the first equation, since the second one is obtained in the same way. By assumption, χ is coupling of φ, ψ, so:

$$\varphi \;=\; \mathcal{N}(\pi_1)(\chi) \;=\; \mathcal{N}(\pi_1)(acc(zip(\vec{x}, \vec{y}))) \;=\; acc(\vec{x}).$$

The last equation is obtained as in the previous point.

3 By:

$$\begin{aligned} \sum_{\chi \in dcpl^{-1}(\varphi,\psi)} (\chi) &= \sum_{\chi \in dcpl^{-1}(\varphi,\psi)} \Big| acc^{-1}(\chi) \Big| && \text{by Proposition 1.7.2} \\ &= \Big| \bigcup_{\chi \in dcpl^{-1}(\varphi,\psi)} acc^{-1}(\chi) \Big| && \text{since } \bigcup \text{ is a disjoint union} \\ &= \Big| (dcpl \circ acc)^{-1}(\varphi, \psi) \Big| \\ &= \Big| (dcpl \circ acc \circ zip)^{-1})(\varphi, \psi) \Big| && \text{since } zip \text{ is an isomorphism} \\ &\overset{(3.5)}{=} \Big| (acc \times acc)^{-1}(\varphi, \psi) \Big| \\ &= \Big| (acc)^{-1}(\varphi) \times (acc)^{-1}(\psi) \Big| \\ &= \Big| (acc)^{-1}(\varphi) \Big| \cdot \Big| (acc)^{-1}(\psi) \Big| \\ &= (\varphi) \cdot (\psi) && \text{by Proposition 1.7.2 again.} \end{aligned}$$

❑

With the results in this lemma we obtain the promised alternative formulation of multizip in terms of couplings.

Proposition 3.1.7 *We can also describe multizip as:*

$$mzip(\varphi, \psi) \;=\; \sum_{\chi \in dcpl^{-1}(\varphi,\psi)} \frac{(\chi)}{(\varphi)\cdot(\psi)} \big|\chi\rangle. \tag{3.6}$$

On the right-hand-side there is a proper distribution by Lemma 3.1.6 (3).

Proof. By the following reasoning, starting with Proposition 1.7.2.

$$\begin{aligned}\sum_{\chi \in dcpl^{-1}(\varphi,\psi)} \frac{(\chi)}{(\varphi)\cdot(\psi)} \big|\chi\rangle \;&=\; \sum_{\vec{z} \in (X\times Y)^K} \frac{1}{(\varphi)\cdot(\psi)} \big|acc(\vec{z})\rangle \\ &=\; \sum_{\vec{x} \in X^K} \sum_{\vec{y} \in Y^K} \frac{1}{(\varphi)\cdot(\psi)} \big|acc(zip(\vec{x},\vec{y}))\rangle \;=\; mzip(\varphi,\psi). \qquad \square\end{aligned}$$

Once we have seen the definition of *mzip*, via 'deconstruction' of multisets into lists, a *zip* operation on lists, and 'reconstruction' to a multiset result, we can try to apply this approach more widely. For instance, instead of using a *zip* on lists we can simply concatenate (++) the lists — assuming they contain elements from the same set. This yields, like in (3.1), a composite channel:

$$\mathcal{N}[K](X) \times \mathcal{N}[L](X) \xrightarrow{arr \otimes arr} \mathcal{D}(X^K \times X^L) \xrightarrow[\cong]{\mathcal{D}(+\!+)} \mathcal{D}(X^{K+L}) \xrightarrow{\mathcal{D}(acc)} \mathcal{D}(\mathcal{N}[K+L](X))$$

It is easy to see that this yields addition of multisets, as a deterministic channel.

In Section 2.11 we have seen formal similarities between accumulation *acc* and matching *mat*. We can try the analogue of *mzip* for matchting. It is not so interesting since it yields a deterministic function.

Lemma 3.1.8 *Let X, Y be finite sets with $|X| \geq K$ and $|Y| \geq K$, for some number $K \in \mathbb{N}$. The composite channel $SP(K) \times SP(K) \rightarrowtail SP(K)$ in:*

$$SP(K) \times SP(L) \xrightarrow{mat^{-1} \otimes mat^{-1}} X^K \times Y^K \xrightarrow[\cong]{zip} (X \times Y)^K \xrightarrow{mat} SP(K)$$

is deterministic; it sends set partitions $P, Q \in SP(K)$ to their 'intersection':

$$P \,\&\, Q \;:=\; \Big\{ B \cap C \;\Big|\; B \in P,\, C \in Q,\, B \cap C \neq \emptyset \Big\}.$$

Proof. Let $D \in mat(zip(\vec{x},\vec{y}))$ be a block, for sequences $\vec{x} \in mat^{-1}(P)$ and $\vec{y} \in mat^{-1}(Q)$, say $D = \{i_1, \ldots, i_n\}$ for $n \geq 2$. Then $zip(\vec{x},\vec{y})_{i_j} = zip(\vec{x},\vec{y})_{i_k}$ for each $1 \leq j, k \leq n$, and thus $x_{i_j} = x_{i_k}$ and $y_{i_j} = y_{i_k}$. This means that $D = B \cap C$, for blocks $B \in P$ and $C \in Q$. The same argument can be used to show that the non-empty intersection of block $B \cap C$, for $B \in P$, $C \in Q$, must be in $mat(zip(\vec{x},\vec{y}))$. The argument does not depend on the choice over $\vec{x}$ or $\vec{y}$, and thus gives a deterministic channel. $\square$

Exercises

3.1.1 Show that:

$$\begin{aligned} mzip[4]\big(1|a\rangle + 2|b\rangle + 1|c\rangle,\ 3|0\rangle + 1|1\rangle\big) &= \tfrac{1}{4}\big|1|a,1\rangle + 2|b,0\rangle + 1|c,0\rangle\big\rangle \\ &\quad + \tfrac{1}{2}\big|1|a,0\rangle + 1|b,0\rangle + 1|b,1\rangle + 1|c,0\rangle\big\rangle \\ &\quad + \tfrac{1}{4}\big|1|a,0\rangle + 2|b,0\rangle + 1|c,1\rangle\big\rangle. \end{aligned}$$

3.1.2 Show in the context of the previous exercise that:

$$\begin{aligned} &Flrn_*(mzip[4]\big(1|a\rangle + 2|b\rangle + 1|c\rangle,\ 3|0\rangle + 1|1\rangle\big)) \\ &= \tfrac{3}{16}|a,0\rangle + \tfrac{1}{16}|a,1\rangle + \tfrac{3}{8}|b,0\rangle + \tfrac{1}{8}|b,1\rangle + \tfrac{3}{16}|c,0\rangle + \tfrac{1}{16}|c,1\rangle. \end{aligned}$$

Compute also: $Flrn(1|a\rangle + 2|b\rangle + 1|c\rangle) \otimes Flrn(3|0\rangle + 1|1\rangle)$, and remember Theorem 3.1.4.

3.1.3 Consider Definition 3.1.5 for the K-fold product functor $F = (-)^K$. Check that every pair of sequences $\vec{x} \in X^K$ and $\vec{y} \in Y^K$ has precisely one coupling, namely $zip(\vec{x}, \vec{y}) \in (X \times Y)^K$.

3.1.4 Consider the two multisets $\varphi = 2|a\rangle + 2|b\rangle \in \mathcal{N}[4](\{a,b\})$ and $\psi = 2|0\rangle + 1|1\rangle + 1|2\rangle \in \mathcal{N}[4](\{0,1,2\})$. Write down all four couplings $\tau \in \mathcal{N}[4](\{a,b\} \times \{0,1,2\})$ of φ and ψ.

3.1.5 Let $\tau \in \mathcal{N}[K](X \times Y)$ be a coupling of multisets $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[K](Y)$. Show that $Flrn(\tau) \in \mathcal{D}(X \times Y)$ is then a coupling of distributions $Flrn(\varphi) \in \mathcal{D}(X)$ and $Flrn(\psi) \in \mathcal{D}(Y)$, see Definition 2.3.8 (3).

3.1.6 1 Check that *mzip* does *not* commute with diagonals, in the sense that the following triangle does *not* commute.

$$\begin{tikzcd} \mathcal{N}[K](X) \arrow[r, "\Delta"] \arrow[dr, bend right, "\langle \mathcal{N}[K](\Delta) \rangle"'] \arrow[dr, phantom, "\neq"] & \mathcal{N}[K](X) \times \mathcal{N}[K](X) \arrow[d, "mzip"] \\ & \mathcal{D}(\mathcal{N}[K](X \times X)) \end{tikzcd}$$

Hint: Consider for instance the multiset $1|a\rangle + 1|b\rangle$.

2 Check that *mzip* and *zip* do *not* commute with accumulation, as in:

$$\begin{tikzcd} X^K \times Y^K \arrow[r, "acc \times acc"] \arrow[d, "zip"', "\cong"] \arrow[dr, phantom, "\neq"] & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[d, "mzip"] \\ (X \times Y)^K \arrow[r, "\langle acc \rangle"] & \mathcal{D}(\mathcal{N}[K](X \times Y)) \end{tikzcd}$$

Hint: Take sequences $[a,b,b]$, $[0,0,1]$ and re-use Example 3.1.2.

3.1.7 Multizip does *not* commute with sums, in the sense that, in general,

$$mzip(\varphi_1 + \varphi_2,\ \psi_1 + \psi_2) \neq mzip(\varphi_1, \psi_1) + mzip(\varphi_2, \psi_2).$$

where the latter + is convolution. Take:

$$\begin{cases} \varphi_1 = 1|0\rangle + 1|1\rangle \\ \psi_1 = 1|a\rangle + 1|b\rangle \end{cases} \qquad \begin{cases} \varphi_2 = 2|0\rangle + 1|1\rangle \\ \psi_2 = 1|a\rangle + 2|b\rangle. \end{cases}$$

Check consecutively:

1 the multizip equations:

$$\begin{aligned}
mzip(\varphi_1, \psi_1) &= \tfrac{1}{2}\big|1|0,b\rangle + 1|1,a\rangle\big\rangle + \tfrac{1}{2}\big|1|0,a\rangle + 1|1,b\rangle\big\rangle \\
mzip(\varphi_2, \psi_2) &= \tfrac{1}{3}\big|2|0,b\rangle + 1|1,a\rangle\big\rangle + \tfrac{2}{3}\big|1|0,a\rangle + 1|0,b\rangle + 1|1,b\rangle\big\rangle \\
mzip(\varphi_1 + \varphi_2,\, \psi_1 + \psi_2) &= \tfrac{1}{10}\big|3|0,b\rangle + 2|1,a\rangle\big\rangle + \tfrac{3}{5}\big|1|0,a\rangle + 2|0,b\rangle + 1|1,a\rangle + 1|1,b\rangle\big\rangle \\
&\quad + \tfrac{3}{10}\big|2|0,a\rangle + 1|0,b\rangle + 2|1,b\rangle\big\rangle.
\end{aligned}$$

2 the convolution:

$$\begin{aligned}
mzip(\varphi_1, \psi_1) + mzip(\varphi_2, \psi_2) &= \tfrac{1}{3}\big|2|0,a\rangle + 1|0,b\rangle + 2|1,b\rangle\big\rangle \\
&\quad + \tfrac{1}{2}\big|1|0,a\rangle + 2|0,b\rangle + 1|1,a\rangle + 1|1,b\rangle\big\rangle \\
&\quad + \tfrac{1}{6}\big|3|0,b\rangle + 2|1,a\rangle\big\rangle.
\end{aligned}$$

3.1.8 Compute the intersection of the two set partitions of 10 below.

$$\Big\{\{1,2,3,4,5,9,10\},\{6,7,8\}\Big\} \qquad \Big\{\{1,6\},\{2,3,10\},\{4,8,9\},\{5\},\{7\}\Big\}.$$

3.2 Single draws

In Section 2.6 we have encountered distributions on draws form an urn, where both the draws and the urns were represented as multisets. As part of a more systematic perspective on draws we look in this section at draws of a single ball from an urn (as a multiset). One can then remove the ball (*i.e.* not return it to the urn) or add another ball of the same colour to the urn. We call these operations draw-delete DD and draw-add DA, respectively. We will describe them as channels. They resemble hypergeometric and Pólya draws, but there are subtle differences. These hypergeometric and Pólya channels are operations from urns to draws. The draw-delete and draw-add maps are channels from urns to urns.

We start with draw-delete and first consider two operations for probabilistically deleting an element, via projection from a sequence, or via drawing from a multiset. They turn out to be closely related, via accumulation and arrangement.

Definition 3.2.1 Fix a set X and a number $K \in \mathbb{N}$.

1 The projection-delete channel $PD\colon X^{K+1} \rightsquigarrow X^K$ is defined as:

$$PD(x_1, \ldots, x_{K+1}) \coloneqq \sum_{1 \le i \le K+1} \frac{1}{K+1}\big|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{K+1}\rangle. \tag{3.7}$$

It forms a uniform distribution over the sequences of length K obtained by separately removing each of the single elements of the original sequence of length $K+1$.

2 The draw-delete channel $DD\colon \mathcal{N}[K+1](X) \rightsquigarrow \mathcal{N}[K](X)$ is defined on a multiset $\psi \in \mathcal{N}[K+1](X)$ as:

$$DD(\psi) \coloneqq \sum_{x \in supp(\psi)} \frac{\psi(x)}{K+1}\Big|\psi - 1|x\rangle\Big\rangle = \sum_{x \in supp(\psi)} Flrn(\psi)(x)\Big|\psi - 1|x\rangle\Big\rangle. \tag{3.8}$$

The two channels are closely related.

Lemma 3.2.2 *The projection-delete and draw-delete channels commute with both accumulation and arrangement, as expressed by the diagrams below.*

$$\begin{tikzcd} X^K \arrow[d, "acc"'] & X^{K+1} \arrow[l, "PD"'] \arrow[d, "acc"] \\ \mathcal{N}[K](X) & \mathcal{N}[K+1](X) \arrow[l, "DD"'] \end{tikzcd} \qquad \begin{tikzcd} X^K & X^{K+1} \arrow[l, "PD"'] \\ \mathcal{N}[K](X) \arrow[u, "arr"] & \mathcal{N}[K+1](X) \arrow[l, "DD"'] \arrow[u, "arr"'] \end{tikzcd}$$

Projection-delete thus commutes with the transposition channel $\mathit{transp} = \mathit{arr} \circledcirc \mathit{acc}$ *from* (2.86)*, in:*

$$\mathit{transp} \circledcirc PD \;=\; PD \circledcirc \mathit{transp}.$$

Proof. For $\vec{x} = (x_1, \ldots, x_{K+1}) \in X^{K+1}$ with $acc(\vec{x}) = \psi$ we have:

$$\begin{aligned} (acc \circledcirc PD)(\vec{x}) &= \sum_{1 \leq i \leq K+1} \frac{1}{K+1} \Big| acc(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{K+1}) \Big\rangle \\ &= \sum_{1 \leq i \leq K+1} \frac{1}{K+1} \Big| \psi - 1|x_i\rangle \Big\rangle \\ &= \sum_{y \in supp(\varphi)} \frac{\psi(y)}{K+1} \Big| \psi - 1|y\rangle \Big\rangle \;=\; DD(\psi) \;=\; (DD \circledcirc acc)(\vec{x}). \end{aligned}$$

Similarly, for $\psi \in \mathcal{N}[K+1](X)$,

$$\begin{aligned} (arr \circledcirc DD)(\psi) &= \sum_{\vec{x} \in X^K} \sum_{y \in supp(\psi)} arr(\psi - 1|y\rangle)(\vec{x}) \cdot \frac{\psi(y)}{K+1} \big|\vec{x}\rangle \\ &= \sum_{y \in supp(\psi)} \sum_{\vec{x} \in acc^{-1}(\psi - 1|y\rangle)} \frac{1}{(\psi - 1|y\rangle)} \cdot \frac{\psi(y)}{K+1} \big|\vec{x}\rangle \\ &= \sum_{y \in supp(\psi)} \sum_{\vec{x} \in acc^{-1}(\psi - 1|y\rangle)} \frac{(\psi(y)-1)! \cdot \prod_{z \neq y} \psi(z)!}{K!} \cdot \frac{\psi(y)}{K+1} \big|\vec{x}\rangle \\ &= \sum_{y \in supp(\psi)} \sum_{\vec{x} \in acc^{-1}(\psi - 1|y\rangle)} \frac{\psi!}{(K+1)!} \big|\vec{x}\rangle \\ &= \sum_{\vec{z} \in acc^{-1}(\psi)} \sum_{1 \leq i \leq K} \frac{1}{(\psi)} \cdot \frac{1}{K+1} \big|z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_K\rangle \\ &= \sum_{\vec{x} \in X^K} \sum_{\vec{z} \in X^{K+1}} arr(\psi)(\vec{z}) \cdot PD(\vec{z})(\vec{x}) \big|\vec{x}\rangle \;=\; (PD \circledcirc arr)(\psi). \end{aligned}$$

We now use commutation of these two rectangles in:

$$\mathit{transp} \circledcirc PD \;=\; arr \circledcirc acc \circledcirc PD \;=\; arr \circledcirc DD \circledcirc acc \;=\; PD \circledcirc arr \circledcirc acc \;=\; PD \circledcirc \mathit{transp}. \qquad \square$$

We turn to draw-add and follow the same pattern, with separate definitions for sequences and for multisets. Projection-addition is now more subtle: for a sequence $(x_1, \ldots, x_K)$ we can consecutively add each element x_i. A separate question is where to insert this x_i in the sequence. There are $K+1$ places, so we include them all. An illustration is given after the definition.

Definition 3.2.3 Let X be a set and let $K \geq 1$.

1 The projection-add channel $PA : X^K \rightsquigarrow X^{K+1}$ is:

$$PA(x_1, \ldots, x_K) \coloneqq \sum_{1 \leq i \leq K} \sum_{1 \leq j \leq K+1} \frac{1}{K(K+1)} \big| x_1, \ldots, x_{j-1}, x_i, x_j, \ldots, x_K \rangle. \tag{3.9}$$

2 The draw-add channel $DA : \mathcal{N}[K](X) \rightsquigarrow \mathcal{N}[K+1](X)$ is defined on a multiset $\varphi \in \mathcal{N}[K](X)$ as:

$$DA(\varphi) \coloneqq \sum_{x \in supp(\varphi)} \frac{\varphi(x)}{K} \Big| \varphi + 1|x\rangle \Big\rangle = \sum_{x \in supp(\varphi)} Flrn(\varphi)(x) \Big| \varphi + 1|x\rangle \Big\rangle. \tag{3.10}$$

For instance,

$$\begin{aligned} PA(a,b,c) &= \tfrac{1}{6}\big|a,a,b,c\rangle + \tfrac{1}{12}\big|a,b,a,c\rangle + \tfrac{1}{6}\big|a,b,b,c\rangle + \tfrac{1}{12}\big|a,b,c,a\rangle + \tfrac{1}{12}\big|a,b,c,b\rangle \\ &\quad + \tfrac{1}{6}\big|a,b,c,c\rangle + \tfrac{1}{12}\big|a,c,b,c\rangle + \tfrac{1}{12}\big|b,a,b,c\rangle + \tfrac{1}{12}\big|c,a,b,c\rangle. \end{aligned}$$

Lemma 3.2.4 *The projection-add channel commutes with transposition:*

$$transp \circ PA = PA \circ transp.$$

In addition, both projection- and draw-add maps commute with accumulation and aggregation:

$$\begin{tikzcd} X^K \arrow[r, "PA"] \arrow[d, "acc"'] & X^{K+1} \arrow[d, "acc"] \\ \mathcal{N}[K](X) \arrow[r, "DA"] & \mathcal{N}[K+1](X) \end{tikzcd} \qquad \begin{tikzcd} X^K \arrow[r, "PA"] & X^{K+1} \\ \mathcal{N}[K](X) \arrow[r, "DA"] \arrow[u, "arr"] & \mathcal{N}[K+1](X) \arrow[u, "arr"'] \end{tikzcd}$$

Proof. The equation $transp \circ PA = PA \circ transp$ clearly holds: a permutation of a sequence $x_1, \ldots, x_{j-1}, x_i, x_j, \ldots, x_K$ in (3.9) can be obtained from a project-add to a permutation of the original sequence $x_1, \ldots, x_K$.

For commutation with accumulation, let $\vec{x} \in X^K$,

$$\begin{aligned} (acc \circ PA)(\vec{x}) &= \sum_{1 \leq i \leq K} \sum_{1 \leq j \leq K+1} \frac{1}{K(K+1)} \Big| acc(x_1, \ldots, x_{j-1}, x_i, x_j, \ldots, x_K) \Big\rangle \\ &= \sum_{1 \leq i \leq K} \frac{1}{K} \Big| acc(x_1, \ldots, x_K) + 1|x_i\rangle \Big\rangle \\ &= \sum_{y \in supp(acc(\vec{x}))} \frac{acc(\vec{x})(y)}{K} \Big| acc(\vec{x}) + 1|y\rangle \Big\rangle = (DA \circ acc)(\vec{x}). \end{aligned}$$

For commutation with arrangement we use that acc is surjective, so that $f = g$ follows from $f \circ acc = g \circ acc$. Using this we are done with:

$$\begin{aligned} (PA \circ arr) \circ acc = PA \circ transp = transp \circ PA &= arr \circ acc \circ PA \\ &= arr \circ DA \circ acc \qquad \text{as just shown} \\ &= (arr \circ DA) \circ acc. \end{aligned}$$

❑

So far we have carefully spelled out the relation between projection operations on sequences and draw operations on multisets, via accumulation and arrangement. In the sequel we shall mostly work with the two draw operations DD and DA acting on multisets.

Remark 3.2.5 The delete and add channels are *not* each other's inverses, in a situation:

$$\begin{tikzcd} \mathcal{N}[K](X) \arrow[r, bend right, "DA"'] & \mathcal{N}[K+1](X) \arrow[l, bend right, "DD"'] \end{tikzcd}$$

For instance:

$$\begin{aligned} DD(3|a\rangle + 1|b\rangle) &= \tfrac{3}{4}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|3|a\rangle\big\rangle \\ DA(2|a\rangle + 1|b\rangle) &= \tfrac{2}{3}\big|3|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{3}\big|2|a\rangle + 2|b\rangle\big\rangle. \end{aligned}$$

In a next step we see that neither $DA \circ DD$ nor $DD \circ DA$ is the identity.

$$\begin{aligned} (DA \circ DD)(3|a\rangle + 1|b\rangle) &= DA_*\Big(\tfrac{3}{4}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|3|a\rangle\big\rangle\Big) \\ &= \tfrac{3}{4} \cdot \Big(\tfrac{2}{3}\big|3|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{3}\big|2|a\rangle + 2|b\rangle\big\rangle\Big) + \tfrac{1}{4} \cdot 1\big|4|a\rangle\big\rangle \\ &= \tfrac{1}{2}\big|3|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|2|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{4}\big|4|a\rangle\big\rangle \\ (DD \circ DA)(2|a\rangle + 1|b\rangle) &= DD_*\Big(\tfrac{2}{3}\big|3|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{3}\big|2|a\rangle + 2|b\rangle\big\rangle\Big) \\ &= \tfrac{2}{3} \cdot \Big(\tfrac{3}{4}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|3|a\rangle\big\rangle\Big) \\ &\qquad + \tfrac{1}{3} \cdot \Big(\tfrac{1}{2}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{2}\big|2|a\rangle + 1|b\rangle\big\rangle\Big) \\ &= \tfrac{1}{2}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{6}\big|3|a\rangle\big\rangle + \tfrac{1}{6}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{6}\big|2|a\rangle + 1|b\rangle\big\rangle. \end{aligned}$$

Since the drawing of a single ball from an urn takes the multiplicities into account, the urns after DD and DA have the same frequentist distribution as before, but only if we interpret 'after' as channel composition $\circ$. This is the content of the following basic result.

Theorem 3.2.6 *One has both:*

$$Flrn \circ DD = Flrn \qquad \textit{and} \qquad Flrn \circ DA = Flrn.$$

Equivalently, the following two triangles of channels commute.

$$\begin{tikzcd} \mathcal{N}[K](X) \arrow[rr, "DD"] \arrow[dr, "Flrn"'] & & \mathcal{N}[K+1](X) \arrow[dl, "Flrn"] \\ & X & \end{tikzcd} \qquad \begin{tikzcd} \mathcal{N}[K](X) \arrow[rr, "DA"] \arrow[dr, "Flrn"'] & & \mathcal{N}[K+1](X) \arrow[dl, "Flrn"] \\ & X & \end{tikzcd}$$

Proof. For the proof of commutation of draw-delete and frequentist learning we take $\psi \in \mathcal{M}[K+1](X)$ and $y \in X$. Then:

$$\begin{aligned} (Flrn \circ DD)(\psi)(y) &= \sum_{\varphi \in \mathcal{N}[K](X)} DD(\psi)(\varphi) \cdot Flrn(\varphi)(y) \\ &\overset{(3.8)}{=} \sum_{x \in supp(\psi)} \frac{\psi(x)}{K+1} \cdot Flrn(\psi - 1|x\rangle)(y) \\ &= \frac{\psi(y)}{K+1} \cdot \frac{\psi(y)-1}{K} + \sum_{x \neq y} \frac{\psi(x)}{K+1} \cdot \frac{\psi(y)}{K} \\ &= \frac{\psi(y)}{K(K+1)} \cdot \left(\psi(y) - 1 + \sum_{x \neq y} \psi(x)\right) \\ &= \frac{\psi(y)}{K(K+1)} \cdot \left(\left(\sum_{x} \psi(x)\right) - 1\right) \\ &= \frac{\psi(y)}{K(K+1)} \cdot ((K+1) - 1) = \frac{\psi(y)}{K+1} = Flrn(\psi)(y). \end{aligned}$$

Similarly, for $\varphi \in \mathcal{N}[K](X)$, where $K > 0$, and $y \in X$,

$$\begin{aligned}
(Flrn \circ DA)(\varphi)(y) &\overset{(3.10)}{=} \sum_{x \in supp(\varphi)} \frac{\varphi(x)}{K} \cdot Flrn(\varphi + 1|x\rangle)(y) \\
&= \frac{\varphi(y)}{K} \cdot \frac{\varphi(y)+1}{K+1} + \sum_{x \neq y} \frac{\varphi(x)}{K} \cdot \frac{\varphi(y)}{K+1} \\
&= \frac{\varphi(y)}{K(K+1)} \cdot \left(\varphi(y) + 1 + \sum_{x \neq y} \varphi(x) \right) \\
&= \frac{\varphi(y)}{K(K+1)} \cdot (K+1) = \frac{\varphi(y)}{K} = Flrn(\varphi)(y).
\end{aligned}$$

❑

When we iterate draw-and-delete and draw-and-add, as channels, we obtain distributions that strongly remind us of the hypergeometric and Pólya distribution. The iterations below describe not what is drawn from the urn — as in the hypergeometric and Pólya cases — but what is left in the urn after such draws. The full picture appears in Propositions 3.5.3 and 3.6.3. The iteration of draws leads to subtraction and addition of the draws φ from and to the urn.

Theorem 3.2.7 *Iterating $K \in \mathbb{N}$ times the draw-and-delete and draw-and-add channels yields channels:*

$$\begin{tikzcd}
\mathcal{N}[L](X) \arrow[r, bend right, "DA^K"'] & \mathcal{N}[L+K](X) \arrow[l, bend right, "DD^K"']
\end{tikzcd}$$

On $\upsilon \in \mathcal{N}[L+K](X)$ and $\psi \in \mathcal{N}[L](X)$ they satisfy:

$$DD^K(\upsilon) = \sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{L+K}{K}} |\upsilon - \varphi\rangle \qquad DA^K(\psi) = \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} |\psi + \varphi\rangle. \tag{3.11}$$

The probabilities in these expressions add up to one by Lemma 1.8.2 and Proposition 1.8.6.

Proof. We use induction on $K \in \mathbb{N}$. In both cases the only option for a draw φ in $\mathcal{N}[0](X)$ is the empty multiset $\mathbf{0}$, so that the sums in (3.11) are equal to $1|\upsilon\rangle$ and $1|\psi\rangle$, and thus equal to the zero-th iteration $DD^0(\upsilon) := 1|\upsilon\rangle$ and $DA^0(\psi) := 1|\psi\rangle$.

For the induction steps we make extensive use of the equations in Exercise 1.8.8. In those cases we shall put 'E' above the equation. We start with the induction step for draw-delete. For

$\upsilon \in \mathcal{N}[L+K+1](X)$,

$$\begin{aligned}
DD^{K+1}(\upsilon) &= DD^K{}_*(DD(\upsilon)) \\
&= DD^K{}_*\Bigl(\sum_{x\in supp(\upsilon)} \frac{\upsilon(x)}{L+K+1}\,\bigl|\upsilon - 1|x\rangle\bigr\rangle\Bigr) \\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{x\in supp(\upsilon)} \frac{\upsilon(x)}{L+K+1}\cdot DD^K(\upsilon - 1|x\rangle)(\varphi)\,\bigl|\varphi\bigr\rangle \\
&\overset{\text{(IH)}}{=} \sum_{x\in supp(\upsilon)}\ \sum_{\varphi\leq_K \upsilon-1|x\rangle} \frac{\upsilon(x)}{L+K+1}\cdot\frac{\binom{\upsilon-1|x\rangle}{\varphi}}{\binom{L+K}{K}}\,\bigl|(\upsilon - 1|x\rangle) - \varphi\bigr\rangle \\
&\overset{\text{(E)}}{=} \sum_{x\in supp(\upsilon)}\ \sum_{\varphi\leq_K \upsilon-1|x\rangle} \frac{\varphi(x)+1}{K+1}\cdot\frac{\binom{\upsilon}{\varphi+1|x\rangle}}{\binom{L+K+1}{K+1}}\,\bigl|\upsilon - (\varphi + 1|x\rangle)\bigr\rangle \\
&= \sum_{\varphi\leq_{K+1}\upsilon}\ \sum_{x\in supp(\upsilon)} \frac{\varphi(x)}{K+1}\cdot\frac{\binom{\upsilon}{\varphi}}{\binom{L+K+1}{K+1}}\,\bigl|\upsilon - \varphi\bigr\rangle \\
&= \sum_{\varphi\leq_{K+1}\upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{L+K+1}{K+1}}\,\bigl|\upsilon - \varphi\bigr\rangle, \qquad \text{since } \|\varphi\| = K+1.
\end{aligned}$$

We reason basically in the same way for draw-add, and now also use Exercise 1.8.12. For $\psi \in \mathcal{N}[L](X)$,

$$\begin{aligned}
DA^{K+1}(\psi) &= DA^K{}_*\Bigl(\sum_{x\in supp(\psi)} \frac{\psi(x)}{L}\,\bigl|\psi + 1|x\rangle\bigr\rangle\Bigr) \\
&\overset{\text{(IH)}}{=} \sum_{x\in supp(\psi)}\ \sum_{\varphi\in\mathcal{N}[K](supp(\psi+1|x\rangle))} \frac{\psi(x)}{L}\cdot\frac{\left(\!\binom{\psi+1|x\rangle}{\varphi}\!\right)}{\left(\!\binom{L+1}{K}\!\right)}\,\bigl|(\psi+1|x\rangle) + \varphi\bigr\rangle \\
&\overset{\text{(E)}}{=} \sum_{x\in supp(\psi)}\ \sum_{\varphi\in\mathcal{N}[K](supp(\psi))} \frac{\varphi(x)+1}{K+1}\cdot\frac{\left(\!\binom{\psi}{\varphi+1|x\rangle}\!\right)}{\left(\!\binom{L}{K+1}\!\right)}\,\bigl|\psi + (\varphi+1|x\rangle)\bigr\rangle \\
&= \sum_{\varphi\in\mathcal{N}[K+1](supp(\psi))}\ \sum_{x\in supp(\psi)} \frac{\varphi(x)}{K+1}\cdot\frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K+1}\!\right)}\,\bigl|\psi + \varphi\bigr\rangle \\
&= \sum_{\varphi\in\mathcal{N}[K+1](supp(\psi))} \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K+1}\!\right)}\,\bigl|\psi + \varphi\bigr\rangle.
\end{aligned}$$

❑

The draw-delete map interacts neatly with the multizip channel from the previous section. This fails for draw-add, see Exercise 3.2.7.

Lemma 3.2.8 *For sets X, Y and a number K, the following diagram commutes.*

$$\begin{tikzcd}
\mathcal{N}[K+1](X)\times\mathcal{N}[K+1](Y) \arrow[r, "DD\otimes DD"] \arrow[d, "mzip"'] & \mathcal{N}[K](X)\times\mathcal{N}[K](Y) \arrow[d, "mzip"] \\
\mathcal{N}[K+1](X\times Y) \arrow[r, "DD"] & \mathcal{N}[K](X\times Y)
\end{tikzcd} \tag{3.12}$$

Proof. We use the formulation of *mzip* from Proposition 3.1.7 in terms of couplings. We will

uses the following equation, for multisets $\varphi \in \mathcal{N}[K+1](X)$, $\psi \in \mathcal{N}[K+1](Y)$.

$$\begin{aligned} &\Big\{\chi \in dcpl^{-1}\big(\varphi - 1|x\rangle,\ \psi - 1|y\rangle\big) \,\Big|\, x \in supp(\varphi),\ y \in supp(\psi)\Big\} \\ &= \Big\{\rho - 1|x,y\rangle \,\Big|\, \rho \in dcpl^{-1}(\varphi,\psi),\ (x,y) \in supp(\rho)\Big\}. \end{aligned} \tag{$*$}$$

For the inclusion ($\subseteq$) we observe that if $\chi \in dcpl^{-1}(\varphi - 1|x\rangle, \psi - 1|y\rangle)$, then $\mathcal{M}(\pi_1)(\chi) = \varphi - 1|x\rangle$ and $\mathcal{M}(\pi_2)(\chi) = \psi - 1|y\rangle$. We take $\rho := \chi + 1|x,y\rangle$, so that $(x,y) \in supp(\rho)$ obviously holds. Moreover,

$$\begin{aligned} \mathcal{M}(\pi_1)(\rho) \;=\; \mathcal{M}(\pi_1)(\chi + 1|x,y\rangle) \;&=\; \mathcal{M}(\pi_1)(\chi) + \mathcal{M}(\pi_1)(1|x,y\rangle) \qquad \text{by Lemma 1.6.3 (3)} \\ &=\; \varphi - 1|x\rangle + 1|x\rangle \;=\; \varphi. \end{aligned}$$

Similarly, $\mathcal{M}(\pi_2)(\rho) = \psi$, so that $\rho \in dcpl^{-1}(\varphi,\psi)$. For the inclusion ($\supseteq$), let $\rho \in dcpl^{-1}(\varphi,\psi)$ have $(x,y) \in supp(\rho)$. Then, clearly, $x \in supp(\varphi)$ and $y \in supp(\psi)$. Moreover, it is easy to see that $\chi := \rho - 1|x,y\rangle$ has $\varphi - 1|x\rangle$ and $\psi - 1|y\rangle$ as marginals.

In order to prove commutation of the rectangle (3.12), let $\varphi \in \mathcal{N}[K+1](X)$, $\psi \in \mathcal{N}[K+1](Y)$.

$$\begin{aligned} &mzip_*(DD(\varphi) \otimes DD(\psi)) \\ &= \sum_{x \in supp(\varphi)} \sum_{y \in supp(\psi)} \frac{\varphi(x)\cdot\psi(y)}{(K+1)^2} \cdot mzip(\varphi - 1|x\rangle, \psi - 1|y\rangle) \\ &\overset{(3.6)}{=} \sum_{x \in supp(\varphi)} \sum_{y \in supp(\psi)} \sum_{\chi \in dcpl^{-1}(\varphi - 1|x\rangle, \psi - 1|y\rangle)} \frac{\varphi(x)\cdot\psi(y)\cdot(\chi)}{(K+1)^2\cdot(\varphi - 1|x\rangle)\cdot(\psi - 1|y\rangle)} \,|\chi\rangle \\ &\overset{(*)}{=} \sum_{\rho \in dcpl^{-1}(\varphi,\psi)} \sum_{(x,y) \in supp(\rho)} \frac{(\rho - 1|x,y\rangle)}{(\varphi)\cdot(\psi)} \,|\rho - 1|x,y\rangle\rangle \\ &= \sum_{\rho \in dcpl^{-1}(\varphi,\psi)} \sum_{(x,y) \in supp(\rho)} \frac{\rho(x,y)}{K+1} \cdot \frac{(\rho)}{(\varphi)\cdot(\psi)} \,|\rho - 1|x,y\rangle\rangle \\ &= \sum_{\rho \in \mathcal{M}[K+1](X\times Y)} \sum_{\chi \in \mathcal{M}[K](X\times Y)} DD(\rho)(\chi)\cdot mzip(\varphi,\psi)(\rho)\,|\chi\rangle \;=\; DD_*(mzip(\varphi,\psi)). \end{aligned}$$ ❑

The draw-delete and draw-add channels both form an infinite chain. We are especially interested in the draw-delete chain, of the form:

$$\mathbf{1} \cong \mathcal{N}[0](X) \xleftarrow{DD} \mathcal{N}[1](X) \xleftarrow{DD} \mathcal{N}[2](X) \xleftarrow{DD} \mathcal{N}[3](X) \xleftarrow{DD} \cdots \tag{3.13}$$

A collection of distributions $\big(\sigma_K \in \mathcal{D}(\mathcal{N}[K](X))\big)_{K\in\mathbb{N}}$ will be called a *draw-delete cone*, or simply a *DD-cone*, if $DD_*(\sigma_{K+1}) = \sigma_K$ for each K. Such cones play an important role in De Finetti's work on exchangeability, see de Finetti (1930) and *e.g.* Jacobs and Staton (2020); Staton and Summer (2022). We shall see a small part of this below, but first we show that multinomial and Pólya distributions form draw-delete cones. Hypergeometric distributions do not give rise to such cones since a hypergeometric urn becomes empty after a certain number of draws so that draws of arbitrary sizes are impossible.

As an aside, the name 'cone' as used here comes from the category-theoretic concept of a limit. It is not related to a 'cone' in the sense of a module over $\mathbb{R}_{\geq 0}$, as used in Lemma 1.6.3 (2).

Proposition 3.2.9 *Multinomial and Pólya distributions form draw-delete cones: for an arbi-*

trary set X and for all $K \in \mathbb{N}$, the following two triangles commute.

$$\begin{tikzcd} \mathcal{N}[K](X) & & \mathcal{N}[K+1](X) \arrow[ll, "DD"'] \\ & \mathcal{D}(X) \arrow[ul, "mn[K]"] \arrow[ur, "mn[K+1]"'] & \end{tikzcd} \qquad \begin{tikzcd} \mathcal{N}[K](X) & & \mathcal{N}[K+1](X) \arrow[ll, "DD"'] \\ & \mathcal{N}_*(X) \arrow[ul, "pl[K]"] \arrow[ur, "pl[K+1]"'] & \end{tikzcd}$$

Proof. For the triangle on the left, let $\omega \in \mathcal{D}(X)$ be an arbitrary distribution.

$$\begin{aligned}
(DD \circ mn[K+1])(\omega) &= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[K+1](X)} mn[K+1](\omega)(\psi) \cdot DD[K](\psi)(\varphi) \,|\varphi\rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in X} mn[K+1](\omega)(\varphi + 1|x\rangle) \cdot \frac{\varphi(x)+1}{K+1} \,|\varphi\rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in X} \frac{(K+1)!}{\prod_y (\varphi + 1|x\rangle)(y)!} \cdot \prod_y \omega(y)^{(\varphi+1|x\rangle)(y)} \cdot \frac{\varphi(x)+1}{K+1} \,|\varphi\rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in X} \frac{K!}{\prod_y \varphi(y)!} \cdot \omega(x) \cdot \prod_y \omega(y)^{\varphi(y)} \,|\varphi\rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \left(\sum_{x \in X} \omega(x) \right) \cdot \boldsymbol{(\varphi)} \cdot \prod_y \omega(y)^{\varphi(y)} \,|\varphi\rangle \;=\; mn[K](\omega).
\end{aligned}$$

In the proof below, of commutation of the above triangle on the right, the marked equation $\overset{(*)}{=}$ indicates the use of Exercises 1.8.12 and 1.8.8. For urn $\upsilon \in \mathcal{N}[L](X)$ of size $L > 0$, and for $\varphi \in \mathcal{N}[K](X)$,

$$\begin{aligned}
(DD \circ pl[K+1])(\psi) &= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\chi \in \mathcal{N}[K+1](supp(\psi))} DD(\chi)(\varphi) \cdot pl[K+1](\psi)(\chi) \,|\varphi\rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} \sum_{x \in supp(\psi)} \frac{\varphi(x)+1}{K+1} \cdot pl[K+1](\psi)(\varphi + 1|x\rangle) \,|\varphi\rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} \sum_{x \in supp(\psi)} \frac{\varphi(x)+1}{K+1} \cdot \frac{\left(\!\binom{\psi}{\varphi+1|x\rangle}\!\right)}{\left(\!\binom{L}{K+1}\!\right)} \,|\varphi\rangle \\
&\overset{(*)}{=} \sum_{\varphi \in \mathcal{N}[K](supp(\psi))} \sum_{x \in supp(\psi)} \frac{\psi(x)+\varphi(x)}{L+K} \cdot \frac{\left(\!\binom{\psi}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} \,|\varphi\rangle \;=\; pl[K](\psi). \qquad \square
\end{aligned}$$

In the literature (transposition-)exchangeable distributions play an important role, notably in relation to De Finetti's result, see de Finetti (1930); Jaynes (1982); Bernardo and Smith (2000). The result below justifies why such (chains of) exchangeable distributions on sequences can be replaced by chains of distributions on multisets. Thus, by using distributions on multisets we sidestep the whole issue of (transposition-)exchangeability.

Proposition 3.2.10 *Let X be an arbitrary set. There is a bijective correspondence between:*

$$\frac{\textit{transposition-exchangeable } \left(\tau_K \in \mathcal{D}(X^K)\right)_{K \in \mathbb{N}} \textit{ with } \mathcal{D}(\pi)(\tau_{K+1}) = \tau_K}{\textit{draw-delete cones } \left(\sigma_K \in \mathcal{D}(\mathcal{N}[K](X))\right)_{K \in \mathbb{N}}}$$

Here, $\pi \colon X^{K+1} \to X^K$ is the function that projects away the last element.

Proof. ($\Downarrow$) Let $\tau_K \in \mathcal{D}(X^K)$, for $K \in \mathbb{N}$, be a collection of transposition-exchangeable distributions such that τ_{K+1} marginalises to τ_K, for each $K \in \mathbb{N}$, that is, $\mathcal{D}(\pi)(\tau_{K+1}) = \tau_K$. We first show that:

$$PD_*(\tau_{K+1}) = \tau_K. \tag{$*$}$$

Indeed, the assumptions about the $\tau's$ give, for each sequence $\vec{x} = (x_1, \ldots, x_K) \in X^K$,

$$\begin{aligned} PD_*(\tau_{K+1})(\vec{x}) &\overset{(3.7)}{=} \sum_{y \in X} \sum_{1 \leq i \leq K+1} \frac{1}{K+1} \cdot \tau_{K+1}(x_1, \ldots, x_{i-1}, y, x_i, \ldots, x_K) \\ &= \sum_{y \in X} \sum_{1 \leq i \leq K+1} \frac{1}{K+1} \cdot \tau_{K+1}(x_1, \ldots, x_K, y) \\ &= \sum_{y \in X} \tau_{K+1}(x_1, \ldots, x_K, y) = \mathcal{D}(\pi)(\tau_{K+1})(x_1, \ldots, x_K) = \tau_K(\vec{x}) \end{aligned}$$

The transposition-exchangeability of τ_K means that $transp_*(\tau_K) = \tau_K = id_*(\tau_K)$ by Lemma 2.11.4 (3). When we view τ_K as a channel $\mathbf{1} \rightsquigarrow X^K$, we can apply the equaliser property of arrangement in Proposition 2.11.5 (2). It gives us a unique map $\mathbf{1} \rightsquigarrow \mathcal{N}[K](X)$, that is a distribution $\overline{\tau}_K \in \mathcal{D}(\mathcal{N}[K](X))$, with $arr_*(\overline{\tau}_K) = \tau_K$. Then:

$$\mathcal{D}(acc)(\tau_K) = acc_*(arr_*(\overline{\tau}_K)) = (acc \circ arr)_*(\overline{\tau}_K) \overset{(2.39)}{=} \overline{\tau}_K. \tag{$**$}$$

We can now show that these $\overline{\tau}_K$ form a draw-delete cone:

$$\begin{aligned} DD_*(\overline{\tau}_{K+1}) \overset{(**)}{=} (DD \circ acc)_*(\tau_{K+1}) &= (acc \circ PD)_*(\tau_{K+1}) \qquad \text{by Lemma 3.2.2} \\ &= acc_*(PD_*(\tau_{K+1})) \\ &\overset{(*)}{=} acc_*(\tau_K) = \mathcal{D}(acc)(\tau_K) \overset{(**)}{=} \overline{\tau}_K. \end{aligned}$$

($\Uparrow$) Now assume we have a draw-delete cone $\sigma_K \in \mathcal{D}(\mathcal{N}[K](X))$, for $K \in \mathbb{N}$. We define $\overline{\sigma}_K := arr_*(\sigma_K) \in \mathcal{D}(X^K)$. This distribution $\overline{\sigma}_K$ is exchangeable by Lemma 2.11.4 (3):

$$transp_*(\overline{\sigma}_K) = (arr \circ acc \circ arr)_*(\sigma_K) = arr_*(\sigma_K) = \overline{\sigma}_K.$$

Moreover, Exercise 3.2.3 gives us:

$$\mathcal{D}(\pi)(\overline{\sigma}_{K+1}) = (\pi \circ arr)_*(\sigma_{K+1}) = (arr \circ DD)_*(\sigma_{K+1}) = arr_*(\sigma_K) = \overline{\sigma}_K.$$

Thus, these distributions $\overline{\sigma}_K$ satisfy the properties above the double lines in the above proposition.

What remains to be shown is that these transformations ($\Downarrow$) and ($\Uparrow$) are each other's inverses. First,

$$\overline{\overline{\tau}}_K = arr_*(\overline{\tau}_K) = \tau_K, \quad \text{by definition of } \overline{\tau}_K.$$

Next $\overline{\overline{\sigma}}_K$ is by definition the unique distribution with $arr_*(\overline{\overline{\sigma}}_K) = \overline{\sigma}_K$. But also, again by definition, $arr_*(\sigma_K) = \overline{\sigma}_K$. Hence $\overline{\overline{\sigma}}_K = \sigma_K$. ❑

Exercises

3.2.1 Check that we could have introduced DD via the coequaliser property of accumulation (from Proposition 2.11.3 (1)) as:

$$\begin{tikzcd}
X^{K+1} \arrow[rr, two heads, "acc"] \arrow[dr, "PD"'] & & \mathcal{N}[K+1](X) \arrow[d, dashed, "DD"] \\
& \mathcal{D}(X^K) \arrow[r, "\mathcal{D}(acc)"'] & \mathcal{D}(\mathcal{N}[K](X))
\end{tikzcd}$$

3.2.2 Calculate yourself an instance of the commutation of project- and draw-add with arrangement in Lemma 3.2.4, in the diagram on the right: take the multiset $\varphi = 1|a\rangle + 2|b\rangle$ and show that:

$$\begin{aligned}
PA_*(arr(\varphi)) &= \tfrac{1}{18}\big|a,a,b,b\rangle + \tfrac{1}{18}\big|a,b,a,b\rangle + \tfrac{1}{18}\big|a,b,b,a\rangle + \tfrac{1}{6}\big|a,b,b,b\rangle + \tfrac{1}{18}\big|b,a,a,b\rangle \\
&\quad + \tfrac{1}{18}\big|b,a,b,a\rangle + \tfrac{1}{6}\big|b,a,b,b\rangle + \tfrac{1}{18}\big|b,b,a,a\rangle + \tfrac{1}{6}\big|b,b,a,b\rangle + \tfrac{1}{6}\big|b,b,b,a\rangle \\
&= arr_*(DA(\varphi)).
\end{aligned}$$

3.2.3 Write $\pi\colon X^{K+1} \to X^K$ for the (standard, non-probabilistic) projection function that throws away the last element.

1 Show that the following diagram of channels commutes.

$$\begin{tikzcd}
\mathcal{N}[K](X) \arrow[d, "arr"'] & \mathcal{N}[K+1](X) \arrow[l, "DD"'] \arrow[d, "arr"] \\
X^K & X^{X+1} \arrow[l, "\pi"']
\end{tikzcd}$$

2 Check that accumulation does *not* commute with draw-and-delete, as in:

$$\begin{tikzcd}
\mathcal{N}[K](X) & \mathcal{N}[K+1](X) \arrow[l, "DD"'] \arrow[d, phantom, "\neq"] \\
X^K \arrow[u, "acc"] & X^{X+1} \arrow[l, "\pi"] \arrow[u, "acc"']
\end{tikzcd}$$

3.2.4 Show that the draw-and-delete and draw-and-add maps DD and DA are natural, from $\mathcal{N}[K+1]$ to $\mathcal{D} \circ \mathcal{N}[K]$, and from $\mathcal{N}[K]$ to $\mathcal{D} \circ \mathcal{N}[K+1]$.

3.2.5 Recall the uniform distributions $unif_{\mathcal{N}[K](X)} \in \mathcal{D}(\mathcal{N}[K](X))$ from Exercise 2.4.5, where the set X is finite.

1 Prove that these uniform distributions form a draw-delete cone, that is, satisfy:

$$DD_*(unif_{\mathcal{N}[K+1](X)}) = unif_{\mathcal{N}[K](X)}.$$

One can calculate this by expanding the definitions, or by combining Proposition 3.2.9 with Exercise 3.6.1.

2 Check that the draw-add maps do *not* preserve uniform distributions. Show for instance that for $X = \{a, b\}$,

$$\begin{aligned}
DA_*(unif_{\mathcal{N}[3](X)}) &= \tfrac{1}{4}\big|4|a\rangle\rangle + \tfrac{1}{6}\big|3|a\rangle + 1|b\rangle\rangle + \tfrac{1}{6}\big|2|a\rangle + 2|b\rangle\rangle \\
&\quad + \tfrac{1}{6}\big|1|a\rangle + 3|b\rangle\rangle + \tfrac{1}{4}\big|4|b\rangle\rangle.
\end{aligned}$$

3.2.6 Check that multinomial channels do *not* commute with DA: take for instance $\omega = \frac{1}{3}|a\rangle + \frac{2}{3}|b\rangle$ and show first that:

$$\begin{aligned} mn[2](\omega) &= \tfrac{1}{9}\big|2|a\rangle\rangle + \tfrac{4}{9}\big|1|a\rangle + 1|b\rangle\rangle + \tfrac{4}{9}\big|2|b\rangle\rangle \\ mn[3](\omega) &= \tfrac{1}{27}\big|3|a\rangle\rangle + \tfrac{2}{9}\big|2|a\rangle + 1|b\rangle\rangle + \tfrac{4}{9}\big|1|a\rangle + 2|b\rangle\rangle + \tfrac{8}{27}\big|3|b\rangle\rangle. \end{aligned}$$

Now show that $DA_*(mn[2](\omega))$ differs from the last distribution.
(Pólya channels do commute with DA, see Exercise 3.5.1.)

3.2.7 Draw-add DA does *not* commute with *mzip*, like DD in Lemma 3.2.8. Take for instance $\varphi = 1|a\rangle + 1|b\rangle$ and $\psi = 1|0\rangle + 1|0\rangle$ and check that:

$$\begin{aligned} DA_*(mzip(\varphi,\psi)) &= \tfrac{1}{4}\Big|2|a,1\rangle + 1|b,0\rangle\Big\rangle + \tfrac{1}{4}\Big|1|a,1\rangle + 2|b,0\rangle\Big\rangle \\ &\quad + \tfrac{1}{4}\Big|2|a,0\rangle + 1|b,1\rangle\Big\rangle + \tfrac{1}{4}\Big|1|a,0\rangle + 2|b,1\rangle\Big\rangle \\ mzip_*(DA(\varphi)\otimes DA(\psi)) &= \tfrac{1}{6}\Big|1|a,0\rangle + 1|a,1\rangle + 1|b,0\rangle\Big\rangle + \tfrac{1}{12}\Big|2|a,1\rangle + 1|b,0\rangle\Big\rangle \\ &\quad + \tfrac{1}{12}\Big|1|a,1\rangle + 2|b,0\rangle\Big\rangle + \tfrac{1}{12}\Big|2|a,0\rangle + 1|b,1\rangle\Big\rangle \\ &\quad + \tfrac{1}{6}\Big|1|a,0\rangle + 1|a,1\rangle + 1|b,1\rangle\Big\rangle \\ &\quad + \tfrac{1}{6}\Big|1|a,0\rangle + 1|b,0\rangle + 1|b,1\rangle\Big\rangle \\ &\quad + \tfrac{1}{6}\Big|1|a,1\rangle + 1|b,0\rangle + 1|b,1\rangle\Big\rangle + \tfrac{1}{12}\Big|1|a,0\rangle + 2|b,1\rangle\Big\rangle. \end{aligned}$$

3.2.8 Let X be a finite set, say with n elements, of the form $X = \{x_1, \ldots, x_n\}$. Define for $K \geq 1$,

$$\sigma_K \coloneqq \sum_{1\leq i\leq n} \frac{1}{n}\big|K|x_i\rangle\rangle \;\in\; \mathcal{D}(\mathcal{N}[K](X)).$$

Thus:

$$\begin{aligned} \sigma_1 &= \tfrac{1}{n}\big|1|x_1\rangle\rangle + \cdots + \tfrac{1}{n}\big|1|x_n\rangle\rangle \\ \sigma_2 &= \tfrac{1}{n}\big|2|x_1\rangle\rangle + \cdots + \tfrac{1}{n}\big|2|x_n\rangle\rangle, \qquad \textit{etc.} \end{aligned}$$

Show that these σ_K form a cone, both for draw-delete and for draw-add:

$$DD_*(\sigma_{K+1}) = \sigma_K \qquad \text{and} \qquad DA_*(\sigma_K) = \sigma_{K+1}.$$

Thus, the whole sequence $(\sigma_K)_{K\in\mathbb{N}_{>0}}$ can be generated from $\sigma_1 = unif_{\mathcal{N}[1](X)}$ by repeated application of DA.

3.2.9 Let X be a finite set and $K \in \mathbb{N}$ be an arbitrary number. Show that for $\sigma \in \mathcal{D}(\mathcal{N}[K+1](X))$ and $\varphi \in \mathcal{N}[K](X)$ one has:

$$\frac{DD_*(\sigma)(\varphi)}{(\varphi)} = \sum_{x\in X} \frac{\sigma(\varphi+1|x\rangle)}{(\varphi+1|x\rangle)}\Big|\varphi + 1|x\rangle\Big\rangle.$$

3.2.10 Show that the probabilistic projection channel PD makes the following diagram commute.

$$\begin{tikzcd} \mathcal{D}(X)^{K+1} \arrow[r, "\otimes", "\circ"'] \arrow[d, "PD"', "\circ"] & X^{K+1} \arrow[d, "PD", "\circ"'] \\ \mathcal{D}(X)^{K} \arrow[r, "\otimes", "\circ"'] & X^{K} \end{tikzcd}$$

3.2.11 Recall the uniform projection channel $unpr[K]\colon X^K \multimap X$ from Exercise 2.4.9. Prove that the following rectangles commute.

$$\begin{tikzcd} X^K \arrow[rr, "PD"] \arrow[dr, "unpr[K]"'] & & X^{K+1} \arrow[dl, "unpr[K+1]"] \\ & X & \end{tikzcd} \qquad \begin{tikzcd} X^K \arrow[rr, "PA"] \arrow[dr, "unpr[K]"'] & & X^{K+1} \arrow[dl, "unpr[K+1]"] \\ & X & \end{tikzcd}$$

3.2.12 We define a draw-store-delete channel DSD of the form:

$$\mathcal{N}[K+1](X) \xrightarrow{DSD} \mathcal{D}\big(X \times \mathcal{N}[K](X)\big)$$

as:

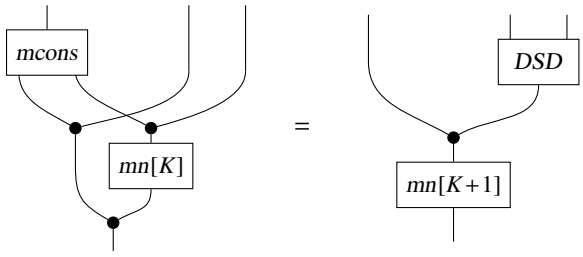
$$DSD(\psi) := \sum_{x \in supp(\psi)} \frac{\psi(x)}{\|\psi\|} \Big| x, \psi - 1|x\rangle \Big\rangle .$$

Thus, frequentist learning is the first marginal of DSD and draw-delete is the second.

We also define a multiset-cons function $mcons\colon X \times \mathcal{N}[K](X) \to \mathcal{N}[K+1](X)$ as $mcons(x, \varphi) = \varphi + 1|x\rangle$.

1 Check that $mcons \circ DSD = id$, so that $DSD \circ mcons$ is split idempotent in **Chan**.

2 Prove the equality below, expressed in terms of string diagrams.



(3.14)

We shall see an extended version in Theorem 3.5.4.

3.2.13 Check that the following diagram does *not* commute.

$$\begin{tikzcd} \mathcal{N}[K+1](X) \times \mathcal{N}[L+1](X) \arrow[r, "DD \otimes DD"] \arrow[d, "+"'] \arrow[dr, phantom, "\neq"] & \mathcal{N}[K](X) \times \mathcal{N}[L](X) \arrow[d, "+"] \\ \mathcal{N}[K+L+2](X) \arrow[r, "DD \circ DD"'] & \mathcal{N}[K+L](X) \end{tikzcd}$$

3.3 The multinomial channel

We have introduced multinomial distributions earlier, in Definition 2.6.6, and since then we have seen them a few times. In this section we take a systematic look at their properties. We standardly describe multinomial distributions in 'multivariate' form, for multiple colours. The binomial distribution is then a special 'bivariate' case, for two colours only, see Lemma 2.6.8. We sometimes allow ourselves the short hands 'multinomials' and 'binomials' for 'multinomial distributions' and 'binomial distributions'.

First we recall that the multinomial channel has the form:

$$\mathcal{D}(X) \xrightarrow{mn[K]} \mathcal{D}\big(\mathcal{N}[K](X)\big) \qquad \text{written as} \qquad \mathcal{D}(X) \xrightarrow{mn[K]} \mathcal{N}[K](X).$$

The number $K \in \mathbb{N}$ represents the number of balls that are drawn, without replacement, from a distribution $\omega \in \mathcal{D}(X)$ that represents an abstract urn. The resulting distribution $mn[K](\omega)$ assigns probabilities to K-ball draws $\varphi \in \mathcal{N}[K](X)$. There is no bound on K, since the multinomial drawing mode involves returning drawn balls to the urn, so that the urn does not become empty.

For many elementary properties of multionomial distributions it is essential that we use the channel form $mn[K] \colon \mathcal{D}(X) \rightsquigarrow \mathcal{N}[K](X)$ with associated sequential and parallel composition $\circ\!\!\!\circ$ and $\otimes$ for channels. This will become clear in the diagrams below, with circles on the shafts of arrows.

For convenience we repeat essential formulations of the multinomial distribution, for a distribution $\omega \in \mathcal{D}(X)$.

$$\begin{aligned} mn[K](\omega) &\overset{(2.54)}{=} \sum_{\varphi \in \mathcal{N}[K](X)} \binom{\varphi}{} \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \,|\varphi\rangle \overset{(2.55)}{=} \mathcal{D}(acc)\big(iid[K](\omega)\big) \\ &= \mathcal{D}(acc)\big(\omega^K\big) \\ &= \mathcal{D}(acc)\big(\underbrace{\omega \otimes \cdots \otimes \omega}_{K \text{ times}}\big). \end{aligned}$$

Thus, the probability assigned to a draw / multiset $\varphi \in \mathcal{N}[K](X)$ is the sum over all probabilities $\omega^K(\vec{x}) = \prod_i \omega(x_i)$ for sequence $\vec{x} \in X^K$ that accumulate to φ, that is, with $acc(\vec{x}) = \varphi$. Recall, from Proposition 1.7.2, that there are $\binom{\varphi}{}$ many such sequences $\vec{x}$ with $acc(\vec{x}) = \varphi$, where $\binom{\varphi}{} = \frac{\|\varphi\|!}{\varphi!} = \frac{K!}{\prod_x \varphi(x)!}$.

We start with a result from Jacobs (2023b) (in this form) that captures a basic relationship between distributions and multisets in a situation that we shall later characterise as: accumulation is a 'sufficient statistic' for iid, see Section 7.7. One can recognise this result in (Bishop, 2006, §2.2), where it is mentioned that a particular sum — amounting to accumulation — is sufficient as statistic for iid. But a formalisation as in Diagram (3.15) does not occur in the traditional literature.

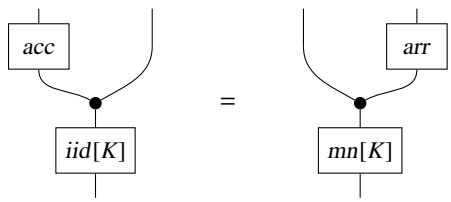
In the first point below we emphasise the crucial property of accumulation that is used. This property occurs in more general form in Proposition 7.7.5 later on. It is formulated in terms of kernel relations $\ker(f) \subseteq X \times X$, for a function $f \colon X \to Y$, where $\ker(f) = \{(x, x') \mid f(x) = f(x')\}$.

Theorem 3.3.1 *Fix a set X and a number $K \in \mathbb{N}$.*

1 *For any distribution $\omega \in \mathcal{D}(X)$, one has $\ker(acc) \subseteq \ker(iid[K](\omega))$. More explicitly, for all sequences $\vec{x}, \vec{y} \in X^K$,*

$$acc(\vec{x}) = acc(\vec{y}) \implies \omega^K(\vec{x}) = \omega^K(\vec{y}).$$

2 *The accumulation function $acc \colon X^K \to \mathcal{N}[K](X)$ and its probabilistic 'arrangement' inverse $arr = acc^{\smile} \colon \mathcal{N}[K](X) \rightsquigarrow X^K$ from (2.38) satisfy the equation between the following two string diagrams.*



(3.15)

By discarding the wires on the left on both sides we get as consequence the equation:

$$iid[K] = arr \circ\!\!\!\circ mn[K]. \tag{3.16}$$

Proof. We choose an arbitrary distribution $\omega \in \mathcal{D}(X)$.

1 Let $\vec{x}, \vec{y} \in X^K$ be given with $acc(\vec{x}) = \varphi = acc(\vec{y})$. This means that the sequences $\vec{x}$ and $\vec{y}$ contain the same elements, with the same multiplicities, but possibly in different order. These orders are irrelevant for multiplications. Thus, as also noted at the end of the proof of Theorem 2.6.7,

$$\omega^K(\vec{x}) \;=\; \prod\nolimits_i \omega(x_i) \;=\; \prod_{z \in X} \omega(z)^{\varphi(z)} \;=\; \prod\nolimits_i \omega(y_i) \;=\; \omega^K(\vec{y}).$$

2 We need to prove an equation (3.15) between tuple channels, see Definition 2.4.4 (3). We use Proposition 1.7.2 and Theorem 2.6.7:

$$\begin{aligned}
\langle acc, id \rangle_*(iid[K](\omega)) &= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in X^K} \langle acc \rangle(\vec{x})(\varphi) \cdot iid[K](\omega)(\vec{x}) \,\big|\, \varphi, \vec{x} \rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in acc^{-1}(\varphi)} \omega^K(\vec{x}) \,\big|\, \varphi, \vec{x} \rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in acc^{-1}(\varphi)} \prod_{x \in X} \omega(x)^{\varphi(x)} \,\big|\, \varphi, \vec{x} \rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in acc^{-1}(\varphi)} \frac{1}{(\varphi)} \cdot (\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \,\big|\, \varphi, \vec{x} \rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in X^K} arr(\varphi)(\vec{x}) \cdot mn[K](\omega)(\varphi) \,\big|\, \varphi, \vec{x} \rangle \\
&= \langle id, arr \rangle_*(mn[K](\omega)).
\end{aligned}$$

❑

Remark 3.3.2 In the above situation we see that the distributions $iid[K](\omega)$ and $mn[K](\omega)$ are related via accumulation, namely as $\mathcal{D}(acc)(iid[K](\omega)) = mn[K](\omega)$, where accumulation satisfies the kernel inclusion property of Theorem 3.3.1 (1). There is a similar situation where this kernel property fails — and where, as a consequence, we do not have a sufficient statistics situation as in (3.15).

Recall from Proposition 2.7.10 that $\mathcal{D}(sum)(mn[K](\omega)) = K \cdot \omega$, for the sum function $sum \colon \mathcal{N}[K](M) \to M$, where M is a commutative monoid, say $M = \mathbb{N}$. In this situation we do not have that $sum(\varphi) = sum(\psi)$ implies $mn[K](\omega)(\varphi) = mn[K](\omega)(\psi)$. Indeed, knowing that two multisets have equal sums, does not ensure that they are equally likely as draws from an urn ω. Thus, *sum* is not a sufficient statistics for multinomial distributions.

We continue with the theory of multinomial distributions. A basic fact is that frequentist learning after a multinomial is the identity. We will soon expand on this, but first we need an auxiliary result.

Lemma 3.3.3 *Fix a distribution $\omega \in \mathcal{D}(X)$ and a number K. For each $y \in X$,*

$$\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \;=\; K \cdot \omega(y).$$

As a result, the validity $mn[K](\omega) \models Flrn(-)(y)$ equals $\omega(y)$.

Proof. The equation holds for $K = 0$, since then $\varphi(y) = 0$. Hence we may assume $K > 0$. Then:

$$\begin{aligned}
&\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \\
&= \sum_{\varphi \in \mathcal{N}[K](X),\, \varphi(y) \neq 0} \varphi(y) \cdot \frac{K!}{\prod_x \varphi(x)!} \cdot \prod_x \omega(x)^{\varphi(x)} \\
&= \sum_{\varphi \in \mathcal{N}[K](X),\, \varphi(y) \neq 0} K \cdot \frac{(K-1)!}{\prod_x (\varphi - 1|y\rangle)(x)!} \cdot \omega(y) \cdot \prod_x \omega(x)^{(\varphi - 1|y\rangle)(x)} \\
&= K \cdot \omega(y) \cdot \sum_{\varphi \in \mathcal{N}[K-1](X)} (\varphi) \cdot \prod_x \omega(x)^{\varphi(x)} \\
&= K \cdot \omega(y) \cdot \sum_{\varphi \in \mathcal{N}[K-1](X)} mn[K-1](\omega)(\varphi) \;=\; K \cdot \omega(y).
\end{aligned}$$

❑

Theorem 3.3.4 *Frequentist learning from a multinomial gives the original distribution:*

$$Flrn_*(mn[K](\omega)) \;=\; \omega. \tag{3.17}$$

This means that the following diagram of channels commutes.

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "mn[K]"] \arrow[dr, bend right, "sam"'] & \mathcal{N}[K](X) \arrow[d, "Flrn"] \\
& X
\end{tikzcd} \tag{3.18}$$

The channel $\mathrm{sam} \colon \mathcal{D}(X) \rightsquigarrow X$ *is the identity function* $\mathcal{D}(X) \to \mathcal{D}(X)$*, considered as channel; it is called the sample channel.*

For a channel $c \colon X \rightsquigarrow Y$*, the channel* $\mathrm{mn}[K](c) := \mathrm{mn}[K] \circ c$ *satisfies:*

$$\begin{tikzcd}
X \arrow[r, "mn[K](c)"] \arrow[dr, bend right, "c"'] & \mathcal{N}[K](Y) \arrow[d, "Flrn"] \\
& Y
\end{tikzcd} \tag{3.19}$$

In Subsection 2.2.1 we have described sampling from a distribution $\omega \in \mathcal{D}(X)$ in programming. For a parameter K, it produces a multiset $\varphi \in \mathcal{N}[K](X)$ such that $Flrn(\varphi)$ is close to ω. Operationally, such sampling involves a random number generator and (mostly) produces different φ's in different runs. Hence such sampling is a randomised algorithm, not a mathematical function.

The multinomial distribution $mn[K](\omega)$ is the mathematical counterpart of such sampling. It does not produce a single sample, but *all* samples of a particular size K, with corresponding probabilities, say written as $mn[K](\omega) = \sum_i r_i|\varphi_i\rangle$. The above theorem says that frequentist learning of all these samples, while taking the probabilities r_i into account, yields the original distribution ω, that is:

$$Flrn_*(mn[K](\omega)) \;=\; \sum\nolimits_i r_i \cdot Flrn(\varphi_i) \;=\; \omega.$$

We read this as a correctness result for multinomial sampling. We shall come accross similar results, see especially Theorem 5.6.1. Diagam (3.19) also occurs in the axiomatic setting of (Jacobs, 2021c, Thm. 2).

Proof. By Lemma 3.3.3:

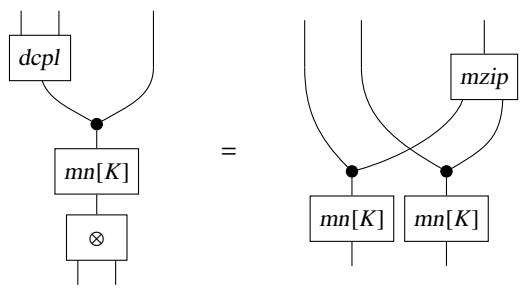
$$\begin{aligned}(Flrn \circ mn[K])(\omega)(y) &= Flrn_*\Big(mn[K](\omega)\Big)(y) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot Flrn(\varphi)(y) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \frac{\varphi(y)}{K} \;=\; K \cdot \omega(y) \cdot \frac{1}{K} \;=\; \omega(y).\end{aligned}$$

An alternative proof is suggested in Exercise 3.3.5. Using the equation that we just proved:

$$(Flrn \circ mn[K](c))(x) \;=\; Flrn_*\Big(mn[K](c(x))\Big) \;=\; c(x). \qquad \square$$

We turn to the multizip channel *mzip*, from Section 3.1, in relation to multinomial distributions. Our main result involves the decouple function $dcpl = \langle \mathcal{N}(\pi_1), \mathcal{N}(\pi_2)\rangle \colon \mathcal{N}[K](X \times Y) \to \mathcal{N}[K](X) \times \mathcal{N}[K](Y)$, consisting of a pair of marginalisations for multisets, see (3.4). We have already seen, in Proposition 3.1.3 (3), that it forms a split idempotent with multizip (in **Chan**). It is in fact a 'sufficient statistic' as expressed below.

Theorem 3.3.5 *Decouple and multizip are related to multinomial distributions via the following equation between string diagrams.*



(3.20)

Proof. We take distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ as inputs on the two incoming wires at the

bottom, on both sides of the equation in (3.20).

$$
\begin{aligned}
&\langle dcpl, id\rangle_*(mn[K](\omega\otimes\rho))\\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{\psi\in\mathcal{N}[K](Y)}\ \sum_{\chi\in\mathcal{N}[K](X\times Y)} \langle dcpl\rangle(\chi)(\varphi,\psi)\cdot mn[K](\omega\otimes\rho)(\chi)\,\Big|\,\varphi,\psi,\chi\Big\rangle\\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{\psi\in\mathcal{N}[K](Y)}\ \sum_{\chi\in dcpl^{-1}(\varphi,\psi)} (\chi)\cdot\prod_{x\in X,\,y\in Y}((\omega\otimes\rho)(x,y))^{\chi(x,y)}\,\Big|\,\varphi,\psi,\chi\Big\rangle\\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{\psi\in\mathcal{N}[K](Y)}\ \sum_{\chi\in dcpl^{-1}(\varphi,\psi)} (\chi)\cdot\prod_{x\in X,\,y\in Y}\omega(x)^{\chi(x,y)}\cdot\prod_{x\in X,\,y\in Y}\rho(y)^{\chi(x,y)}\,\Big|\,\varphi,\psi,\chi\Big\rangle\\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{\psi\in\mathcal{N}[K](Y)}\ \sum_{\chi\in dcpl^{-1}(\varphi,\psi)} (\chi)\cdot\prod_{x\in X}\omega(x)^{\sum_{y\in Y}\chi(x,y)}\cdot\prod_{y\in Y}\rho(y)^{\sum_{x\in X}\chi(x,y)}\,\Big|\,\varphi,\psi,\chi\Big\rangle\\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{\psi\in\mathcal{N}[K](Y)}\ \sum_{\chi\in dcpl^{-1}(\varphi,\psi)} (\chi)\cdot\prod_{x\in X}\omega(x)^{\varphi(x)}\cdot\prod_{y\in Y}\rho(y)^{\psi(y)}\,\Big|\,\varphi,\psi,\chi\Big\rangle\\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{\psi\in\mathcal{N}[K](Y)}\ \sum_{\chi\in dcpl^{-1}(\varphi,\psi)} \frac{(\chi)}{(\varphi)\cdot(\psi)}\cdot mn[K](\omega)(\varphi)\cdot mn[K](\rho)(\psi)\,\Big|\,\varphi,\psi,\chi\Big\rangle\\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{\psi\in\mathcal{N}[K](Y)}\ \sum_{\chi\in\mathcal{N}[K](X\times Y)} mzip(\varphi,\psi)(\chi)\cdot\Big(mn[K](\omega)\otimes mn[K](\rho)\Big)(\varphi,\psi)\,\Big|\,\varphi,\psi,\chi\Big\rangle\\
&= \langle id, mzip\rangle_*(mn[K](\omega)\otimes mn[K](\rho)).
\end{aligned}
$$

❑

Equation (3.20) contains a lot of information. When we discard (that is, marginalise out) the wires at the top-left, on both sides of the equation in (3.20), we get the result below. It demonstrates that the multinomial channel interacts nicely with multizip and tensors (of distributions). Later in Theorey 3.8.11 a categorical interpretation will be given: multinomial is a monoidal transformation.

Corollary 3.3.6 *Multinomial channels commute with tensor and multizip:*

$$mzip_*\Big(mn[K](\omega)\otimes mn[K](\rho)\Big) = mn[K](\omega\otimes\rho). \tag{3.21}$$

Diagrammatically this amounts to:

$$
\begin{tikzcd}
\mathcal{D}(X)\times\mathcal{D}(Y) \arrow[r, "mn[K]\otimes mn[K]"] \arrow[d, "\otimes"'] & \mathcal{N}[K](X)\times\mathcal{N}[K](Y) \arrow[d, "mzip"] \\
\mathcal{D}(X\times Y) \arrow[r, "mn[K]"] & \mathcal{N}[K](X\times Y)
\end{tikzcd}
$$

Proof. As mentioned, Equation (3.21) follows from (3.20), via marginalisation on the left. Alternatively, one can prove the equation directly via (categorical) diagram-chasing: commutation of the outer diagram below follows from the commuting subparts of the diagram, which arise by

unfolding the definition of *mzip* in (3.1), below on the right.

$$\begin{tikzcd}
\mathcal{D}(X) \times \mathcal{D}(Y) \arrow[r, "mn[K]\otimes mn[K]"] \arrow[dd, "\otimes"'] \arrow[drr, bend right, "iid\otimes iid"'] & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[d, "arr\otimes arr"] \arrow[ddd, bend left, "mzip"] \\
& X^K \times Y^K \arrow[d, "zip"] \\
& (X \times Y)^K \arrow[d, "acc"] \\
\mathcal{D}(X \times Y) \arrow[r, "mn[K]"] \arrow[uur, bend left, "iid"] & \mathcal{N}[K](X \times Y)
\end{tikzcd}$$

The three subdiagrams commute by (3.16), twice, and by Lemma 2.3.5 (3). ❑

Next we look at multinomials and flattening of multisets and distributions. We first obtain an 'average' or 'mean' result for multinomials (see also Definition 4.2.1). We can describe it via a flatten map $flat\colon \mathcal{M}(\mathcal{M}(X)) \to \mathcal{M}(X)$, using inclusions $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ and $\mathcal{N}[K](X) \hookrightarrow \mathcal{M}(X)$. This is another consequence of Lemma 3.3.3.

Proposition 3.3.7 *Fix a distribution $\omega \in \mathcal{D}(X)$ and a number K. Each natural multiset $\varphi \in \mathcal{N}[K](X)$ can be regarded as an element of the set of all multisets $\mathcal{M}(X)$. Similarly, ω can be understood as an element of $\mathcal{M}(X)$. With these inclusions in mind one has:*

$$flat\Big(mn[K](\omega)\Big) \;=\; \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi \;=\; K \cdot \omega \;\in\; \mathcal{M}(X).$$

The first equation expands the definition of 'flat'. The second equation is the relevant new fact.

Proof. Since:

$$\begin{aligned}
\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi \;&=\; \sum_{x \in X} \left(\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x) \right) \big| x \rangle \\
&=\; \sum_{x \in X} K \cdot \omega(x) \,\big| x \rangle \qquad \text{by Lemma 3.3.3} \\
&=\; K \cdot \sum_{x \in X} \omega(x) \,\big| x \rangle \;=\; K \cdot \omega.
\end{aligned}$$

❑

Since a multinomial $mn[K](\omega)$ is a distribution we can use it as an abstract urn, not containing single balls, but containing multisets of balls (draws). Hence we can draw from $mn[K](\omega)$ as well, giving a distribution on draws of draws. The first item belows shows that this can also be done with a single multinomial. The second item does not have such a clear explanation; it will be useful later on.

Theorem 3.3.8 *Fix a set X and numbers $K, L \in \mathbb{N}$.*

1 *Multinomial channels compose, with a bit of help of the (fixed-size) flatten operation for multisets, as in:*

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "mn[K]"] \arrow[drr, bend right, "mn[L\cdot K]"'] & \mathcal{D}(\mathcal{M}[K](X)) \arrow[r, "mn[L]"] & \mathcal{D}\Big(\mathcal{M}[L](\mathcal{M}[K](X))\Big) \arrow[d, "\mathcal{D}(flat)"] \\
& & \mathcal{D}\Big(\mathcal{M}[L\cdot K](X)\Big)
\end{tikzcd}$$

2 *The following diagram commutes:*

$$\begin{tikzcd}
\mathcal{D}(\mathcal{N}[K](X)) \arrow[r, "mn[L]"] \arrow[dr, bend right, "\mathcal{D}(Flrn)"'] & \mathcal{D}\big(\mathcal{M}[L](\mathcal{M}[K](X))\big) \arrow[r, "\mathcal{D}(flat)"] & \mathcal{D}\big(\mathcal{M}[L\cdot K](X)\big) \arrow[dl, bend left, "\mathcal{D}(Flrn)"] \\
& \mathcal{D}(\mathcal{D}(X)) \arrow[dr, bend right, "flat"'] & \mathcal{D}(\mathcal{D}(X)) \arrow[dl, bend left, "flat"] \\
& & \mathcal{D}(X)
\end{tikzcd}$$

Proof. 1 We use the following diagram chase:

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "iid[K]"] \arrow[ddr, bend right, "iid[K\cdot L]"'] & \mathcal{D}(X^K) \arrow[r, "\mathcal{D}(acc)"] \arrow[d, "iid[L]"'] & \mathcal{D}(\mathcal{N}[K](X)) \arrow[d, "iid[L]"] \\
& \mathcal{D}((X^K)^L) \arrow[r, "\mathcal{D}(acc^L)"] \arrow[d, "\mathcal{D}(flat^{\mathcal{L}})"'] & \mathcal{D}(\mathcal{N}[K](X)^L) \arrow[d, "\mathcal{D}(acc)"] \\
& \mathcal{D}(X^{K\cdot L}) \arrow[dr, bend right, "\mathcal{D}(acc)"'] & \mathcal{D}(\mathcal{N}[L](\mathcal{N}[K](X))) \arrow[d, "\mathcal{D}(flat^{\mathcal{N}})"] \\
& & \mathcal{D}(\mathcal{N}[K\cdot L](X))
\end{tikzcd}$$

The triangle on the left commutes by Exercise 2.3.14. The rectangle on the upper-right commutes by naturality of $iid[L]$, see Lemma 2.3.5 (1). The lower-right subdiagram commutes because $acc\colon \mathcal{L} \Rightarrow \mathcal{N}$ is a map of monads, see Exercise 1.12.7.

2 For $\omega \in \mathcal{D}(\mathcal{N}[K](X))$ and $x \in X$ we have:

$$\begin{aligned}
&(flat \circ \mathcal{D}(Flrn) \circ \mathcal{D}(flat) \circ mn[L])(\omega)(x) \\
&= (Flrn \circ flat)_*(mn[L](\omega))(x) \\
&= \sum_{\Psi \in \mathcal{N}[L](\mathcal{N}[K](X))} Flrn(flat(\Psi))(x) \cdot mn[L](\omega)(\Psi) \\
&= \sum_{\Psi \in \mathcal{N}[L](\mathcal{N}[K](X))} \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\Psi(\varphi)\cdot\varphi(x)}{L\cdot K} \cdot mn[L](\omega)(\Psi) \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\varphi(x)}{L\cdot K} \cdot \sum_{\Psi \in \mathcal{N}[L](\mathcal{N}[K](X))} \Psi(\varphi)\cdot mn[L](\omega)(\Psi) \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\varphi(x)}{L\cdot K} \cdot L \cdot \omega(\varphi) \qquad \text{by Lemma 3.3.3} \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} Flrn(\varphi)(x)\cdot\omega(\varphi) \;=\; Flrn_*(\omega)(x) \;=\; (flat \circ \mathcal{D}(Flrn))(\omega)(x).
\end{aligned}$$

❑

The relation between multinomial and Poisson distributions is worth making explicit. We shall do so in the next theorem in a string diagrammatic manner. The essentials are (re)formulated in terms of updating and disintegration, see Example 7.3.2 (**??**).

The formulation below involves some bookkeeping, especially to turn a sequence of numbers into frequencies (*i.e.* multiplicities) of a multiset. We write this operation as *Freq*, both for

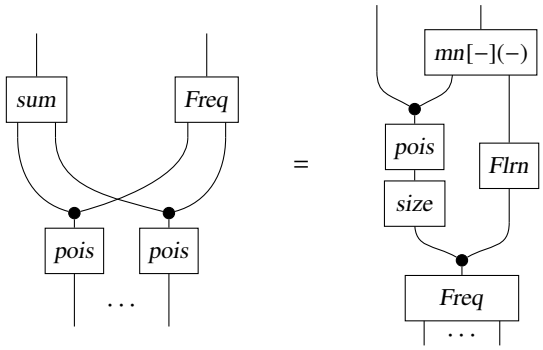
arbitrary and for natural multisets, of the form:

$$\begin{array}{rcl} (\mathbb{R}_{\geq 0})^m & \xrightarrow[\cong]{Freq} & \mathcal{M}(\boldsymbol{m}) \\ \vec{r} & \longmapsto & \sum_{0 \leq i < m} r_i | i \rangle \end{array} \qquad \begin{array}{rcl} \mathbb{N}^m & \xrightarrow[\cong]{Freq} & \mathcal{N}(\boldsymbol{m}) \\ \vec{n} & \longmapsto & \sum_{0 \leq i < m} n_i | i \rangle \end{array} \tag{3.22}$$

We have seen these isomorphisms in Exercise 1.6.5.

The next result relates multinomial and Poisson distributions. This relationship is known, see *e.g.* (Johnson et al., 1997, Ch. 35, §2), but not as in this diagram.

Theorem 3.3.9 *Poisson and multinomial channels are related as expressed by the following equation between channels of type* $(\mathbb{R}_{>0})^m \to \mathcal{D}_\infty(\mathbb{N} \times \mathcal{N}(\boldsymbol{m}))$.



(3.23)

The 'bookkeeping' in this result happens especially on the right-hand-side, where the first three channels (*Freq*, *size* and *Flrn*) are essentially only reordering the input, see also Lemma 2.2.4.

By marginalising out the wires on the right in (3.23) one obtains that Poisson distributions commute with sums, as we have seen in Proposition 2.7.9.

Proof. Let $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_{>0}$ be given, for which we abbreviate $\lambda := \sum_i \lambda_i \in \mathbb{R}_{>0}$. We start reasoning from the left-hand-side in (3.23).

$$\begin{aligned} & \sum_{\vec{n} \in \mathbb{N}^m} \prod_{0 \leq i < m} pois[\lambda_i](n_i) \Big| sum(\vec{n}), Freq(\vec{n}) \Big\rangle \\ &= \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{m})} \prod_{0 \leq i < m} e^{-\lambda_i} \cdot \frac{\lambda_i^{\varphi(i)}}{\varphi(i)!} \Big| K, \varphi \Big\rangle \\ &= \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{m})} e^{-\lambda} \cdot \frac{\prod_{0 \leq i < m} \lambda_i^{\varphi(i)}}{\varphi ! \!\!\!\circ} \Big| K, \varphi \Big\rangle \\ &= \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{m})} e^{-\lambda} \cdot \frac{\lambda^K}{K!} \cdot \frac{K!}{\varphi ! \!\!\!\circ} \cdot \prod_{0 \leq i < m} \left(\frac{\lambda_i}{\lambda}\right)^{\varphi(i)} \Big| K, \varphi \Big\rangle \\ &= \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{m})} pois[\lambda](K) \cdot mn[K]\Big(\textstyle\sum_i \frac{\lambda_i}{\lambda} | i \rangle \Big)(\varphi) \Big| K, \varphi \Big\rangle . \end{aligned}$$

The latter expression is the right-hand-side of (3.23). ❑

Since $mn[K](\omega)$ is a distribution, the sum over *all* draws $\sum_\varphi mn[K](\omega)(\varphi)$ equals one. But what if we restrict this sum to draws φ of certain colours only, that is, with $supp(\varphi) \subseteq S$, for a proper subset $S \subseteq supp(\omega)$? And what if we then let the size of these draws K go to infinity? The

result below describes what happens: the sum of the probabilities of such restricted draws goes to zero, as the size goes to infinity. It turns out that the same behaviour exists in the hypergeometric and Pólya cases, see Proposition 3.6.7.

Proposition 3.3.10 *Let $\omega \in \mathcal{D}(X)$ be given with a proper, non-empty subset $S \subseteq supp(\omega)$, so $S \neq \emptyset$ and $S \neq supp(\omega)$. For $K \in \mathbb{N}$, write:*

$$M_K := \sum_{\varphi \in \mathcal{M}[K](S)} mn[K](\omega)(\varphi).$$

Then $M_K > M_{K+1}$ and $\lim_{K\to\infty} M_K = 0$.

Proof. Write $r := \sum_{x\in S} \omega(x)$ in:

$$M_K = \sum_{\varphi \in \mathcal{M}[K](S)} mn[K](\omega)(\varphi) = \sum_{\varphi \in \mathcal{M}[K](S)} \binom{}{\varphi} \cdot \prod_{x\in S} \omega(x)^{\varphi(x)} \overset{(1.43)}{=} r^K.$$

Since $0 < r < 1$ we get $M_K = r^K > r^{K+1} = M_{K+1}$ and $\lim_{K\to\infty} M_K = \lim_{K\to\infty} r^K = 0$. ❑

Exercises

3.3.1 Let's throw a fair dice 12 times. What is the probability that each number appears twice? Show that it is $\frac{12!}{72^6}$.

3.3.2 Check that the multiset coefficient distribution from (2.5) is an instance of the multinomial distribution: for a non-empty finite set X, with uniform distribution $\upsilon \in \mathcal{D}(X)$, and for a number K,

$$mc[X, K] = mn[K](\upsilon).$$

3.3.3 Use Exercise 1.8.6 to show that for $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[L](X)$ one has:

$$mn[K+L](\omega)(\varphi + \psi) = \frac{\binom{K+L}{K}}{\binom{\varphi+\psi}{\varphi}} \cdot mn[K](\omega)(\varphi) \cdot mn[L](\omega)(\psi),$$

and if φ, ψ have disjoint support, then:

$$mn[K+L](\omega)(\varphi + \psi) = \binom{K+L}{K} \cdot mn[K](\omega)(\varphi) \cdot mn[L](\omega)(\psi).$$

3.3.4 Let X be finite set. A family of distributions $\sigma_K \in \mathcal{D}(\mathcal{N}[K](X))$ is called *multinomial* when:

- $\sum_{x\in X} \sigma_1(1|x\rangle) = 1$;
- $\sigma_K(K|x\rangle) = \sigma_1(1|x\rangle)^K$, for $x \in X$ and $K \in \mathbb{N}$;
- for multisets $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[L](X)$ with disjoint support, $\sigma_{K+L}(\varphi + \psi) = \binom{K+L}{K} \cdot \sigma_K(\varphi) \cdot \sigma_L(\psi)$.

1 Verify that for each state $\omega \in \mathcal{D}(X)$, the multinomial distributions $mn[K](\omega)$, for $K \in \mathbb{N}$, form a multinomial family.

2 In the other direction, let $\sigma_K \in \mathcal{D}(\mathcal{N}[K](X))$ form a multinomial family. Show that $\sigma_K = mn[K](\omega)$ for a distribution $\omega \in \mathcal{D}(X)$.

3.3.5 Give an alternative proof of the equation $Flrn_*(mn[K](\omega)) = \omega$ in Theorem 3.3.4 via Theorem 2.6.7 and Exercise 2.4.9.

3.3.6 Let X be a set with a subset $A \subseteq X$. Consider the sum-evaluation function sum_A, defined as $sum_A(f) = \sum_{x \in A} f(x)$. We use it as map with the following two types.

$$\mathcal{D}(X) \xrightarrow{sum_A} [0,1] \qquad\qquad \mathcal{N}[K](X) \xrightarrow{sum_A} \{0,1,\ldots,K\}$$

Show that the following diagram commutes.

$$\begin{tikzcd} \mathcal{D}(X) \arrow[r, "sum_A"] \arrow[d, "mn[K]"'] & {[0,1]} \arrow[d, "bn[K]"] \\ \mathcal{D}(\mathcal{N}[K](X)) \arrow[r, "\mathcal{D}(sum_A)"] & \mathcal{D}(\{0,1,\ldots,K\}) \end{tikzcd}$$

3.3.7 The aim of this exercise is to prove recurrence relations for multinomials: for each $\omega \in \mathcal{D}(X)$ and $\varphi \in \mathcal{N}[K](X)$ with $K > 0$ one has:

$$mn[K](\omega)(\varphi) \;=\; \sum_{x \in supp(\varphi)} \omega(x) \cdot mn[K-1](\omega)(\varphi - 1|x\rangle).$$

Here are two possible avenues:

1 use the recurrence relations (1.41) for multiset coefficients $\binom{\varphi}{}$;
2 show first:

$$\omega(x) \cdot mn[K-1](\omega)(\varphi - 1|x\rangle) \;=\; Flrn(\varphi)(x) \cdot mn[K](\omega)(\varphi).$$

Follow-up both avenues.

3.3.8 Prove the following items, in the style of Lemma 3.3.3, for a distribution $\omega \in \mathcal{D}(X)$ and a number $K > 1$.

1 For two elements $y \neq z$ in X,

$$\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \cdot \varphi(z) \;=\; K \cdot (K-1) \cdot \omega(y) \cdot \omega(z).$$

2 For a single element $y \in X$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y) \cdot (\varphi(y) - 1) \;=\; K \cdot (K-1) \cdot \omega(y)^2.$$

3 Again, for a single element $y \in X$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(y)^2 \;=\; K \cdot (K-1) \cdot \omega(y)^2 + K \cdot \omega(y).$$

Hint: Write $a^2 = a \cdot (a-1) + a$ and then use the previous item and Lemma 3.3.3.

3.3.9 Fix a distribution $\omega \in \mathcal{D}(X)$ and a number K. We write τ as abbreviation for the following distribution.

$$\tau \;=\; \langle Flrn, Flrn\rangle_*\Big(mn[K](\omega)\Big) \;\in\; \mathcal{D}(X \times X).$$

1 Show that τ is coupling of ω and ω, that is:

$$\tau[1,0] \;=\; \omega \;=\; \tau[0,1].$$

2 Use the previous exercise to prove that τ is a convex sum of the form:

$$\tau \;=\; \tfrac{K-1}{K} \cdot (\omega \otimes \omega) + \tfrac{1}{K} \cdot \mathcal{D}(\Delta)(\omega).$$

3.3.10 1 Generalise the closure of binomials under convolution from Proposition 2.7.6 (2) to multinomials: for $K, L \in \mathbb{N}$,

$$\begin{tikzcd} \mathcal{D}(X) \arrow[r, "\Delta"] \arrow[drrr, bend right, "mn[K+L]"'] & \mathcal{D}(X) \times \mathcal{D}(X) \arrow[r, "mn[K]\otimes mn[L]"] & \mathcal{N}[K](X) \times \mathcal{N}[L](X) \arrow[d, "+"] \\ & & \mathcal{N}[K+L](X) \end{tikzcd}$$

One can use Exercise 1.7.15 and Lemma 2.7.8 for a proof. One can also wait for Theorem 3.5.4, since this closure under convolution is a consequence.
Apart from that, conclude that for a fixed distribution ω the channel $mn[-](\omega)\colon \mathbb{N} \to \mathcal{D}(\mathcal{N}(X))$ is a map of (commutative) monoids — like in Proposition 2.7.9 and Exercise 2.7.14.

2 Generalise these convolutions to K-ary form, as $K \cdot mn[1](\omega) = mn[K](\omega)$, and conclude that K-sized draws can be reduced to parallel single draws, as in:

$$\begin{tikzcd} \mathcal{D}(X) \arrow[r, "\Delta"] \arrow[drr, bend right, "mn[K]"'] & \mathcal{D}(X)^K \arrow[r, "mn[1]^K"] & \mathcal{N}[1](X)^K \arrow[d, "+"] \\ & & \mathcal{N}[K](X) \end{tikzcd}$$

Notice that this is essentially Theorem 2.6.7, via the isomorphism $\mathcal{N}[1](X) \cong X$.

3 Generalise the convolution property in the first point also in the following manner: for distributions $\omega, \rho \in \mathcal{D}(X)$ and numbers $K, L \geq 1$ one has:

$$\mathcal{D}(acc)\Big(\omega^K \otimes \rho^L\Big) \;=\; mn[K](\omega) + mn[L](\rho).$$

3.3.11 Show that for a natural multiset φ of size K one has:

$$flat\Big(mn[K](Flrn(\varphi))\Big) \;=\; \varphi.$$

3.3.12 Check that the following diagram does *not* commute, in general:

$$\begin{tikzcd} \mathcal{D}^2(X) \arrow[r, "\mathcal{D}(mn[K])"] \arrow[d, "flat"'] & \mathcal{D}^2(\mathcal{N}[K](X)) \arrow[d, "flat"] \\ \mathcal{D}(X) \arrow[r, "mn[K]"] & \mathcal{D}(\mathcal{N}[K](X)) \end{tikzcd}$$

Take for instance as distribution of distributions $\frac{3}{4}\Big|\frac{1}{3}|a\rangle + \frac{2}{3}|b\rangle\Big\rangle + \frac{1}{3}\Big|1|a\rangle\Big\rangle$, with $K = 2$.

3.3.13 Check that multinomial channels do *not* commute with tensors, as in:

$$\begin{tikzcd} \mathcal{D}(X) \times \mathcal{D}(Y) \arrow[r, "mn[K]\otimes mn[L]"] \arrow[d, "\otimes"'] \arrow[dr, phantom, "\neq"] & \mathcal{M}[K](X) \times \mathcal{M}[L](Y) \arrow[d, "\otimes"] \\ \mathcal{D}(X \times Y) \arrow[r, "mn[K\cdot L]"'] & \mathcal{M}[K\cdot L](X \times Y) \end{tikzcd}$$

3.4 Divergence and multinomials

Recall from Section 2.10 that KL-divergence tells how different two distributions are. This divergence plays a role in the context of multinomial distributions, especially when we wish to find a distribution that matches a number of data points, found in for instance in an experiment. The

data points can be organised as a multiset, say ψ. We can then try to find a distribution ω with minimal divergence $D_{KL}(Flrn(\psi), \omega)$. As we shall see, this corresponds to what is commonly called maximal likelihood. An obvious solution is to take $\omega = Flrn(\psi)$. In practice, ω is required to be a certain type of distribution, for instance binomial.

We start with an isometry-style result for multinomial distributions from Jacobs (2025c).

Proposition 3.4.1 *Let distributions $\omega, \omega' \in \mathcal{D}(X)$ be given with a draw-size number $K \in \mathbb{N}$. Then:*

$$\begin{aligned} D_{KL}\Big(mn[K](\omega),\, mn[K](\omega')\Big) &= K \cdot D_{KL}(\omega, \omega') \\ &= D_{KL}\Big(\mathcal{D}(Flrn)(mn[K](\omega)),\, \mathcal{D}(Flrn)(mn[K](\omega'))\Big). \end{aligned}$$

Proof. We may assume $supp(\omega) \subseteq supp(\omega')$ and write $supp(\omega') = \{x_1, \ldots, x_n\} \subseteq X$. We use Lemma 3.3.3 for the first equation:

$$\begin{aligned} D_{KL}\Big(mn[K](\omega),\, mn[K](\omega')\Big) &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \ln\left(\frac{mn[K](\omega)(\varphi)}{mn[K](\omega')(\varphi)}\right) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \ln\left(\frac{\binom{\varphi} \cdot \prod_i \omega(x_i)^{\varphi(x_i)}}{\binom{\varphi} \cdot \prod_i \omega'(x_i)^{\varphi(x_i)}}\right) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \ln\left(\prod_i \left(\frac{\omega(x_i)}{\omega'(x_i)}\right)^{\varphi(x_i)}\right) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \sum_i \varphi(x_i) \cdot \ln\left(\frac{\omega(x_i)}{\omega'(x_i)}\right) \\ &= \sum_i \left(\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x_i)\right) \cdot \ln\left(\frac{\omega(x_i)}{\omega'(x_i)}\right) \\ &= \sum_i K \cdot \omega(x_i) \cdot \ln\left(\frac{\omega(x_i)}{\omega'(x_i)}\right) = K \cdot D_{KL}(\omega, \omega'). \end{aligned}$$

The second equation follows from the first one via Exercise 2.10.3, using that for a fixed draw-size K, the frequentist learning function $Flrn \colon \mathcal{N}[K](X) \to \mathcal{D}(X)$ is injective. ❑

We next describe a situation that is common in statistics and in machine learning: we have data and like to know which distribution best fits the data. This task is sometimes called density estimation. Commonly, the data is described as a list (or as a row vector), but in this book we describe it as a multiset, since the order of the elements does not matter. It is assumed that the same distribution is used for each of the data points, so that we are in a situation with independent and identically distributed elements. This is reflected in the first point below, in multiset style.

Theorem 3.4.2 *Let $\psi \in \mathcal{N}(X)$ be a fixed (natural) multiset of data points, over some set X, say of size $K := \|\psi\| > 0$. The distribution $\omega \in \mathcal{D}(X)$ that best matches ψ can be described equivalently as the ω with:*

1 *maximal likelihood* $\displaystyle\prod_{x \in supp(\psi)} \omega(x)^{\psi(x)}$;

2 *maximal multinomial probability* $mn[K](\omega)(\psi)$;

3 *miminal error* $\displaystyle -\ln\left(\prod_{x \in supp(\psi)} \omega(x)^{\psi(x)})\right)$;

4 *minimal divergence* $D_{KL}(Flrn(\psi), \omega)$.

Often, the distribution ω is required to be of a certain type, e.g. Poisson, indexed by the rate, but that is irrelevant for these equivalences.

We shall freely switch between these four equivalent formulations. But we often use the last one in formulating general results, since it is very intuitive, involving minimal divergence between the (normalised) data and the distribution that we seek. Notice that this fourth formulation is so natural because we use data as multisets, with the associated frequentist learning function *Flrn*. When the data is represented as a list or a vector $\ell \in \mathcal{L}(X)$, one can equivalently use $Flrn(acc(\ell))$, but this is more cumbersome and assumes an order on the data elements that is irrelevant.

The whole approach in this theorem fits what we identify as Jeffrey style validity (and updating) in Section 8.3 — in contrast to Pearl style. The above fourth formulation will play a key role in, for instance, Jeffrey's update rule, see Theorem 7.9.3, and in expectation maximisation, see Section 8.5. The name 'error' in the third point for the negative logarithm is common in machine learning, see *e.g.* Bishop and Bishop (2024).

A alternative formulation of Theorem 3.4.2 is in Exercise 3.4.1.

Proof. Theres is an equivalence between points (1) and (2) because, by definition:

$$mn[K](\omega)(\psi) = \binom{\psi}{} \cdot \prod_{x \in supp(\psi)} \omega(x)^{\psi(x)}.$$

The multiset coefficient $\binom{\psi}{}$ is constant, since the data ψ is fixed, so it does not play a role in maximisation.

The equivalence between points (1) and (3) is obtained from the fact that the (natural) logarithm preserves and reflects the order (see Exercise 1.4.2), so that maximising the product $\prod_x \omega(x)^{\psi(x)}$ from point (1) corresponds to maximising its logarithm, and corresponds to minimising the negation of this logarithm.

The equivalence between the last two points (3) and (4) follows easily by rewriting the divergence definition:

$$\begin{aligned} D_{KL}(Flrn(\psi), \omega) &= \sum_{x \in X} Flrn(\psi)(x) \cdot \ln\left(\frac{Flrn(\psi)(x)}{\omega(x)}\right) \\ &= \sum_{x \in X} Flrn(\psi)(x) \cdot \ln\big(Flrn(\psi)(x)\big) - \sum_{x \in X} Flrn(\psi)(x) \cdot \ln\big(\omega(x)\big) \\ &= -H(Flrn(\psi)) - \ln\left(\prod_{x \in supp(\psi)} \omega(x)^{\psi(x)}\right). \end{aligned}$$

The entropy $H(Flrn(\psi))$ is constant, since the data ψ is fixed, so that only the error term is relevant in the last line. Minimising it thus corresponds to minimising the divergence. ❑

Examples 3.4.3 We consider several situations where we seek a distribution that best matches given data.

1 Let a natural multiset $\psi \in \mathcal{N}[K](X)$ be given, with size $K > 0$. The distribution $\omega \in \mathcal{D}(X)$ that assigns the highest multinomial draw probability to ψ is $Flrn(\psi)$. That is:

$$Flrn(\psi) \in \operatorname*{argmax}_{\omega \in \mathcal{D}(X)} mn[K](\omega)(\psi).$$

This is obvious if we look at the formulation in item (4) of Theorem 3.4.2 since for $\omega = Flrn(\psi)$ the divergence $D_{KL}(Flrn(\psi), \omega)$ is minimal, namely zero.

This is a standard result, see *e.g.* (Bishop, 2006, §2.2) or (Koller and Friedman, 2009, Ex. 17.5), but it is not always formulated in the same terms. The proof is usually done via some basic analysis, in particular, via Lagrange's multiplier method (like in Bishop (2006)). In our case we use the previous lemma.

2 Let the data now form a non-empty multiset $\psi \in \mathcal{N}(\mathbb{N})$ on the natural numbers. Somehow we know that it matches a binomial distribution $bn[K](r) \in \mathcal{D}(\mathbb{N})$, where the size parameter K is known, but the bias $r \in [0, 1]$ is not. Our aim is thus to find the bias r so that $bn[K](r)$ best matches ψ, for instance because we get the highest likelihood $\prod_n bn[K](r)(n)^{\psi(n)}$. We assume that $K \geq \max(supp(\psi))$, so that $supp(\psi) \subseteq \{0, \ldots, K\} = supp(bn[K](r))$.

We approach this matter by defining the following function $f\colon [0, 1] \to \mathbb{R}$, as logarithm of the expression that we wish to maximise.

$$f(r) \coloneqq \ln\left(\prod_{0 \leq n \leq K} bn[K](r)(n)^{\psi(n)}\right) = \sum_{0 \leq n \leq K} \psi(n) \cdot \ln\left(\binom{K}{n} \cdot r^n \cdot (1 - r)^{K-n}\right).$$

The derivative is:

$$f'(r) = \sum_{0 \leq n \leq K} \frac{\psi(n) \cdot n}{r} - \frac{\psi(n) \cdot (K - n)}{1 - r} = \frac{\sum_n \psi(n) \cdot n}{r} - \frac{\|\psi\| \cdot K - \sum_n \psi(n) \cdot n}{1 - r}.$$

Solving $f'(r) = 0$ yields:

$$r = \frac{1}{K \cdot \|\psi\|} \cdot \sum_{0 \leq n \leq K} \psi(n) \cdot n = \frac{1}{K} \cdot \sum_{0 \leq n \leq K} Flrn(\psi)(n) \cdot n = \frac{Mean(Flrn(\psi))}{K}.$$

The mean function will be introduced only later, in Section 4.2. In Example 4.2.2 (2) we shall see that the mean of the binomial distribution $bn[K](r)$ is $K \cdot r$. The above equation thus shows that our solution r yields $Mean(bn[K](r)) = Mean(Flrn(\psi))$.

3 We consider a similar situation for Poisson distributions: we have data $\psi \in \mathcal{D}(\mathbb{N}) \subseteq \mathcal{D}_\infty(\mathbb{N})$ and ask ourselves for which rate $\lambda \in \mathbb{R}_{\geq 0}$ we have a best match with the Poisson distribution $pois[\lambda] \in \mathcal{D}_\infty(\mathbb{N})$.

We use a similar function:

$$f(\lambda) \coloneqq \ln\left(\prod_{k \in \mathbb{N}} pois[\lambda](k)^{\psi(k)}\right) = \sum_{k \in \mathbb{N}} \psi(k) \cdot \ln\left(e^{-\lambda} \cdot \frac{\lambda^k}{k!}\right) = -\|\psi\| \cdot \lambda + \sum_{k \in \mathbb{N}} \psi(k) \cdot \ln\left(\frac{\lambda^k}{k!}\right).$$

The derivative is thus:

$$f'(\lambda) = -\|\psi\| + \sum_{k \in \mathbb{N}} \frac{\varphi(k)}{\lambda^k} \cdot k \cdot \lambda^{k-1} = -\|\psi\| + \frac{\sum_k \varphi(k) \cdot k}{\lambda}.$$

Solving $f'(\lambda) = 0$ gives as solution:

$$\lambda = \frac{\sum_k \varphi(k) \cdot k}{\|\psi\|} = \sum_{k \in \mathbb{N}} Flrn(\varphi)(k) \cdot k = Mean(Flrn(\varphi)).$$

The mean shows up again, in a similar way as in the previous item, since $Mean(pois[\lambda]) = \lambda$, see Example 4.2.2 (3).

In the results so far we started from a *single* multiset / draw $\psi \in \mathcal{N}(X)$ and looked for the distribution ω that best matches ψ. What if we have a *distribution* of such multisets, of the form $\Omega \in \mathcal{D}(\mathcal{N}[K](X))$? Then we can look for a distribution ω such that the divergence between Ω and $mn[K](\omega)$ is minimal. Such kind of approximations are the topic of the area of 'variational inference', see *e.g.* Mackay (2003); Murphy (2012). It turns out that the distribution $Flrn_*(\Omega)$ gives the least divergence from Ω. This builds on Jacobs and Stein (2023c) and is expressed here via a Pythagereon equation. Later on, in Section 7.8, we shall explain such Pythagereon equations in terms of shapes and factorisations.

Theorem 3.4.4 *Fix a set X and a number X and let $\Omega \in \mathcal{D}(\mathcal{N}[K](X))$ be an arbitrary distribution.*

1 For all $\omega \in \mathcal{D}(X)$ there is an equation:

$$\begin{aligned} D_{KL}\Big(\Omega,\ mn[K](\omega)\Big) &= D_{KL}\Big(\Omega,\ mn[K](Flrn_*(\Omega))\Big) + D_{KL}\Big(mn[K](Flrn_*(\Omega)),\ mn[K](\omega)\Big) \\ &= D_{KL}\Big(\Omega,\ mn[K](Flrn_*(\Omega))\Big) + K \cdot D_{KL}\Big(Flrn_*(\Omega),\ \omega\Big). \end{aligned}$$

This last equation is a consequence of Proposition 3.4.1.

2 As a result,

$$Flrn_*(\Omega) \in \operatorname*{argmin}_{\omega \in \mathcal{D}(X)} D_{KL}(\Omega,\ mn[K](\omega)).$$

Proof. 1 We compute the difference:

$$\begin{aligned} & D_{KL}\Big(\Omega,\ mn[K](\omega)\Big) - D_{KL}\Big(\Omega,\ mn[K](Flrn_*(\Omega))\Big) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} \Omega(\varphi) \cdot \ln\left(\frac{\Omega(\varphi)}{mn[K](\omega)(\varphi)}\right) \\ &\qquad - \sum_{\varphi \in \mathcal{N}[K](X)} \Omega(\varphi) \cdot \ln\left(\frac{\Omega(\varphi)}{mn[K](Flrn_*(\Omega))(\varphi)}\right) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} \Omega(\varphi) \cdot \ln\left(\frac{mn[K](Flrn_*(\Omega))(\varphi)}{mn[K](\omega)(\varphi)}\right) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} \Omega(\varphi) \cdot \ln\left(\frac{\binom{\varphi} \cdot \prod_x Flrn_*(\Omega)(x)^{\varphi(x)}}{\binom{\varphi} \cdot \prod_x \omega(x)^{\varphi(x)}}\right) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in X} \Omega(\varphi) \cdot \varphi(x) \cdot \ln\left(\frac{Flrn_*(\Omega)(x)}{\omega(x)}\right) \\ &= \sum_{\varphi \in \mathcal{N}[K](X)} K \cdot \sum_{x \in X} Flrn_*(\Omega)(x) \ln\left(\frac{Flrn_*(\Omega)(x)}{\omega(x)}\right) \\ &= K \cdot D_{KL}\Big(Flrn_*(\Omega),\ \omega\Big) \\ &= D_{KL}\Big(mn[K](Flrn_*(\Omega)),\ mn[K](\omega)\Big), \qquad \text{by Proposition 3.4.1.} \end{aligned}$$

2 The previous point implies, for each $\omega \in \mathcal{D}(X)$,

$$D_{KL}\Big(\Omega,\ mn[K](\omega)\Big) \geq D_{KL}\Big(\Omega,\ mn[K](Flrn_*(\Omega))\Big). \qquad \square$$

Exercises

3.4.1 Let $\psi \in \mathcal{N}[K](X)$ be given natural multiset of size $K > 0$. Consider two distributions $\omega, \omega' \in \mathcal{D}(X)$ with the same support as ψ. Prove the following addition to Theorem 3.4.2.

$$\begin{aligned}\prod_{x\in X} \omega(x)^{\psi(x)} \leq \prod_{x\in X} \omega'(x)^{\psi(x)} &\iff mn[K](\omega)(\psi) \leq mn[K](\omega')(\psi) \\ &\iff -\ln\left(\prod_{x\in X} \omega(x)^{\psi(x)}\right) \geq -\ln\left(\prod_{x\in X} \omega'(x)^{\psi(x)}\right) \\ &\iff D_{KL}(Flrn(\psi), \omega) \geq D_{KL}(Flrn(\psi), \omega').\end{aligned}$$

3.5 The hypergeometric channel

A multinomial distribution assigns probabilities to draws, with replacement. Two variations have been distinguished without such replacement, namely the 'draw-and-delete' (or -1) hypergeometric mode where a drawn ball is actually removed from the urn, and the 'draw-and-duplicate' (or +1) Pólya mode where the drawn ball is returned to the urn together with an additional ball of the same colour. This section describes the main properties of these '-1' draws, building on the earlier description of hypergeometric distributions in Definition 2.6.1.

Since drawn balls are removed in the hypergeometric mode, the urn in question contains fewer balls with each draw. The urn is thus not a distribution, like in the multinomial case, but a multiset, say of size $L \in \mathbb{N}_{>0}$ that changes as a result of draws. Draws are described as multisets of size K, with the restriction $K \leq L$ in hypergeometric mode since one cannot 'overdraw'. The hypergeometric channel thus takes the form:

$$\mathcal{N}[L](X) \xrightarrow{hg[K]} \mathcal{N}[K](X) \tag{3.24}$$

We recall the definition with draws of size $K \leq L$ from an urn $\upsilon \in \mathcal{N}[L](X)$ of size L.

$$hg[K](\upsilon) \overset{(2.48)}{=} \sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{L}{K}} |\varphi\rangle = \sum_{\varphi \leq_K \psi} \frac{\prod_x \binom{\upsilon(x)}{\varphi(x)}}{\binom{L}{K}} |\varphi\rangle.$$

Recall that we write $\varphi \leq_K \upsilon$ for: $\|\varphi\| = K$ and $\varphi \leq \upsilon$, see Definition 1.7.1 (2). An alternative description in terms of sequences of draws is given in Theorem 2.6.2.

A subtle point in the formulation of the hypergeometric distribution is that the draws φ should be multisets over the support $supp(\upsilon)$ of the urn υ. Indeed, only balls with colours that occur in the urn can be drawn. This is handled implicitly via the requirement $\varphi \leq_K \upsilon$. It ensures an inclusion of supports $supp(\varphi) \subseteq supp(\upsilon)$.

A basic fact about hypergeometric channels is that they can be described as iterations of single draws, that is, as iterations of the draw-and-delete channels DD from Definition 3.2.1 (2). As we shall see soon afterwards, this fact has many consequences.

Theorem 3.5.1 *For $L, K \in \mathbb{N}$, the hypergeometric channel $hg[K]\colon \mathcal{N}[K+L](X) \rightsquigarrow \mathcal{N}[K](X)$*

equals L consecutive draw-and-delete's DD^L in:

$$\begin{tikzcd}
\mathcal{N}[K+L](X) \arrow[rr, "hg[K]"] \arrow[dr, "DD"'] & & \mathcal{N}[K](X) \\
& \mathcal{N}[K+L-1](X) \arrow[dr, dotted, bend right, "\underbrace{DD \circ \cdots \circ DD}_{L-2 \text{ times}}"'] & \\
& & \mathcal{N}[K+1](X) \arrow[uur, "DD"']
\end{tikzcd} \tag{3.25}$$

This result is remarkable. It says that if you have an urn with $K+L$ balls, then the distribution of draws of size K appears when you consecutively draw-and-delete L balls. A full picture involving both the draw and the remaining urn will be described in Proposition 3.5.3 below.

The diagram (3.25) involves iterations $DD^L = DD \circ \cdots \circ DD$ *as channels*, given explicitly by:

$$DD^0 = unit \qquad DD^{n+1} = DD \circ DD^n = DD^n \circ DD.$$

Proof. Write $\upsilon \in \mathcal{N}[K+L](X)$ for the urn. The proof proceeds by induction on the number of iterations L, starting with $L = 0$. Then $\varphi \leq_K \upsilon$ means $\varphi = \upsilon$. Hence:

$$hg[K](\upsilon) = \sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{K+0}{K}} \big|\varphi\rangle = \frac{\binom{\upsilon}{\upsilon}}{\binom{K}{K}} \big|\upsilon\rangle = 1\big|\upsilon\rangle = unit(\upsilon) = DD^0(\upsilon).$$

For the induction step we use $\upsilon \in \mathcal{N}[K+(L+1)](X)$ in:

$$\begin{aligned}
DD^{L+1}(\upsilon) &= \sum_{\varphi \leq_K \upsilon} DD_*(DD^L(\upsilon))(\varphi) \big|\varphi\rangle \\
&= \sum_{\varphi \leq_K \upsilon} \sum_{\psi \in \mathcal{N}[K+1](X)} DD^L(\upsilon)(\psi) \cdot DD(\psi)(\varphi) \big|\varphi\rangle \\
&= \sum_{y \in X} \sum_{\varphi \in \mathcal{N}[K](X)} DD^L(\upsilon)(\varphi + 1|y\rangle) \cdot \frac{\varphi(y)+1}{K+1} \big|\varphi\rangle \\
&\overset{\text{(IH)}}{=} \sum_{y \in X} \sum_{\varphi \in \mathcal{N}[K](X), \varphi+1|y\rangle \leq \upsilon} \frac{\binom{\upsilon}{\varphi+1|y\rangle}}{\binom{K+L+1}{K+1}} \cdot \frac{\varphi(y)+1}{K+1} \big|\varphi\rangle \\
&= \sum_{\varphi \leq_K \upsilon} \sum_{y, \varphi(y) < \upsilon(y)} \frac{(\upsilon(y) - \varphi(y)) \cdot \binom{\upsilon}{\varphi}}{(L+1) \cdot \binom{K+L+1}{K}} \big|\varphi\rangle \quad \text{by Exercise 1.8.8} \\
&= \sum_{\varphi \leq_K \upsilon} \frac{((K+L+1) - K) \cdot \binom{\upsilon}{\varphi}}{(L+1) \cdot \binom{K+L+1}{K}} \big|\varphi\rangle = \sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{K+L+1}{K}} \big|\varphi\rangle = hg[K](\upsilon). \qquad \square
\end{aligned}$$

From this result we can deduce many additional facts about hypergeometric distributions.

Corollary 3.5.2 *1 Hypergeometric channels (3.24) are natural in X.*

2 Frequentist learning form hypergeometric draws is like learning from the urn: for $L \geq K$,

$$\begin{tikzcd}
\mathcal{N}[L](X) \arrow[rr, "hg[K]"] \arrow[dr, bend right, "Flrn"'] & & \mathcal{N}[K](X) \arrow[dl, bend left, "Flrn"] \\
& X &
\end{tikzcd}$$

3 For $L \geq K$ one has:

$$\begin{tikzcd} \mathcal{N}[L](X) \arrow[rr, "DD"] \arrow[dr, "hg[K]"'] & & \mathcal{N}[L+1](X) \arrow[dl, "hg[K]"] \\ & \mathcal{N}[K](X) & \end{tikzcd}$$

4 Also, for $L \geq K+1$,

$$\begin{tikzcd} \mathcal{N}[K](X) & & \mathcal{N}[K+1](X) \arrow[ll, "DD"'] \\ & \mathcal{N}[L](X) \arrow[ul, "hg[K]"] \arrow[ur, "hg[K+1]"'] & \end{tikzcd}$$

5 Hypergeometric channels compose, as in:

$$\begin{tikzcd} \mathcal{N}[K+L+M](X) \arrow[rr, "hg[K]"] \arrow[dr, "hg[K+L]"'] & & \mathcal{N}[K](X) \\ & \mathcal{N}[K+L](X) \arrow[ur, "hg[K]"'] & \end{tikzcd}$$

6 Hypergeometric and multinomial channels commute, as in:

$$\begin{tikzcd} \mathcal{N}[K+L](X) \arrow[rr, "hg[K]"] & & \mathcal{N}[K](X) \\ & \mathcal{D}(X) \arrow[ul, "mn[K+L]"] \arrow[ur, "mn[K]"'] & \end{tikzcd}$$

7 Hypergeometric channels commute with multizip: for $L \geq K$,

$$\begin{tikzcd} \mathcal{N}[L](X) \times \mathcal{N}[L](Y) \arrow[r, "hg[K] \otimes hg[K]"] \arrow[d, "mzip"'] & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[d, "mzip"] \\ \mathcal{N}[L](X \times Y) \arrow[r, "hg[K]"] & \mathcal{N}[K](X \times Y) \end{tikzcd}$$

Proof. 1 By naturality of draw-and-delete, see Exercise 3.2.4.
2 By Theorem 3.2.6.
3 By Theorem 3.5.1.
4 Idem.
5 Similarly, since $DD^{L+M} = DD^L \circ DD^M$.
6 By Proposition 3.2.9.
7 By Lemma 3.2.8. ❑

In Proposition 3.2.9 we have seen that multinomial and Pólya channels form draw-delete cones, via commutation with draw-delete *for every number* K, where K is the size of the draws. Item (4) of the above result describes a similar commutation with draw-delete, but this time *not for every* K, but only for $K+1 \leq L$, where L is the size of the urn. Hence hypergeometric channels do not form draw-delete cones.

Aside: although physically impossible, mathematically one can 'overdraw', that is, draw more balls than are present in an urn. In Jacobs and Stein (2023b) this is elaborated in terms of 'signed' distributions, in which probabilities may be negative. The resulting 'signed' hypergeometric channel does form a draw-delete cone.

In Exercise 3.2.12 we have seen the draw-store-delete channel, of the form $DSD\colon \mathcal{N}[K+$

$1](X) \rightsquigarrow X \times \mathcal{N}[K](X)$. We can extend it to the hypergeometric case, for which we introduce the *ad hoc* name hypergeometric-store, abbreviated as $hgs[K] \colon \mathcal{N}[K{+}L](X) \rightsquigarrow \mathcal{N}[K](X) \times \mathcal{N}[L](X)$. It keeps track both of the draw and the remaining urn:

$$hgs[K](\upsilon) \coloneqq \sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{K+L}{K}} \Big| \varphi, \upsilon - \varphi \Big\rangle . \tag{3.26}$$

We consider the two marginals.

Proposition 3.5.3 *The two marginals of the above hypergeometric-store channel* (3.26) *are iterated draw-delete's in:*

$$\begin{tikzcd}
 & \mathcal{N}[K+L](X) \arrow[dl, bend right, "hg[K] = DD^L"'] \arrow[d, "hgs[K]"] \arrow[dr, bend left, "DD^K"] & \\
\mathcal{N}[K](X) & \mathcal{N}[K](X) \times \mathcal{N}[L](X) \arrow[l, "\pi_1"'] \arrow[r, "\pi_2"'] & \mathcal{N}[L](X)
\end{tikzcd}$$

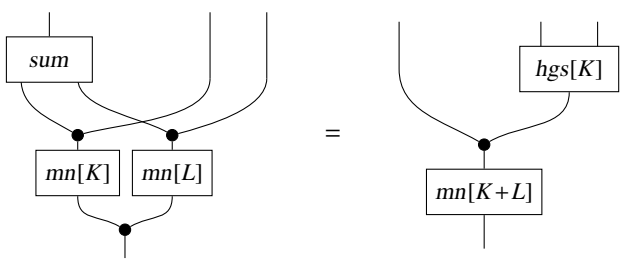
Proof. The triangle on the left commutes by Theorem 3.5.1:

$$\pi_{1*}(hgs[K](\upsilon)) \overset{(3.26)}{=} \mathcal{D}(\pi_1)\left(\sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{K+L}{K}} \Big| \varphi, \upsilon - \varphi \Big\rangle \right) = \sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{K+L}{K}} \big| \varphi \rangle = hg[K](\upsilon) = DD^L(\upsilon).$$

For the triangle on the right we use Theorem 3.2.7:

$$\pi_{2*}(hgs[K](\upsilon)) \overset{(3.26)}{=} \mathcal{D}(\pi_2)\left(\sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{K+L}{K}} \Big| \varphi, \upsilon - \varphi \Big\rangle \right) = \sum_{\varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{K+L}{K}} \big| \upsilon - \varphi \rangle \overset{(3.11)}{=} DD^K(\upsilon). \qquad \square$$

Theorem 3.5.4 *The sum of multisets function and the hypergeometric-store channel* (3.26) *are related in the following way:*



(3.27)

The string diagram equation that we have seen earlier in Exercise 3.2.12 for draw-store-delete is a special case of this theorem, for $K = 1$. A consequence of (3.27), obtained by marginalising out the outputs on the right, is the closure of multinomials under convolution, that we saw earlier in Exercise 3.3.10.

Proof. For a distribution $\omega \in \mathcal{D}(X)$ one has:

$$\begin{aligned}
&\langle sum, id\rangle_*(mn[K](\omega)\otimes mn[L](\omega))\\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{\psi\in\mathcal{N}[L](X)} mn[K](\omega)(\varphi)\cdot mn[L](\omega)(\psi)\,\Big|\,\varphi+\psi,\varphi,\psi\Big\rangle\\
&= \sum_{\varphi\in\mathcal{N}[K](X)}\ \sum_{\psi\in\mathcal{N}[L](X)} (\varphi)\cdot\prod_x \omega(x)^{\varphi(x)}\cdot(\psi)\cdot\prod_x\omega(x)^{\psi(x)}\,\Big|\,\varphi+\psi,\varphi,\psi\Big\rangle\\
&= \sum_{\psi\in\mathcal{N}[L](X)}\ \sum_{\varphi\in\mathcal{N}[K](X)} (\varphi)\cdot(\psi)\cdot\prod_x \omega(x)^{(\varphi+\psi)(x)}\,\Big|\,\varphi+\psi,\varphi,\psi\Big\rangle\\
&= \sum_{\upsilon\in\mathcal{N}[K+L](X)}\ \sum_{\varphi\leq_K\upsilon} \frac{(\varphi)\cdot(\upsilon-\varphi)}{(\upsilon)}\cdot(\upsilon)\cdot\prod_x\omega(x)^{\upsilon(x)}\,\Big|\,\upsilon,\varphi,\upsilon-\varphi\Big\rangle\\
&= \sum_{\upsilon\in\mathcal{N}[K+L](X)}\ \sum_{\varphi\leq_K\upsilon} \frac{K!\cdot L!\cdot\upsilon!}{\varphi!\cdot(\upsilon-\varphi)!\cdot(K+L)!}\cdot mn[K+L](\omega)(\upsilon)\,\Big|\,\upsilon,\varphi,\upsilon-\varphi\Big\rangle\\
&= \sum_{\upsilon\in\mathcal{N}[K+L](X)}\ \sum_{\varphi\leq_K\upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{K+L}{K}}\cdot mn[K+L](\omega)(\upsilon)\,\Big|\,\upsilon,\varphi,\upsilon-\varphi\Big\rangle\\
&= \langle id, hgs[K]\rangle_*(mn[K+L](\omega)).
\end{aligned}$$

❑

There is the following analogue of Lemma 3.3.3.

Lemma 3.5.5 *Let $\upsilon\in\mathcal{N}(X)$ be a non-empty urn of size $L=\|\upsilon\|>0$.*

1 For a fixed element $y\in X$,

$$\sum_{\varphi\leq_K\upsilon} hg[K](\upsilon)(\varphi)\cdot\varphi(y) \;=\; K\cdot Flrn(\upsilon)(y), \tag{3.28}$$

This means that: $hg[K](\upsilon)\models Flrn(-)(y) \;=\; Flrn(\upsilon)(y)$.

2 For $L\geq K\geq 1$,

$$flat\Big(hg[K](\upsilon)\Big) \;=\; \sum_{\varphi\leq_K\upsilon} hg[K](\upsilon)(\varphi)\cdot\varphi \;=\; \frac{K}{L}\cdot\upsilon \;=\; K\cdot Flrn(\upsilon).$$

Proof. 1 We use Exercises 1.8.12 and 1.8.8 in the marked equation $\overset{(*)}{=}$ below. Let $K\leq L$ in:

$$\begin{aligned}
\sum_{\varphi\leq_K\upsilon} hg[K](\upsilon)(\varphi)\cdot\varphi(y) &= \sum_{\varphi\leq_K\upsilon}\frac{\binom{\upsilon}{\varphi}\cdot\varphi(y)}{\binom{L}{K}}\\
&\overset{(*)}{=} \sum_{\varphi\leq_K\upsilon}\frac{\upsilon(y)\cdot\binom{\upsilon-1|y\rangle}{\varphi-1|y\rangle}}{\frac{L}{K}\cdot\binom{L-1}{K-1}}\\
&= K\cdot\frac{\upsilon(y)}{L}\cdot\sum_{\varphi\leq_{K-1}\upsilon-1|y\rangle}\frac{\binom{\upsilon-1|y\rangle}{\varphi}}{\binom{L-1}{K-1}} \;=\; K\cdot Flrn(\upsilon)(y).
\end{aligned}$$

2 We use the previous point in:

$$\begin{aligned}
\sum_{\varphi\leq_K\upsilon} hg[K](\upsilon)(\varphi)\cdot\varphi \;&=\; \sum_{x\in X}\left(\sum_{\varphi\leq_K\upsilon} hg[K](\upsilon)(\varphi)\cdot\varphi(x)\right)\Big|x\Big\rangle \overset{(3.28)}{=} \sum_{x\in X} K\cdot Flrn(\upsilon)(x)\,\Big|x\Big\rangle\\
&=\; K\cdot Flrn(\upsilon).
\end{aligned}$$

❑

Exercises

3.5.1 We have seen that draw-delete DD commutes with hypergeometric and Pólya distributions, see Corollary 3.5.2 (4) and Proposition 3.2.9. The aim of this exercise is to check that draw-add DA does *not* commute with hypergeometric, nor with Pólya, distributions, for instance by checking that for $\varphi = 3|a\rangle + 1|b\rangle$,

$$\begin{aligned} hg[2](\varphi) &= \tfrac{1}{2}\big|2|a\rangle\big\rangle + \tfrac{1}{2}\big|1|a\rangle + 1|b\rangle\big\rangle \\ hg[3](\varphi) &= \tfrac{1}{4}\big|3|a\rangle\big\rangle + \tfrac{3}{4}\big|2|a\rangle + 1|b\rangle\big\rangle \\ DA_*(hg[2](\varphi)) &= \tfrac{1}{2}\big|3|a\rangle\big\rangle + \tfrac{1}{4}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{4}\big|1|a\rangle + 2|b\rangle\big\rangle, \end{aligned}$$

And:

$$\begin{aligned} pl[2](\varphi) &= \tfrac{3}{5}\big|2|a\rangle\big\rangle + \tfrac{3}{10}\big|1|a\rangle + 1|b\rangle\big\rangle + \tfrac{1}{10}\big|2|b\rangle\big\rangle \\ pl[3](\varphi) &= \tfrac{1}{2}\big|3|a\rangle\big\rangle + \tfrac{3}{10}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{3}{20}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{20}\big|3|b\rangle\big\rangle \\ DA_*(pl[2](\varphi)) &= \tfrac{3}{5}\big|3|a\rangle\big\rangle + \tfrac{3}{20}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{3}{20}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{10}\big|3|b\rangle\big\rangle \end{aligned}$$

3.5.2 1 Give a direct proof of Theorem 3.5.1 for $L = 1$.

2 Elaborate also the case $K = 1$ in Theorem 3.5.1 and show that in that case $DD^L(\psi) = Flrn(\psi)$, at least when we identify the (isomorphic) sets $\mathcal{N}[1](X)$ and X.

3.5.3 This exercise considers two trivial cases.

1 Show that $hg[K](L|x\rangle) = 1\big|K|x\rangle\big\rangle$, for $L \geq K$.

2 And also that the following diagram commutes, for $L \geq 1$, using the identification $X \cong \mathcal{N}[1](X)$.

$$\begin{tikzcd} & \mathcal{D}(\mathcal{N}[1](X)) \arrow[d, equal, "\wr"] \\ \mathcal{N}[L](X) \arrow[ur, bend left, "hg[1]"] \arrow[r, bend right, "Flrn"'] & \mathcal{D}(X) \end{tikzcd}$$

3.5.4 Let $unif_L \in \mathcal{D}(\mathcal{N}[L](X))$ be the uniform distribution on multisets of size L, from Exercise 3.2.5, where X is a finite set. Show that $hg[K]_*(unif_L) = unif_K$, for $K \leq L$.

3.5.5 Prove in analogy with Exercise 3.3.8, for an urn $\upsilon \in \mathcal{N}[L](X)$ and elements $y, z \in X$, the following points.

1 When $y \neq z$ and $K \leq L$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} hg[K](\upsilon)(\varphi) \cdot \varphi(y) \cdot \varphi(z) \;=\; K \cdot (K-1) \cdot Flrn(\upsilon)(y) \cdot \frac{\upsilon(z)}{L-1}.$$

2 When $K \leq L$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} hg[K](\upsilon)(\varphi) \cdot \varphi(y)^2 \;=\; K \cdot Flrn(\upsilon)(y) \cdot \frac{(K-1) \cdot \upsilon(y) + (L-K)}{L-1}.$$

3 Use the previous two points to prove:

$$\begin{aligned} &\langle Flrn, Flrn\rangle_*\Big(hg[K](\upsilon)\Big) \\ &= \frac{L \cdot (K-1)}{(L-1) \cdot K} \cdot \Big(Flrn(\upsilon) \otimes Flrn(\upsilon)\Big) \;+\; \frac{L-K}{(L-1) \cdot K} \cdot \mathcal{D}(\Delta)\Big(Flrn(\upsilon)\Big). \end{aligned}$$

3.5.6 Use Theorem 3.5.1 to prove the following two recurrence relations for hypgeometric distributions.

$$hg[K](\upsilon) = \sum_{x \in supp(\upsilon)} Flrn(\upsilon)(x) \cdot hg[K-1](\upsilon - 1|x\rangle)$$
$$hg[K](\upsilon)(\varphi) = \sum_{x} \frac{\varphi(x)+1}{K+1} \cdot hg[K-1](\upsilon)(\varphi + 1|x\rangle)$$

3.5.7 Fix numbers $N, M \in \mathbb{N}$ and write $\upsilon = N|0\rangle + M|1\rangle$ for an urn with N balls of colour 0 and M of colour 1. Let $n \leq N$. Show that:

$$\sum_{0 \leq m \leq M} hg[n+m](\upsilon)(n|0\rangle + m|1\rangle) = \frac{N+M+1}{N+1}.$$

Hint: Use Exercise 1.8.11. Notice that the right-hand side does not depend on n.

3.6 The Pólya channel

This section takes a closer look at Pólya distributions, from Definition 2.6.3. They resemble the hypergeometric ones, but there are essential differences. Hypergeometric distributions are based on the draw-and-delete mode whereas Pólya uses draw-and-duplicate: a drawn ball is returned to the urn, together with an extra ball of the same colour. Such an additional ball has a strengthening effect that can capture situations with a cluster dynamics Hayhoe et al. (2017); Lau et al. (2020); Blei et al. (2003).

The Pólya distribution for draws of size K can be described as a channel of the form:

$$\mathcal{N}_*(X) \xrightarrow{pl[K]} \mathcal{N}[K](X) \tag{3.29}$$

The domain is the set $\mathcal{N}_*(X)$ of non-empty multisets, over the set of colours X.

In its definition Pólya uses multichoose $\left(\!\binom{-}{-}\!\right)$, instead of ordinary binomial coefficients $\binom{-}{-}$ used for hypergeometric distributions. Its draw sizes may exceed the size of the urn, in the sense that more balls may be drawn then are actually in the urn, since the urn grows in size with the draw of each single ball. There still is an obvious restriction, namely that only colours that exist in the urn can occur in draws. Indeed, in the formulation below the draws φ are multisets over the support of the urn υ.

$$pl[K](\upsilon) \overset{(2.50)}{=} \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} \frac{\left(\!\binom{\upsilon}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} |\varphi\rangle = \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} \frac{\prod_{x \in supp(\upsilon)} \left(\!\binom{\upsilon(x)}{\varphi(x)}\!\right)}{\left(\!\binom{L}{K}\!\right)} |\varphi\rangle.$$

Theorem 2.6.4 describes these Pólya distributions in terms of sequences of draws.

These urns to which one applies the draw-and-duplicate form of drawing are sometimes called *Pólya urns*, see *e.g.* Mahmoud (2008). Later on in Section **??**, within a continuous setting, Pólya distributions appear as multinomials over Dirichlet. Therefore, the name Dirchlet-multinomial is also used. Here it is reserved for the generalisation to non-natural multisets in Remark 2.6.5.

We start with an analogue of Lemma 3.5.5.

Lemma 3.6.1 *Fix a non-empty urn $\upsilon \in \mathcal{N}(X)$.*

1 For an arbitrary element $y \in X$,

$$\sum_{\varphi \in \mathcal{M}[K](supp(\upsilon))} pl[K](\upsilon)(\varphi) \cdot \varphi(y) \;=\; K \cdot Flrn(\upsilon)(y). \tag{3.30}$$

As a result, $pl[K](\upsilon) \models Flrn(-)(y) \;=\; Flrn(\upsilon)(y)$.

2 For $K \geq 1$,

$$flat\Big(pl[K](\upsilon)\Big) \;=\; \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} pl[K](\upsilon)(\varphi) \cdot \varphi \;=\; \frac{K}{L} \cdot \upsilon \;=\; K \cdot Flrn(\upsilon).$$

Proof. 1 We use Exercises 1.8.12 and 1.8.8 in the marked equation $\overset{(*)}{=}$ in:

$$\begin{aligned}
\sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} pl[K](\upsilon)(\varphi) \cdot \varphi(y) \;&=\; \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} \frac{\binom{\upsilon}{\varphi} \cdot \varphi(y)}{\binom{L}{K}} \\
&\overset{(*)}{=} \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} \frac{\upsilon(y) \cdot \binom{\upsilon+1|y\rangle}{\varphi-1|y\rangle}}{\frac{L}{K} \cdot \binom{L+1}{K-1}} \\
&=\; K \cdot \frac{\upsilon(y)}{L} \cdot \sum_{\varphi \in \mathcal{N}[K-1](supp(\upsilon+1|y\rangle))} \frac{\binom{\upsilon+1|y\rangle}{\varphi}}{\binom{L+1}{K-1}} \;=\; K \cdot Flrn(\upsilon)(y).
\end{aligned}$$

2 We proceed as follows.

$$\begin{aligned}
\sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} pl[K](\upsilon)(\varphi) \cdot \varphi \;&=\; \sum_{x \in X} \left(\sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} pl[K](\upsilon)(\varphi) \cdot \varphi(x) \right) \big| x \rangle \\
&\overset{(3.30)}{=} \sum_{x \in X} K \cdot Flrn(\upsilon)(x) \big| x \rangle \;=\; K \cdot Flrn(\upsilon).
\end{aligned}$$

❑

In the previous section we have seen that hypergeometric channels are in fact iterated draw-and-delete's, see Theorem 3.5.1. One may expect that Pólya draw channels arise analogously as repeted draw-and-add's. This is not the case. But we do have the following Pólya analogues of Theorem 3.5.2 (2) and (3).

Proposition 3.6.2 *Let $K > 0$.*

1 Pólya functions $pl[K] \colon \mathcal{N}_*(X) \to \mathcal{D}(\mathcal{N}[K](X))$ are natural in X.

2 Frequentist learning and Pólya satisfy the following equation:

$$\begin{tikzcd}
\mathcal{N}_*(X) \arrow[rr, "pl[K]"] \arrow[dr, "Flrn"'] & & \mathcal{N}[K](X) \arrow[dl, "Flrn"] \\
& X &
\end{tikzcd}$$

3 The hypergeometric channel preserves Pólya distributions:

$$\begin{tikzcd}
\mathcal{N}[K](X) & & \mathcal{N}[K+L](X) \arrow[ll, "hg[K]"'] \\
& \mathcal{N}_*(X) \arrow[ul, "pl[K]"] \arrow[ur, "pl[K+L]"'] &
\end{tikzcd}$$

4 *Doing a draw-and-add before Pólya has no effect: for $L > 0$,*

$$\begin{tikzcd} \mathcal{N}[L](X) \arrow[rr, "DA"] \arrow[dr, "pl[K]"'] & & \mathcal{N}[L+1](X) \arrow[dl, "pl[K]"] \\ & \mathcal{N}[K](X) & \end{tikzcd}$$

Proof. 1 Easy.

2 For a multiset/urn $\upsilon \in \mathcal{N}(X)$ with $\|\upsilon\| = L > 0$, and for $x \in X$,

$$\begin{aligned} \Big(Flrn \circ pl[K]\Big)(\upsilon)(x) &= \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} Flrn(\varphi)(x) \cdot pl[K](\upsilon)(\varphi) \\ &= \frac{1}{K} \cdot \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} \varphi(x) \cdot pl[K](\upsilon)(\varphi) \overset{(3.30)}{=} Flrn(\upsilon)(x). \end{aligned}$$

3 The hypergeometric channel $hg[K] \colon \mathcal{N}[K+L](X) \to \mathcal{N}[K](X)$ is an iteration of L draw-delete's, see Theorem 3.5.1. These draw-delete's preserve Pólya, see Proposition 3.2.9.

4 We use Exercises 1.8.12 and 1.8.8 in the marked equation $\overset{(*)}{=}$ below. For urn $\upsilon \in \mathcal{N}[L](X)$ and draw $\varphi \in \mathcal{N}[K](supp(\upsilon))$,

$$\begin{aligned} \Big(pl[K] \circ DA\Big)(\upsilon)(\varphi) &= \sum_{x \in supp(\upsilon)} \frac{\upsilon(x)}{L} \cdot pl[K](\upsilon + 1|x\rangle)(\varphi) \\ &= \sum_{x \in supp(\upsilon)} \frac{\upsilon(x)}{L} \cdot \frac{\left(\!\binom{\upsilon+1|x\rangle}{\varphi}\!\right)}{\left(\!\binom{L+1}{K}\!\right)} \\ &\overset{(*)}{=} \sum_{x \in supp(\upsilon)} \frac{\upsilon(x) + \varphi(x)}{L+K} \cdot \frac{\left(\!\binom{\upsilon}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} = \frac{\left(\!\binom{\upsilon}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} = pl[K](\upsilon)(\varphi). \end{aligned}$$

❑

In (3.26) we have seen the hypergeometric-store channel that records both the draw and the remaining urn. There is a similar Pólya-store channel of the form $pls \colon \mathcal{N}[L](X) \rightsquigarrow \mathcal{N}[K](X) \times \mathcal{N}[K+L](X)$, defined as:

$$pls[K](\upsilon) := \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} \frac{\left(\!\binom{\upsilon}{\varphi}\!\right)}{\left(\!\binom{L}{K}\!\right)} \Big| \varphi, \upsilon + \varphi \Big\rangle. \tag{3.31}$$

In the hypergeometric case both the marginals of the 'store' channel are iterations of draw-delete's, see Proposition 3.5.3. In the Pólya case only the second marginal is an iteration, of draw-add's.

Proposition 3.6.3 *The two marginals of the Pólya-store channel* (3.31) *are iterated draw-delete's in:*

$$\begin{tikzcd} & \mathcal{N}[L](X) \arrow[dl, bend right, "pl[K]"'] \arrow[d, "pls[K]"] \arrow[dr, bend left, "DA^K"] & \\ \mathcal{N}[K](X) & \mathcal{N}[K](X) \times \mathcal{N}[L+K](X) \arrow[l, "\pi_1"] \arrow[r, "\pi_2"'] & \mathcal{N}[L+K](X) \end{tikzcd}$$

Proof. Commutation of the triangle on the left holds trivially. Commutation on the right follows from Theorem 3.2.7. ❑

The hypergeometric channel can be expressed as an iteration of draw-delete's, see Theorem 3.5.1. This does not work for the Pólya channel: it is not such an iteration of draw-add's. We do have the following result.

Proposition 3.6.4 *Let $\upsilon \in \mathcal{N}[L](X)$ be an urn, where $L > 0$. For each $K \in \mathbb{N}$ one can write:*

$$\mathcal{D}((-) + \upsilon)\big(pl[K](\upsilon)\big) \;=\; DA^K(\upsilon).$$

The expression $(-) + \upsilon$ is used for the multiset addition function $\varphi \mapsto \varphi + \upsilon$, from $\mathcal{N}[K](X)$ to $\mathcal{N}[L + K](X)$. The operation $\mathcal{D}((-) + \upsilon)$ performs a shift, as in Exercise 2.7.8. The result may also be described as $1|\upsilon\rangle + pl[K](\upsilon)$.

We can now write:

$$pl[K](\upsilon) \;=\; \mathcal{D}((-) - \upsilon)\big(DA^K(\upsilon)\big).$$

Proof. We only prove the first equation, because the second one is a direct consequence. We use induction on $K \geq 0$, starting at $K = 0$. Then:

$$\begin{aligned}
\mathcal{D}((-) + \upsilon)\big(pl[0](\upsilon)\big) &= \mathcal{D}((-) + \upsilon)\left(\sum_{\varphi \in \mathcal{N}[0](supp(\upsilon))} \frac{\binom{\upsilon}{\varphi}}{\binom{L}{0}} \big|\varphi\rangle\right) \\
&= \mathcal{D}((-) + \upsilon)\left(\frac{\binom{\upsilon}{\mathbf{0}}}{\binom{L}{0}} \big|\mathbf{0}\rangle\right) \\
&= \mathcal{D}((-) + \upsilon)(1\big|\mathbf{0}\rangle) \;=\; 1|\upsilon\rangle \;=\; unit(\upsilon) \;=\; DA^0(\upsilon).
\end{aligned}$$

Next, using Exercises 1.8.12 and 1.8.8 in the marked equation $\overset{(E)}{=}$ below.

$$\begin{aligned}
DA^{K+1}(\upsilon) &= DA_*(DA^K(\upsilon)) \\
&\overset{(IH)}{=} DA_*\Big(\mathcal{D}((-) + \upsilon)\big(pl[K](\upsilon)\big)\Big) \\
&= \sum_{\psi \in \mathcal{N}[L+K](X)} \sum_{x \in supp(\psi)} \frac{\psi(x)}{K + L} \cdot \mathcal{D}((-) + \upsilon)\big(pl[K](\upsilon)\big)(\psi)\,\big|\psi + 1|x\rangle\rangle \\
&= \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} \sum_{x \in supp(\varphi)} \frac{\varphi(x) + \upsilon(x)}{K + L} \cdot \frac{\binom{\upsilon}{\varphi}}{\binom{L}{K}} \big|\varphi + \upsilon + 1|x\rangle\rangle \\
&\overset{(E)}{=} \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} \sum_{x \in supp(\varphi)} \frac{\varphi(x) + 1}{K + 1} \cdot \frac{\binom{\upsilon}{\varphi + 1|x\rangle}}{\binom{L}{K+1}} \big|\varphi + 1|x\rangle + \upsilon\rangle \\
&= \sum_{\psi \in \mathcal{N}[K+1](supp(\upsilon))} \sum_{x \in supp(\varphi)} \frac{\psi(x)}{K + 1} \cdot \frac{\binom{\upsilon}{\psi}}{\binom{L}{K+1}} \big|\psi + \upsilon\rangle \;=\; \mathcal{D}((-) + \upsilon)\big(pl[K + 1](\upsilon)\big). \qquad \square
\end{aligned}$$

3.6.1 Drawing from large urns

In the remainder of this section we look several aspects of both hypergeometric and Pólya distributions in 'limiting' situations, for instance, when the urn becomes very large. In that case it matters very little if balls are removed or added to the urn, via hypergeometric draw-delete or Pólya draw-add. Indeed, in the limit, both hypergeometric and Pólya distributions look like multinomials. This is called the law of large urns in Jacobs (2025c), the source of the limit results given below.

For the following result we use Lemma 1.2.3, specifically in the form:

$$\lim_{n\to\infty} \frac{n!}{(n-m)! \cdot n^m} \;=\; 1 \;=\; \lim_{n\to\infty} \frac{(n+m-1)!}{(n-1)! \cdot n^m}. \tag{3.32}$$

Proposition 3.6.5 *Let $X = \{x_1, \ldots, x_N\}$ be a finite set with $N \geq 2$ elements and let $\varphi \in \mathcal{N}[K](X)$ be a draw of size $K \in \mathbb{N}_{>0}$.*

1 *The hypergeometrical probability of drawing φ from an urn becomes equal to the multinomial probability of drawing φ from the frequentist learning of the urn, when the number of balls in the urn for each color x_i goes to infinity:*

$$\lim_{n_1,\ldots n_N\to\infty} \frac{mn[K]\Big(Flrn(\sum_i n_i|x_i\rangle)\Big)(\varphi)}{hg[K]\Big(\sum_i n_i|x_i\rangle\Big)(\varphi)} \;=\; 1.$$

2 *Similarly, Pólya draw probabilities from a large urn are like multinomial probabilities:*

$$\lim_{n_1,\ldots n_N\to\infty} \frac{mn[K]\Big(Flrn(\sum_i n_i|x_i\rangle)\Big)(\varphi)}{pl[K]\Big(\sum_i n_i|x_i\rangle\Big)(\varphi)} \;=\; 1.$$

The binary case of point (1) is discussed in (Jaynes, 2003, §3.7). Later, in Proposition 4.8.2 we shall reformulate these results in terms of distances between hypergeometric / Pólya distributions and multinomial distributions.

Proof. 1 By unpacking the relevant definitions:

$$\begin{aligned}
&\lim_{n_1,\ldots n_N\to\infty} \frac{mn[K]\Big(Flrn(\sum_i n_i|x_i\rangle)\Big)(\varphi)}{hg[K]\Big(\sum_i n_i|x_i\rangle\Big)(\varphi)} \\
&= \lim_{n_1,\ldots n_N\to\infty} \frac{K!}{\prod_i \varphi(x_i)!} \cdot \prod_i \Big(\tfrac{n_i}{\sum_j n_j}\Big)^{\varphi(x_i)} \cdot \frac{\binom{\sum_j n_j}{K}}{\prod_i \binom{n_i}{\varphi(x_i)}} \\
&= \lim_{n_1,\ldots n_N\to\infty} \frac{(\sum_j n_j)!}{((\sum_j n_j) - K)! \cdot (\sum_j n_j)^K} \cdot \prod_i \frac{(n_i - \varphi(x_i))! \cdot n_i^{\varphi(x_i)}}{n_i!} \\
&= \Big(\lim_{n\to\infty} \frac{n!}{(n-K)! \cdot n^K}\Big) \cdot \Big(\prod_i \lim_{n_i\to\infty} \frac{(n_i - \varphi(x_i))! \cdot n_i^{\varphi(x_i)}}{n_i!}\Big) \\
&\overset{(3.32)}{=} 1.
\end{aligned}$$

2 Similarly:

$$\begin{aligned}
&\lim_{n_1,\ldots n_N\to\infty} \frac{mn[K]\Big(Flrn(\sum_i n_i|x_i\rangle)\Big)(\varphi)}{pl[K]\Big(\sum_i n_i|x_i\rangle\Big)(\varphi)} \\
&= \lim_{n_1,\ldots n_N\to\infty} \frac{K!}{\prod_i \varphi(x_i)!} \cdot \prod_i \Big(\tfrac{n_i}{\sum_j n_j}\Big)^{\varphi(x_i)} \cdot \frac{\Big(\binom{\sum_j n_j}{K}\Big)}{\prod_i \Big(\binom{n_i}{\varphi(x_i)}\Big)} \\
&= \lim_{n_1,\ldots n_N\to\infty} \frac{((\sum_j n_j)+K-1)!}{((\sum_j n_j)-1)!\cdot(\sum_j n_j)^K} \cdot \prod_i \frac{(n_i-1)!\cdot n_i^{\varphi(x_i)}}{((n_i+\varphi(x_i)-1)!} \\
&= \Big(\lim_{n\to\infty} \frac{(n-1)!}{(n+K-1)!\cdot n^K}\Big) \cdot \Big(\prod_i \lim_{n_i\to\infty} \frac{(n_i-1)!\cdot n_i^{\varphi(x_i)}}{((n_i+\varphi(x_i)-1)!}\Big) \\
&\overset{(3.32)}{=} 1.
\end{aligned}$$

❑

There is another point of analogy with multinomial distributions that we wish to elaborate, namely the (limit) behaviour when balls are drawn of specific colours only, see Proposition 3.3.10. In the hypergeometric case it does not make sense to look at limit behaviour since after a certain number of steps the urn is empty. But in the Pólya case one can continue drawing indefinitely, making not trivial what happens in the limit. We use the following auxiliary result[2].

Lemma 3.6.6 *Let $0 < N < M$ be given. Define for $n \in \mathbb{N}$,*

$$a_n := \prod_{i<n} \frac{N+i}{M+i}.$$

Then: $\lim_{n\to\infty} a_n = 0$.

Proof. We switch to the (natural) logarithm ln and prove the equivalent statement $\lim_{n\to\infty} \ln(a_n) = -\infty$. We use that the logarithm turns products into sums, see Exercise 1.4.2, and that the derivative of $\ln(x)$ is $\frac{1}{x}$. Then:

$$\begin{aligned}
\ln(a_n) &= \sum_{i<n} \ln\Big(\frac{N+i}{M+i}\Big) = \sum_{i<n} \ln(N+i) - \ln(M+i) \\
&= -\sum_{i<n} \int_{N+1}^{M+1} \frac{1}{x}\,\mathrm{d}x \\
&\overset{(*)}{\leq} -\sum_{i<n} \frac{(M+i)-(N+i)}{M+i} = (N-M)\cdot\sum_{i<n} \frac{1}{M+i}.
\end{aligned}$$

It is well known that the harmonic series $\sum_{n>0} \frac{1}{n}$ is infinite. Since $M > N$ the above sequence $\ln(a_n)$ thus goes to $-\infty$.

The validity of the marked inequality $\leq$ follows from an inspection of the graph of the function $\frac{1}{x}$: the integral from $N+i$ to $M+i$ is the surface under $\frac{1}{x}$ between the points $N+i < M+i$. Since $\frac{1}{x}$ is a decreasing function, this surface is bigger than the rectangle with height $\frac{1}{M+i}$ and length $(M+i)-(N+i)$. ❑

Proposition 3.6.7 *Consider an urn $\upsilon \in \mathcal{N}[L](X)$ with a proper non-empty subset $S \subseteq supp(\upsilon)$.*

[2] With thanks to Bas and Bram Westerbaan for help.

1 *Write for $K \leq L$,*

$$H_K := \sum_{\varphi \in \mathcal{N}[K](S),\, \varphi \leq \upsilon} hg[K](\upsilon)(\varphi).$$

Then $H_K > H_{K+1}$; this stops at $K = L$, when the urn is empty.

2 *Now write, for arbitrary $K \in \mathbb{N}$,*

$$P_K := \sum_{\varphi \in \mathcal{N}[K](S)} pl[K](\upsilon)(\varphi).$$

Then $P_K > P_{K+1}$ and $\lim_{K\to\infty} P_K = 0$.

Proof. We write $L_S := \sum_{x\in S} \upsilon(x)$ is for the number of balls in the urn whose colour is in S.

1 By separating S and its complement $\neg S$ we can write, via Vandermonde's formula from Lemma 1.8.2,

$$H_K = \sum_{\varphi \in \mathcal{N}[K](S),\, \varphi \leq \upsilon} \frac{\left(\prod_{x\in S} \binom{\upsilon(x)}{\varphi(x)}\right) \cdot \left(\prod_{x\notin S} \binom{\upsilon(x)}{0}\right)}{\binom{L}{K}} = \frac{\binom{L_S}{K}}{\binom{L}{K}} = \frac{L_S!}{L!} \cdot \frac{(L-K)!}{(L_S-K)!}.$$

Using a similar description for H_{K+1} we get:

$$H_K > H_{K+1} \iff \frac{L_S!}{L!} \cdot \frac{(L-K)!}{(L_S-K)!} > \frac{L_S!}{L!} \cdot \frac{(L-(K+1))!}{(L_S-(K+1))!} \iff \frac{L-K}{L_S-K} > 1.$$

The latter holds because $L > L_S$, since S is a proper subset of $supp(\upsilon)$.

2 In the Pólya case we get, as before, but now via the Vandermonde formula for multichoose, see Proposition 1.8.6,

$$P_K = \frac{\left(\!\binom{L_S}{K}\!\right)}{\left(\!\binom{L}{K}\!\right)} = \frac{(L-1)!}{(L_S-1)!} \cdot \frac{(L_S+K-1)!}{(K+L-1)!}.$$

We define:

$$\begin{aligned} a_K := \frac{P_{K+1}}{P_K} &= \frac{(L-1)!}{(L_S-1)!} \cdot \frac{(L_S+(K+1)-1)!}{(L+(K+1)-1)!} \cdot \frac{(L_S-1)!}{(L-1)!} \cdot \frac{(K+L-1)!}{(L_S+K-1)!} \\ &= \frac{L_S+K}{K+L} < 1, \qquad \text{since } L_S < L. \end{aligned}$$

Thus $P_{K+1} = a_K \cdot P_K < P_K$ and also:

$$P_K = a_{K-1} \cdot P_{K-1} = a_{K-1} \cdot a_{K-2} \cdot P_{K-2} = \cdots = a_{K-1} \cdot a_{K-2} \cdot \ldots \cdot a_0 \cdot P_0.$$

Our goal is to prove $\lim_{K\to\infty} P_K = 0$. This follows from $\lim_{K\to\infty} \prod_{i<K} a_i = 0$, which we obtain from Lemma 3.6.6. ❑

Exercises

3.6.1 This exercise is the Pólya analogue of Exercise 3.5.3 for hypergeometric distributions.

1 Show that $pl[K](L|x\rangle) = 1\big|K|x\rangle\rangle$, for $L \geq 1$.

2 And also, for $L \geq 1$,

$$\begin{tikzcd} \mathcal{N}[L](X) \arrow[r, bend left, "pl[1]"] \arrow[r, bend right, "Flrn"'] & \mathcal{D}(\mathcal{N}[1](X)) \arrow[d, equal, "\wr"] \\ & \mathcal{D}(X) \end{tikzcd}$$

3.6.2 Recall Exercises 3.3.8 and 3.5.5, and show for an urn $\upsilon \in \mathcal{N}[L](X)$ and elements $y, z \in X$, the following two points.

1 When $y \neq z$,

$$\sum_{\varphi \in \mathcal{N}[K](X)} pl[K](\upsilon)(\varphi) \cdot \varphi(y) \cdot \varphi(z) \;=\; K \cdot (K-1) \cdot Flrn(\upsilon)(y) \cdot \frac{\upsilon(z)}{L+1}.$$

2 Prove also:

$$\sum_{\varphi \in \mathcal{N}[K](X)} pl[K](\upsilon)(\varphi) \cdot \varphi(y)^2 \;=\; K \cdot Flrn(\upsilon)(y) \cdot \frac{(K-1) \cdot \upsilon(y) + (L+K)}{L+1}.$$

3 Use the previous two points to prove:

$$\begin{aligned} &\langle Flrn, Flrn \rangle_*\big(pl[K](\upsilon)\big) \\ &= \frac{L \cdot (K-1)}{(L+1) \cdot K} \cdot \big(Flrn(\upsilon) \otimes Flrn(\upsilon)\big) \;+\; \frac{L+K}{(L+1) \cdot K} \cdot \mathcal{D}(\Delta)\big(Flrn(\upsilon)\big). \end{aligned}$$

3.6.3 Fix a non-empty urn $\upsilon \in \mathcal{N}(X)$ and consider the associated channel $DA_\upsilon \colon \mathcal{M}[N](X) \to \mathcal{D}(\mathcal{M}[N+1](X))$, defined as:

$$DA_\upsilon(\varphi) \coloneqq \sum_{x \in supp(\upsilon+\varphi)} Flrn(\upsilon+\varphi)(x) \big| \varphi + 1|x\rangle\rangle.$$

Prove that for $K \geq 0$,

$$pl[K](\upsilon) \;=\; \big(\underbrace{DA_\upsilon \circ \cdots \circ DA_\upsilon}_{K \text{ times}}\big)(\mathbf{0}).$$

3.6.4 This exercise elaborates that draws from an urn excluding one particular colour can be expressed in binary form. This works for all three modes of drawing: multinomial, hypergeometric, and Pólya.

Let X be a set with at least two elements, and let $x \in X$ be an arbitrary but fixed element. We write $x^\perp$ for an element not in X. Assume $k \leq K$.

1 For $\omega \in \mathcal{D}(X)$ with $x \in supp(\omega)$, use the Multinomial Theorem (1.43) to show that:

$$\begin{aligned} &\sum_{\varphi \in \mathcal{N}[K-k](X-x)} mn[K](\omega)(k|x\rangle + \varphi) \\ &= mn[K]\big(\omega(x)|x\rangle + (1-\omega(x))|x^\perp\rangle\big)\big(k|x\rangle + (K-k)|x^\perp\rangle\big) \;=\; bn[K](\omega(x))(k). \end{aligned}$$

2 Prove, via Vandermonde's formula, that for be an urn υ of size $L \geq K$ one has:

$$\begin{aligned} &\sum_{\varphi \leq_{K-k} \upsilon - \upsilon(x)|x\rangle} hg[K](\upsilon)(k|x\rangle + \varphi) \\ &= hg[K]\big(\upsilon(x)|x\rangle + (L-\upsilon(x))|x^\perp\rangle\big)\big(k|x\rangle + (K-k)|x^\perp\rangle\big). \end{aligned}$$

3 Show again for an urn υ,

$$\begin{aligned}&\sum_{\varphi\in\mathcal{N}[K-k](supp(\upsilon)-x)} pl[K](\upsilon)(k|x\rangle + \varphi)\\ &= pl[K]\Big(\upsilon(x)|x\rangle + (L-\upsilon(x))|x^{\perp}\rangle\Big)\Big(k|x\rangle + (K-k)|x^{\perp}\rangle\Big).\end{aligned}$$

3.6.5 This exercise develops a 'sequence' variant of Pólya draws, in analogy with the projection-add map PA from Definition 3.2.3. We call this map PSA, for 'projection-store-add'. It takes the form:

$$X^N \times X^L \xrightarrow{PSA} \mathcal{D}(X^{N+1} \times X^{L+1})$$

It is defined on $\vec{x} \in X^N$ and $\vec{y} \in X^L$,

$$PSA(\vec{x},\vec{y}) := \sum_{1\leq k\leq L}\ \sum_{1\leq i\leq L+1}\ \sum_{1\leq j\leq N+1} \frac{1}{L(L+1)(N+1)}\Big|\begin{array}{l}(x_1,\ldots,x_{i-1},y_k,x_i,\ldots,x_L),\\ (y_1,\ldots,y_{j-1},y_k,y_j,\ldots,y_N)\rangle.\end{array}$$

1 Show that the following diagram commutes.

$$\begin{tikzcd} X^N \times X^L \arrow[r, "PSA"] \arrow[d, "acc\times acc"'] & X^{N+1} \times X^{L+1} \arrow[d, "acc\times acc"] \\ \mathcal{N}[N](X) \times \mathcal{N}[L](X) \arrow[r, "PLSA"] & \mathcal{N}[N+1](X) \times \mathcal{N}[L+1](X) \end{tikzcd}$$

where $PLSA(\psi,\upsilon) := \sum_{x\in supp(\upsilon)} Flrn(\upsilon)(x)\big|\psi + 1|x\rangle, \upsilon + 1|x\rangle\rangle$.

2 Check that the (channel) iteration for $K \geq 1$ is:

$$PLSA^K(\psi,\upsilon) = \sum_{\varphi\in\mathcal{N}[K](supp(\upsilon))} \frac{\left(\!\binom{\upsilon}{\varphi}\!\right)}{\left(\!\binom{L+K}{K}\!\right)}\big|\psi+\varphi, \upsilon+\varphi\rangle.$$

3 Conclude that we obtain the Pólya-store map from (3.31) as:

$$pls[K](\upsilon) = PLSA^K(\mathbf{0},\upsilon),$$

and thus:

$$pl[K](\upsilon) = PLSA^K(\mathbf{0},\upsilon)[1,0].$$

3.6.6 Entropy $H(\omega)$ for a distribution ω is introduced in Definition 2.10.7. Entropy can also be defined for a non-empty natural multiset $\varphi \in \mathcal{N}(X)$, see *e.g.* (Bishop and Bishop, 2024, §2.5.2), namely as:

$$H(\varphi) := \frac{\ln\big(\!(\varphi)\!\big)}{\|\varphi\|} \quad \text{where, recall,} \quad (\!(\varphi)\!) = \frac{\|\varphi\|!}{\prod_x \varphi(x)!}.$$

Use Stirling's approximations $\ln(n!) \sim n\cdot\ln(n) - n$ for large n to prove that for a large multiset φ, its entropy as defined above, equals the entropy $H(Flrn(\varphi))$ of the associated distribution $Flrn(\varphi)$, obtained via frequentist learning $Flrn$.

3.7 The parallel multinomial law

We have already seen the close connection between multisets and distributions. This section focuses on a very special 'distributivity' relation between them. It shows how a (natural) multiset of distributions can be transformed into a distribution over multisets. This is a rather complicated operation, but it is a fundamental one. It will be called the parallel multinomial law, written as *pml*.

This operation *pml* is an instance of what is called a *distributive law* in category theory. It has popped up, without an explicit description, in Keimel and Plotkin (2017); Dahlqvist et al. (2018) and also in Dash and Staton (2020), for continuous probability, to describe 'point processes' as distributions over multisets (and lists), see also Section 3.10. The explicit descriptions of the law that we use below come from Jacobs (2021a), where the name parallel multinomial law was introduced. This law satisfies several elementary properties that combine basic elements of discrete probability theory.

The parallel multinomial law *pml* that we are after has the following type. For a number $K \in \mathbb{N}$ and a set X it is a function:

$$\mathcal{N}[K]\big(\mathcal{D}(X)\big) \xrightarrow{pml[K]} \mathcal{D}\big(\mathcal{N}[K](X)\big). \tag{3.33}$$

The dependence of *pml* on K (and X) is often left implicit. Notice that *pml* can also be written as channel $\mathcal{N}[K](\mathcal{D}(X)) \rightsquigarrow \mathcal{N}[K](X)$. We shall frequently encounter it in this form in commuting diagrams.

This map *pml* turns a K-element multiset of distributions over X into a distribution over K-element multisets over X. It is not immediately clear how to do this. It turns out that there are several ways to describe *pml*, as described in Jacobs (2021a). Here we use a slightly more abstract approach, starting from the characterisation of accumulation as a coequaliser, from Section 2.11. The fact that there are multiple equivalent formulations of the same law allows us to switch freely and use whichever formulation is most convenient in a particular situation.

Definition 3.7.1 Let X be an arbitrary set and K a natural number. We use the coequaliser property of accumulation, see Proposition 2.11.3 (1) to introduce a map *pml* as the unique (dashed) map in:

$$\begin{tikzcd}
\mathcal{D}(X)^K \arrow[r, bend left, "\text{transpositions}"] \arrow[r, bend right] & \mathcal{D}(X)^K \arrow[r, "acc"] \arrow[dr, "\bigotimes"'] & \mathcal{N}[K](\mathcal{D}(X)) \arrow[dd, dashed, "pml"] \\
& & \\
& \mathcal{D}(X^K) \arrow[dr, "\mathcal{D}(acc)"'] & \\
& & \mathcal{D}(\mathcal{N}[K](X))
\end{tikzcd} \tag{3.34}$$

This description exploits that the map $\mathcal{D}(acc) \circ \bigotimes \colon \mathcal{D}(X)^K \to \mathcal{D}(\mathcal{N}[K](X))$ is stable under transposition, as observed in Exercise 2.11.4.

This definition includes $K = 0$, when $\mathcal{N}[K](\mathcal{D}(X))$ contains only the empty multiset $\mathbf{0}$, for which we get $pml(\mathbf{0}) = 1|\mathbf{0}\rangle$.

The diagram in this definition already occcurred in Exercise 2.11.4, but there we did not introduce a name yet for the map that we now call *pml*. The above definition is quite abstract, but it does allow us to derive a more concrete descriptions below.

In Kozen and Silva (2024) the coequaliser property (3.34) is used in a slightly different way,

namely not for $\mathcal{D}$ but for $\mathcal{M}$, that is for multisets with multiplicities in $\mathbb{R}_{\geq 0}$. This yields maps $\mathcal{N}[K](\mathcal{M}(X)) \to \mathcal{M}(\mathcal{N}[K](X))$, which are shown to (1) form a distributive law $\mathcal{N}\mathcal{M} \Rightarrow \mathcal{M}\mathcal{N}$ of monads, and (2) restrict to $\mathcal{N}\mathcal{D} \Rightarrow \mathcal{D}\mathcal{N}$. This is a variation to obtain *pml*. We postpone its role as distributive law of monads to Section 3.9.

Proposition 3.7.2 *In the situation of Definition 3.7.1,*

1 The map pml *is determined by the following diagram:*

$$\begin{tikzcd}
\mathcal{N}[K](\mathcal{D}(X)) \arrow[rr, "pml"] \arrow[d, "arr"'] & & \mathcal{D}(\mathcal{N}[K](X)) \\
\mathcal{D}(\mathcal{D}(X)^K) \arrow[r, "\mathcal{D}(\bigotimes)"'] & \mathcal{D}^2(X^K) \arrow[r, "flat"'] & \mathcal{D}(X^K) \arrow[u, "\mathcal{D}(acc)"']
\end{tikzcd}$$

Expressed in terms of channel composition $\circledcirc$*, this becomes:*

$$pml \;=\; acc \circledcirc \bigotimes \circledcirc arr. \tag{3.35}$$

2 The parallel multinomial law pml *is the unique channel making both rectangles below commute.*

$$\begin{tikzcd}
\mathcal{D}(X)^K \arrow[r, "acc", "\circ"'] \arrow[d, "\bigotimes"', "\circ"] & \mathcal{N}[K](\mathcal{D}(X)) \arrow[r, "arr", "\circ"'] \arrow[d, "pml", "\circ"'] & \mathcal{D}(X)^K \arrow[d, "\circ", "\bigotimes"'] \\
X^K \arrow[r, "acc", "\circ"'] & \mathcal{N}[K](X) \arrow[r, "arr", "\circ"'] & X^K
\end{tikzcd}$$

Recall from Lemma 2.11.4 that we write transp $=$ arr $\circledcirc$ acc *for the transposition idempotent that occurs twice as the horizontal composite in this diagram.*

The description of *pml* in (3.35) follows the same patter that we used earlier for multizip in Diagram (3.1): first turn a multiset into sequences, apply a function (like zip, or tensor), and then accumulate the resulting sequences back into multisets. The second result shows that *pml* is squeezed between $\bigotimes$, both on the left and on the right.

Proof. 1 We use naturality of *flat* and *unit* and also the monad laws in:

$$\begin{aligned}
\mathcal{D}(acc) \circ flat \circ \mathcal{D}(\textstyle\bigotimes) \circ arr &= flat \circ \mathcal{D}^2(acc) \circ \mathcal{D}(\textstyle\bigotimes) \circ arr \\
&\overset{(3.34)}{=} flat \circ \mathcal{D}(pml) \circ \mathcal{D}(acc) \circ arr \\
&\overset{(2.39)}{=} flat \circ \mathcal{D}(pml) \circ unit \;=\; flat \circ unit \circ pml \;=\; pml.
\end{aligned}$$

2 The rectangle on the left repeats the defining equation of *pml* in (3.34) and thus provides uniqueness. Commutation of the rectangle on the right is obtained as follows.

$$\begin{aligned}
arr \circledcirc pml \overset{(3.35)}{=} arr \circledcirc acc \circledcirc \textstyle\bigotimes \circledcirc arr &= \textstyle\bigotimes \circledcirc arr \circledcirc acc \circledcirc arr \qquad \text{by Exercise 2.11.5} \\
&\overset{(2.39)}{=} \textstyle\bigotimes \circledcirc arr.
\end{aligned}$$

❑

There is a more concrete description of the parallel multinomial law *pml* as a convolution of multionomials. This involves a sum of parallel multinomial distributions, which explains the name.

Proposition 3.7.3 *For distributions* $\omega_i \in \mathcal{D}(X)$ *and numbers* $n_i \geq 1$,

$$pml\Big(\textstyle\sum_i n_i|\omega_i\rangle\Big) \;=\; \textstyle\sum_i\; mn[n_i](\omega_i), \tag{3.36}$$

where the sum $\sum_i$ *on the right-hand-side is the convolution with respect to the commutative monoid* $\mathcal{N}(X)$*, using that* $\mathrm{mn}[n_i](\omega_i) \in \mathcal{D}(\mathcal{N}(X))$.

Proof. Let's assume the index i runs over $\{1, \ldots, m\}$. Then:

$$\begin{aligned} pml\big(\textstyle\sum_i n_i|\omega_i\rangle\big) &= pml\big(acc(\underbrace{\omega_1, \ldots, \omega_1}_{n_1 \text{ times}}, \ldots, \underbrace{\omega_m, \ldots, \omega_m}_{n_m \text{ times}})\big) \\ &\overset{(3.34)}{=} \mathcal{D}(acc)\big(\bigotimes(\omega_1, \ldots, \omega_1, \ldots, \omega_m, \ldots, \omega_m)\big) \\ &= \mathcal{D}(acc)\big(\omega_1^{n_1} \otimes \cdots \otimes \omega_m^{n_m}\big) = \mathrm{mn}[n_1](\omega_1) + \cdots + \mathrm{mn}[n_m](\omega_m). \end{aligned}$$

The latter equation holds by (a generalisation of) Exercise 3.3.10 (3). ❑

The formulation in this last result will be used in the next illustration.

Example 3.7.4 Let $X = \{a, b\}$ be a sample space with two distributions $\omega, \rho \in \mathcal{D}(X)$, given by:

$$\omega = \tfrac{1}{3}|a\rangle + \tfrac{2}{3}|b\rangle \qquad \text{and} \qquad \rho = \tfrac{3}{4}|a\rangle + \tfrac{1}{4}|b\rangle. \tag{3.37}$$

We will define *pml* on the multiset of distributions $2|\omega\rangle + 1|\rho\rangle$ of size $K = 3$. The result should be a distribution on multisets of size $K = 3$ over X. There are four such multisets, namely:

$$3|a\rangle \qquad 2|a\rangle + 1|b\rangle \qquad 1|a\rangle + 2|b\rangle \qquad 3|b\rangle.$$

The goal is to assign a probability to each of them. The map *pml* does this in the following way.

$$\begin{aligned} pml\big(2|\omega\rangle + 1|\rho\rangle\big) &\overset{(3.36)}{=} \mathrm{mn}[2](\omega) + \mathrm{mn}[1](\rho) \\ &\overset{(2.60)}{=} \sum_{\varphi \in \mathcal{N}[2](X), \psi \in \mathcal{N}[1](X)} \mathrm{mn}[2](\omega)(\varphi) \cdot \mathrm{mn}[1](\rho)(\psi) \,\Big|\, \varphi + \psi \Big\rangle \\ &= \mathrm{mn}[2](\omega)(2|a\rangle) \cdot \mathrm{mn}[1](\rho)(1|a\rangle) \,\Big|\, 3|a\rangle \Big\rangle \\ &\quad + \Big(\mathrm{mn}[2](\omega)(2|a\rangle) \cdot \mathrm{mn}[1](\rho)(1|b\rangle) \\ &\qquad + \mathrm{mn}[2](\omega)(1|a\rangle + 1|b\rangle) \cdot \mathrm{mn}[1](\rho)(1|a\rangle)\Big) \,\Big|\, 2|a\rangle + 1|b\rangle \Big\rangle \\ &\quad + \Big(\mathrm{mn}[2](\omega)(1|a\rangle + 1|b\rangle) \cdot \mathrm{mn}[1](\rho)(1|b\rangle) \\ &\qquad + \mathrm{mn}[2](\omega)(2|b\rangle) \cdot \mathrm{mn}[1](\rho)(1|a\rangle)\Big) \,\Big|\, 1|a\rangle + 2|b\rangle \Big\rangle \\ &\quad + \mathrm{mn}[2](\omega)(2|b\rangle) \cdot \mathrm{mn}[1](\rho)(1|b\rangle) \,\Big|\, 3|b\rangle \Big\rangle \\ &= \tbinom{2}{2} \cdot \omega(a)^2 \cdot \tbinom{1}{1,0} \cdot \rho(a) \,\Big|\, 3|a\rangle \Big\rangle \\ &\quad + \Big(\tbinom{2}{2} \cdot \omega(a)^2 \cdot \tbinom{1}{1} \cdot \rho(b) + \tbinom{2}{1} \cdot \omega(a) \cdot \omega(b) \cdot \tbinom{1}{1} \cdot \rho(a)\Big) \,\Big|\, 2|a\rangle + 1|b\rangle \Big\rangle \\ &\quad + \Big(\tbinom{2}{1} \cdot \omega(a) \cdot \omega(b) \cdot \tbinom{1}{1} \cdot \rho(b) + \tbinom{2}{2} \cdot \omega(b)^2 \cdot \tbinom{1}{1} \cdot \rho(a)\Big) \,\Big|\, 1|a\rangle + 2|b\rangle \Big\rangle \\ &\quad + \tbinom{2}{2} \cdot \omega(b)^2 \cdot \tbinom{1}{1} \cdot \rho(b) \,\Big|\, 3|b\rangle \Big\rangle \\ &= (\tfrac{1}{3})^2 \cdot \tfrac{3}{4} \,\Big|\, 3|a\rangle \Big\rangle + \Big((\tfrac{1}{3})^2 \cdot \tfrac{1}{4} + 2 \cdot \tfrac{1}{3} \cdot \tfrac{2}{3} \cdot \tfrac{3}{4}\Big) \,\Big|\, 2|a\rangle + 1|b\rangle \Big\rangle \\ &\quad + \Big(2 \cdot \tfrac{1}{3} \cdot \tfrac{2}{3} \cdot \tfrac{1}{4} + (\tfrac{2}{3})^2 \cdot \tfrac{3}{4}\Big) \,\Big|\, 1|a\rangle + 2|b\rangle \Big\rangle + (\tfrac{2}{3})^2 \cdot \tfrac{1}{4} \,\Big|\, 3|b\rangle \Big\rangle \\ &= \tfrac{1}{12} \,\Big|\, 3|a\rangle \Big\rangle + \tfrac{13}{36} \,\Big|\, 2|a\rangle + 1|b\rangle \Big\rangle + \tfrac{4}{9} \,\Big|\, 1|a\rangle + 2|b\rangle \Big\rangle + \tfrac{1}{9} \,\Big|\, 3|b\rangle \Big\rangle. \end{aligned}$$

We see that actually calculating the parallel multinomial law is quite involved, already on a simply input. However, we do have explicit descriptions (3.35) and (3.36) that can be implemented.

We discuss a next description of the parallel multinomial law *pml*, building on the commutative monoid structure on $\mathcal{N}(X)$, and thus on $\mathcal{D}(\mathcal{N}(X))$, that we used in Proposition 3.7.3. With multiple functor applications we often spare on parentheses and simply write $\mathcal{DN}(X)$ for $\mathcal{D}(\mathcal{N}(X))$. Proposition 1.6.6 this gives a $\mathcal{N}$-algebra $sum\colon \mathcal{N}\big(\mathcal{D}(\mathcal{N}(X))\big) \to \mathcal{D}(\mathcal{N}(X))$. We can describe it in terms of the (iterated) convolution sum $+$ on the diagonal in:

$$\begin{tikzcd} \mathcal{D}(\mathcal{N}(X))^K \arrow[r, "acc"] \arrow[d, "\bigotimes"'] \arrow[dr, "+"] & \mathcal{N}\big(\mathcal{D}(\mathcal{N}(X))\big) \arrow[d, "sum"] \\ \mathcal{D}(\mathcal{N}(X)^K) \arrow[r, "\mathcal{D}(+)"'] & \mathcal{D}(\mathcal{N}(X)) \end{tikzcd} \tag{3.38}$$

This diagonal is expressed via the multiset-sum $+$ and the tensor, by definition of convolution (2.60).

Proposition 3.7.5 *In the context just described:*

1 one can describe the parallel multinomial law as:

$$pml \;=\; \Big(\mathcal{ND}(X) \xrightarrow{\mathcal{ND}(unit)} \mathcal{NDN}(X) \xrightarrow{sum} \mathcal{DN}(X)\Big). \tag{3.39}$$

The unit in this diagram is the singleton-multiset function $unit\colon X \to \mathcal{N}(X)$*, satisfying* $unit(x) = 1|x\rangle$.

2 The map $pml\colon \mathcal{ND}(X) \to \mathcal{DN}(X)$ *is a homomorphism of monoids.*

Proof. 1 We use the property that uniquely defines *pml* to prove that $pml = sum \circ \mathcal{ND}(unit)$.

$$\begin{aligned} sum \circ \mathcal{ND}(unit) \circ acc &= sum \circ acc \circ \mathcal{D}(unit)^K && \text{by naturality of } acc \\ &= \mathcal{D}(+) \circ \bigotimes \circ \mathcal{D}(unit)^K && \text{see (3.38)} \\ &= \mathcal{D}(+) \circ \mathcal{D}(unit^K) \circ \bigotimes && \text{by naturality of } \bigotimes \\ &= \mathcal{D}(acc) \circ \bigotimes && \text{by Exercise 1.7.3.} \end{aligned}$$

2 One has $pml(\mathbf{1}) = 1|\mathbf{1}\rangle$, see Definition 3.7.1. We use that $\mathcal{ND}(unit)$ is a map of monoids, see Lemma 1.6.3 (3), and the function *sum* in (3.38) too, by construction. Then, for $\varphi, \psi \in \mathcal{ND}(X)$,

$$\begin{aligned} pml(\varphi + \psi) \;=\; sum\Big(\mathcal{ND}(unit)(\varphi + \psi)\Big) &= sum\Big(\mathcal{ND}(unit)(\varphi) + \mathcal{ND}(unit)(\psi)\Big) \\ &= sum\Big(\mathcal{ND}(unit)(\varphi)\Big) + sum\Big(\mathcal{ND}(unit)(\psi)\Big) \\ &= pml(\varphi) + pml(\psi). \end{aligned}$$

❑

Exercises

3.7.1 Check that the multinomial distribution can be obtained via the parallel multinomial law, in two different ways.

1 Use the convolution formulation (3.36) of *pml* to show that for a distribution $\omega \in \mathcal{D}(X)$ and number $K \in \mathbb{N}$ one has:

$$mn[K](\omega) \;=\; pml(K|\omega\rangle),$$

that is:

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[rr, "mn[K]"] \arrow[dr, "K\cdot unit"'] & & \mathcal{D}(\mathcal{N}[K](X)) \\
& \mathcal{N}[K](\mathcal{D}(X)) \arrow[ur, "pml"'] &
\end{tikzcd}$$

2 Prove the same thing via the definition (3.34) of *pml*, and via Theorem 2.6.7.

3.7.2 Apply the multiset flatten map $flat\colon \mathcal{M}(\mathcal{M}(X)) \to \mathcal{M}(X)$ in the setting of Example 3.7.4 to show that:

$$flat\Big(2|\omega\rangle + 1|\rho\rangle\Big) = \tfrac{17}{12}|a\rangle + \tfrac{19}{12}|b\rangle = flat\Big(pml(2|\omega\rangle + 1|\rho\rangle)\Big).$$

(The general formulation appears later on in Proposition 3.8.2.)

3.7.3 We claim *pml* is natural: for each function $f\colon X \to Y$ the following diagram commutes.

$$\begin{tikzcd}
\mathcal{N}[K](\mathcal{D}(X)) \arrow[r, "pml"] \arrow[d, "\mathcal{N}[K](\mathcal{D}(f))"'] & \mathcal{D}(\mathcal{N}[K](X)) \arrow[d, "\mathcal{D}(\mathcal{N}[K](f))"] \\
\mathcal{N}[K](\mathcal{D}(Y)) \arrow[r, "pml"] & \mathcal{D}(\mathcal{N}[K](Y)).
\end{tikzcd}$$

1 An abstract argument uses the uniqueness part of the definition (3.34), as suggested in the diagram:

$$\begin{tikzcd}
\mathcal{D}(X)^K \arrow[rrr, "acc"] \arrow[dr, "\mathcal{D}(f)^K"'] & & & \mathcal{N}[K](\mathcal{D}(X)) \arrow[ddd, dashed, bend left] \\
& \mathcal{D}(Y)^K \arrow[dr, "\otimes"'] & & \\
& & \mathcal{D}(Y^K) \arrow[dr, "\mathcal{D}(acc)"'] & \\
& & & \mathcal{D}(\mathcal{N}[K](Y))
\end{tikzcd}$$

2 Prove naturality of *pml* also via one (or all) of the formulations (3.35), (3.36) and (3.39).

3.7.4 Check that the sum-preservation in Proposition 3.7.5 (2) can (also) be expressed via the following diagram.

$$\begin{tikzcd}
\mathcal{N}[K](\mathcal{D}(X)) \times \mathcal{N}[L](\mathcal{D}(X)) \arrow[r, "pml\otimes pml"] \arrow[d, "+"'] & \mathcal{N}[K](X) \times \mathcal{N}[L](X) \arrow[d, "+"] \\
\mathcal{N}[K+L](\mathcal{D}(X)) \arrow[r, "pml"] & \mathcal{N}[K+L](X)
\end{tikzcd}$$

This generalises Exercise 3.3.10 from multinomials to parallel multinomials.

3.8 The parallel multinomial law: basic properties

This section continues the investigation of the parallel multinomial law *pml*, introduced in the previous section, in various forms. This section focuses on some key properties of this law, including its interaction with frequentist learning, multizip and hypergeometric distributions. These properties are expressed in the language of category theory.

As we have seen, actual calculations with the parallel multinomial law *pml* quickly grow out of hand. So we might worry that proving properties also becomes tedious. But abstraction will

help us. Since there is a characterisation of *pml* with a uniqueness property, in Diagram (3.34), we can reason with the associated uniqueness proof principle. In most general form it says that $f = g$ follows from $f \circ acc = g \circ acc$, where acc is the accumulation map. This uses the coequaliser property of accumulation, see Proposition 2.11.3 (1).

First we show how *pml* interacts with frequentist learning.

Theorem 3.8.1 *The distributive law* pml *commutes with frequentist learning, in the sense that for* $\Psi \in \mathcal{N}[K](\mathcal{D}(X))$*,*

$$Flrn_*(pml(\Psi)) = flat(Flrn(\Psi)).$$

Equivalently, in diagrammatic form:

$$\begin{tikzcd}
\mathcal{N}[K](\mathcal{D}(X)) \arrow[r, "pml", "\circ"'] \arrow[d, "Flrn"', "\circ"] & \mathcal{N}[K](X) \arrow[d, "Flrn", "\circ"'] \\
\mathcal{D}(X) \arrow[r, "sam", "\circ"'] & X
\end{tikzcd}$$

The channel sam$\colon \mathcal{D}(X) \rightsquigarrow X$ *at the bottom is the identity function* $\mathcal{D}(X) \to \mathcal{D}(X)$*, used as sample operation.*

Proof. We use the formulation $pml = acc \circ \bigotimes \circ arr$ from (3.35) in the following diagram, using the uniform projection channel *unpr*.

$$\begin{tikzcd}
\mathcal{N}[K](\mathcal{D}(X)) \arrow[r, "arr", "\circ"'] \arrow[dr, bend right, "Flrn"'] & \mathcal{D}(X)^K \arrow[r, "\bigotimes", "\circ"'] \arrow[d, "unpr", "\circ"'] & X^K \arrow[r, "acc", "\circ"'] \arrow[dr, "unpr", "\circ"'] & \mathcal{N}[K](X) \arrow[d, "Flrn", "\circ"'] \\
& \mathcal{D}(X) \arrow[rr, "sam", "\circ"'] & & X
\end{tikzcd}$$

The subdiagrams commute by Exercise 2.4.9. ❑

We include a result that generalises Exercise 3.7.2. It is a discrete version of (Dash and Staton, 2020, Lem. 13). The proof below uses the 'mean' of multinomials, from Proposition 3.3.7.

Proposition 3.8.2 *Consider inclusions* $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ *and* $\mathcal{N}[K](X) \hookrightarrow \mathcal{M}(X)$*, together with the multiset flatten map* flat$\colon \mathcal{M}(\mathcal{M}(X)) \to \mathcal{M}(X)$*. Via these inclusions, one has, for* $\Psi \in \mathcal{N}[K](\mathcal{D}(X))$*,*

$$flat(pml(\Psi)) = flat(\Psi).$$

Proof. For $\Psi = n_1|\omega_1\rangle + \cdots + n_k|\omega_k\rangle \in \mathcal{N}[K](\mathcal{D}(X))$,

$$\begin{aligned}
&\mathit{flat}(\mathit{pml}(\Psi)) \overset{(3.36)}{=} \mathit{flat}\Big(\textstyle\sum_i \mathit{mn}[n_i](\omega_i)\Big) \\
&= \mathit{flat}\left(\sum_{\varphi_1 \in \mathcal{N}[n_1](X), \ldots, \varphi_k \in \mathcal{N}[n_k](X)} \Big(\textstyle\prod_i \mathit{mn}[n_i](\omega)(\varphi_i)\Big) \,\Big|\, \textstyle\sum_i \varphi_i \rangle\right) \\
&= \sum_{\varphi_1 \in \mathcal{N}[n_1](X), \ldots, \varphi_k \in \mathcal{N}[n_k](X)} \Big(\textstyle\prod_i \mathit{mn}[n_i](\omega)(\varphi_i)\Big) \cdot \Big(\textstyle\sum_i \varphi_i\Big) \\
&= \sum_{\varphi_1 \in \mathcal{N}[n_1](X), \ldots, \varphi_k \in \mathcal{N}[n_k](X)} \Big(\textstyle\prod_i \mathit{mn}[n_i](\omega)(\varphi_i)\Big) \cdot \varphi_1 \\
&\qquad + \cdots + \sum_{\varphi_1 \in \mathcal{N}[n_1](X), \ldots, \varphi_k \in \mathcal{N}[n_k](X)} \Big(\textstyle\prod_i \mathit{mn}[n_i](\omega)(\varphi_i)\Big) \cdot \varphi_k \\
&= \sum_{\varphi_1 \in \mathcal{N}[n_1](X)} \mathit{mn}[n_1](\omega_1)(\varphi_1) \cdot \varphi_1 + \cdots + \sum_{\varphi_k \in \mathcal{N}[n_k](X)} \mathit{mn}[n_k](\omega_k)(\varphi_k) \cdot \varphi_k \\
&= n_1 \cdot \omega_1 + \cdots + n_k \cdot \omega_k \qquad \text{by Proposition 3.3.7} \\
&= \mathit{flat}(\Psi).
\end{aligned}$$

❑

As we have seen in Proposition 3.7.3, the parallel multinomial law *pml* can be expressed as convolution sum of multinomial distributions. But it also commutes with the multinomial channel, in several ways.

Theorem 3.8.3 *1 The parallel multinomial law* pml *commutes with multinomials in the following manner.*

$$\begin{tikzcd}
\mathcal{D}(\mathcal{D}(X)) \arrow[r, "\mathit{mn}[K]", "\circ"{marking}] \arrow[d, "\mathit{flat}"'] & \mathcal{N}[K](\mathcal{D}(X)) \arrow[d, "\mathit{pml}", "\circ"{marking}] \\
\mathcal{D}(X) \arrow[r, "\mathit{mn}[K]", "\circ"{marking}] & \mathcal{N}[K](X)
\end{tikzcd}$$

2 There is a second form of exchange between pml *and multinomials:*

$$\begin{tikzcd}
& \mathcal{D}\mathcal{N}[K](X) \arrow[r, "\mathit{mn}[L]"] & \mathcal{D}\mathcal{N}[L]\mathcal{N}[K](X) \arrow[dr, bend left, "\mathcal{D}(\mathit{flat})"] & \\
\mathcal{N}[K]\mathcal{D}(X) \arrow[ur, bend left, "\mathit{pml}"] \arrow[dr, bend right, "\mathcal{N}(\mathit{mn}[L])"'] & & & \mathcal{D}\mathcal{N}[L{\cdot}K](X) \\
& \mathcal{N}[K]\mathcal{D}\mathcal{N}[L](X) \arrow[r, "\mathit{pml}"'] & \mathcal{D}\mathcal{N}[K]\mathcal{N}[L](X) \arrow[ur, bend right, "\mathcal{D}(\mathit{flat})"'] &
\end{tikzcd}$$

Proof. 1 We use that $\mathit{mn}[K] = \mathit{acc} \circ \mathit{iid}[K]$, see Theorem 2.6.7, in:

$$\begin{tikzcd}
\mathcal{D}(\mathcal{D}(X)) \arrow[r, "\mathit{iid}", "\circ"{marking}] \arrow[d, "\mathit{flat}"'] & \mathcal{D}(X)^K \arrow[r, "\mathit{acc}", "\circ"{marking}] \arrow[d, "\circ"{marking}, "\otimes"] & \mathcal{N}[K](\mathcal{D}(X)) \arrow[d, "\mathit{pml}", "\circ"{marking}] \\
\mathcal{D}(X) \arrow[r, "\mathit{iid}", "\circ"{marking}] & X^K \arrow[r, "\mathit{acc}", "\circ"{marking}] & \mathcal{N}[K](X)
\end{tikzcd}$$

The rectangle on the left commutes by Exercise 2.4.10, and the one on the right by construction, see Definition 3.7.1.

2 We show that precomposing both legs in the diagram with the accumulation map $\mathit{acc}\colon \mathcal{D}(X)^K \to \mathcal{M}[K]\mathcal{D}(X)$ yields an equality. This suffices by Proposition 2.11.3 (1). We make essential use

of the commutation of accumulution and tensors with the swap isomorphism $(X^K)^L \xrightarrow{\cong} (X^L)^K$ from Exercises 1.7.18 and 2.4.10 (6).

$$\begin{aligned}
&\mathcal{D}(flat) \circ mn[L] \circ pml \circ acc \\
&\overset{(3.34)}{=} \mathcal{D}(flat) \circ \mathcal{D}(acc) \circ iid[L] \circ \mathcal{D}(acc) \circ \bigotimes[K] \\
&= \mathcal{D}(flat) \circ \mathcal{D}(acc) \circ \mathcal{D}(acc^L) \circ iid[L] \circ \bigotimes[K] \\
&= \mathcal{D}(flat) \circ \mathcal{D}(acc) \circ \mathcal{D}(acc^K) \circ \mathcal{D}(swap) \circ iid[L] \circ \bigotimes[K] \\
&= \mathcal{D}(flat) \circ \mathcal{D}(acc) \circ \mathcal{D}(acc^K) \circ \bigotimes[K] \circ iid[L]^K \\
&= \mathcal{D}(flat) \circ \mathcal{D}(acc) \circ \bigotimes[K] \circ \mathcal{D}(acc)^K \circ iid[L]^K \\
&\overset{(3.34)}{=} \mathcal{D}(flat) \circ pml \circ acc \circ mn[L]^K \;=\; \mathcal{D}(flat) \circ pml \circ \mathcal{N}(mn[L]) \circ acc. \qquad \square
\end{aligned}$$

We turn to *pml* and hypergeometric channels, which, as we have seen in Theorem 3.5.1, are composites of draw-and-delete channels. We know from Proposition 3.2.9 that multinomial channels commute with draw-and-delete. The same holds for the parallel multinomial law.

Proposition 3.8.4 *The parallel multinomial law commutes with draw-delete DD, as in:*

$$\begin{tikzcd}
\mathcal{N}[K](\mathcal{D}(X)) \arrow[d, "pml"'] & \mathcal{N}[K+1](\mathcal{D}(X)) \arrow[l, "DD"'] \arrow[d, "pml"] \\
\mathcal{N}[K](X) & \mathcal{N}[K+1](X) \arrow[l, "DD"']
\end{tikzcd}$$

Proof. We use the probabilistic projection channel $PD \colon X^{K+1} \rightsquigarrow X^K$ from Definition 3.2.1 and its interaction with $\bigotimes$ in Exercise 3.2.10.

$$\begin{aligned}
DD \circ pml \overset{(3.35)}{=} DD \circ acc \circ \bigotimes \circ arr &= acc \circ PD \circ \bigotimes \circ arr && \text{by Lemma 3.2.2} \\
&= acc \circ \bigotimes \circ PD \circ arr && \text{by Exercise 3.2.10} \\
&= pml \circ acc \circ PD \circ arr && \text{by (3.34)} \\
&= pml \circ DD \circ acc \circ arr && \text{by Lemma 3.2.2} \\
&= pml \circ DD && \text{by (2.39).} \qquad \square
\end{aligned}$$

Corollary 3.8.5 *The parallel multinomial law commutes with the hypergeometric channel: for $L \geq K$ one has:*

$$\begin{tikzcd}
\mathcal{N}[L](\mathcal{D}(X)) \arrow[r, "hg[K]"] \arrow[d, "pml"'] & \mathcal{N}[K](\mathcal{D}(X)) \arrow[d, "pml"] \\
\mathcal{N}[L](X) \arrow[r, "hg[K]"] & \mathcal{N}[K](X)
\end{tikzcd}$$

Proof. Theorem 3.5.1 shows that the hypergeometric distribution can be expressed via iterated draw-and-deletes. Hence the result follows from (iterated application of) Proposition 3.8.4. $\square$

We continue to show that the parallel multinomial law *pml* commutes with the unit and flatten operations of the distribution monad. This shows that *pml* is an instance of what is called a distributive law in category theory. Such laws are important in combining different forms of computation. A notorious result, noted around 2000 by Gordon Plotkin, is that the powerset monad $\mathcal{P}$ does not distribute over the probability distributions monad $\mathcal{D}$. Plotkin never published this important no-go result himself. Instead, it appeared in Varacca (2003); Varacca and Winskel (2006) (with full credits), see also Zwart and Marsden (2019); Zwart (2020). This negative result

is interpreted as: there is no semantically solid way to combine non-deterministic and probabilistic computation. The fact that a distributive law for multisets and distributions does exist shows that multiset-computations and probability can be combined. Indeed, in Corollary 3.8.7 we shall see that the K-sized multiset functor $\mathcal{N}[K]$ can be 'extended' to the category of probabilistic channels.

But first we have to show that *pml* is a distributive law. We have already seen in Exercise 3.7.3 that it is natural.

Proposition 3.8.6 *The parallel multinomial law* pml *is a* distributive law *of the K-sized multiset functor $\mathcal{N}[K]$ over the distribution monad $\mathcal{D}$. This means that* pml *commutes with the unit and flatten operations of $\mathcal{D}$, as expressed by the following two diagrams.*

$$\begin{tikzcd} \mathcal{N}[K](X) \arrow[d, "\mathcal{N}(unit)"'] \arrow[dr, bend left, "unit"] & \\ \mathcal{N}[K](\mathcal{D}(X)) \arrow[r, "pml"] & \mathcal{D}(\mathcal{N}[K](X)) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{N}[K](\mathcal{D}^2(X)) \arrow[r, "pml"] \arrow[d, "\mathcal{N}(flat)"'] & \mathcal{D}(\mathcal{N}[K](\mathcal{D}(X))) \arrow[r, "\mathcal{D}(pml)"] & \mathcal{D}^2(\mathcal{N}[K](X)) \arrow[d, "flat"] \\ \mathcal{N}[K](\mathcal{D}(X)) \arrow[rr, "pml"] & & \mathcal{D}(\mathcal{N}[K](X)) \end{tikzcd}$$

Proof. In Exercise 2.4.11 we have seen that the big tensor $\bigotimes\colon \mathcal{D}(X)^K \to \mathcal{D}(X^K)$ is a distributive law. These properties will be used to show that *pml* is a distributive law too. We exploit the uniqueness property from 3.34.

$$\begin{aligned} pml \circ \mathcal{N}(unit) \circ acc &= pml \circ acc \circ unit^K && \text{by naturality of } acc\text{, see Exercise 1.7.15} \\ &= \mathcal{D}(acc) \circ \bigotimes \circ unit^K && \text{by (3.34)} \\ &= \mathcal{D}(acc) \circ unit && \text{via the first diagram in Exercise 2.4.11} \\ &= unit \circ acc && \text{by naturality of } unit. \end{aligned}$$

Similarly for the flatten-diagram:

$$\begin{aligned} flat \circ \mathcal{D}(pml) \circ pml \circ acc &= flat \circ \mathcal{D}(pml) \circ \mathcal{D}(acc) \circ \bigotimes && \text{by (3.34)} \\ &= flat \circ \mathcal{D}(\mathcal{D}(acc)) \circ \mathcal{D}(\bigotimes) \circ \bigotimes && \text{again by (3.34)} \\ &= \mathcal{D}(acc) \circ flat \circ \mathcal{D}(\bigotimes) \circ \bigotimes && \text{by naturality of } flat \\ &= \mathcal{D}(acc) \circ \bigotimes \circ flat^K && \text{via Exercise 2.4.11} \\ &= pml \circ acc \circ flat^K && \text{once again by (3.34)} \\ &= pml \circ \mathcal{N}(flat) \circ acc && \text{by naturality of } acc. \end{aligned}$$

❑

This result says that *pml* is a so-called $\mathcal{K}\ell$-law, of the functor $\mathcal{N}[K]$ over the monad $\mathcal{D}$. In general such a $\mathcal{K}\ell$-law corresponds to an extension of the functor to the Kleisli category **Chan** = **Chan**($\mathcal{D}$) of channels for the monad, see Jacobs (2016b); Jacobs et al. (2015) for details.

Corollary 3.8.7 *The K-fold product $(-)^K\colon$* **Sets** $\to$ **Sets** *and the K-sized multiset functor $\mathcal{N}[K]\colon$* **Sets** $\to$ **Sets** *extend to functors $(-)^K\colon$* **Chan** $\to$ **Chan** *and $\mathcal{N}[K]\colon$* **Chan** $\to$ **Chan**, *in commuting diagrams:*

$$\begin{tikzcd} \mathbf{Chan} \arrow[r, "(-)^K"] & \mathbf{Chan} \\ \mathbf{Sets} \arrow[u] \arrow[r, "(-)^K"] & \mathbf{Sets} \arrow[u] \end{tikzcd} \qquad \begin{tikzcd} \mathbf{Chan} \arrow[r, "\mathcal{N}[K]"] & \mathbf{Chan} \\ \mathbf{Sets} \arrow[u] \arrow[r, "\mathcal{N}[K]"] & \mathbf{Sets} \arrow[u] \end{tikzcd} \tag{3.40}$$

Both versions of $(-)^K$ and of $\mathcal{N}[K]$ coincide on objects. On a channel / morphism $c\colon X \rightarrowtail Y$ the

extended functor $(-)^K \colon \mathbf{Chan} \rightsquigarrow \mathbf{Chan}$ *is defined as:*

$$c^K \;:=\; \Big(X^K \xrightarrow{c \times \cdots \times c} \mathcal{D}(Y) \times \cdots \times \mathcal{D}(Y) \xrightarrow{\bigotimes} \mathcal{D}(Y^K)\Big). \tag{3.41}$$

The extension $\mathcal{N}[K] \colon \mathbf{Chan} \to \mathbf{Chan}$ *on a channel* $c \colon X \rightsquigarrow Y$ *be described in two (equivalent) ways as:*

$$\mathcal{N}[K](c) \;:=\; \left(\begin{tikzcd} & \mathcal{N}[K](\mathcal{D}(Y)) \arrow[dr, "pml", "\circ"'] & \\ \mathcal{N}[K](X) \arrow[ur, bend left, "\mathcal{N}(c)"] \arrow[dr, "arr", "\circ"'] & & \mathcal{N}[K](Y) \\ & X^K \arrow[r, "c^K", "\circ"'] & Y^K \arrow[ur, "acc", "\circ"'] \end{tikzcd} \right) \tag{3.42}$$

The fact that we use the same notation $\mathcal{N}[K]$ for two different functors may be confusing, but usually the context will tell which one is meant. When confusion is likely, we may drop the parameter K for the multiset functor $\mathcal{N} \colon \mathbf{Sets} \to \mathbf{Sets}$, as in the above diagram (3.42). The same confusion may arise for c^K, since it may mean $c \times \cdots \times c$ and $c \otimes \cdots \otimes c$. Some authors write $c^{\otimes K}$ for the latter parallel product.

Proof. The fact that products extend to channels has already occurred, for instance in (2.37) and Exercise 2.4.10 (3). Commutation of the above rectangle on the left in (3.40) is trivial on sets / objects. Now let g be a function, that is, a morphism in **Sets**. The upgoing functors in (3.40) sends g to the deterministic channel $\langle g \rangle = unit \circ g$. We have:

$$\begin{aligned} \langle g \rangle^K &\overset{(3.41)}{=} \bigotimes \circ \Big((unit \circ g) \times \cdots \times (unit \circ g)\Big) \\ &= \bigotimes \circ (unit \times \cdots \times unit) \circ (g \times \cdots \times g) \\ &= (unit \otimes \cdots \otimes unit) \circ (g \times \cdots \times g) \overset{(2.37)}{=} unit \circ (g \times \cdots \times g) \;=\; \langle g^K \rangle. \end{aligned}$$

The functor $\mathcal{N}[K] \colon \mathbf{Chan} \to \mathbf{Chan}$ is defined on objects / sets as $X \mapsto \mathcal{N}[K](X)$. A morphism $c \colon X \rightsquigarrow Y$ in **Chan** is sent to the channel $\mathcal{N}[K](X) \rightsquigarrow \mathcal{N}[K](Y)$, obtained in (3.42). We check that these two formulations coincide, via the formulation of *pml* in Equation (3.35) and naturality of arrangement:

$$\begin{aligned} pml \circ \mathcal{N}(c) &= (acc \circledcirc \bigotimes \circledcirc arr) \circ \mathcal{N}(c) \\ &= \mathcal{D}(acc) \circ flat \circ \mathcal{D}(\bigotimes) \circ arr \circ \mathcal{N}(c) \\ &= \mathcal{D}(acc) \circ flat \circ \mathcal{D}(\bigotimes) \circ \mathcal{D}(c \times \cdots \times c) \circ arr \\ &= \mathcal{D}(acc) \circ flat \circ \mathcal{D}(c \otimes \cdots \otimes c) \circ arr \;=\; acc \circledcirc c^K \circledcirc arr. \end{aligned}$$

This extension $\mathcal{N}[K]$ preseves identities and composition by Proposition 3.8.6.

The rectangle on the right in (3.40) commutes too, via the same proposition:

$$\mathcal{N}[K](\langle g \rangle) \;=\; pml \circ \mathcal{N}(unit \circ g) \;=\; pml \circ \mathcal{N}(unit) \circ \mathcal{N}(g) \;=\; unit \circ \mathcal{N}(g) \;=\; \langle \mathcal{N}(g) \rangle. \qquad \square$$

Next we show that parallel multinomials and multizip commute. This is a non-trivial technical result, which plays a crucial role in what follows.

Lemma 3.8.8 *The following diagram commutes.*

$$\begin{tikzcd}
\mathcal{N}[K](\mathcal{D}(X)) \times \mathcal{N}[K](\mathcal{D}(Y)) \arrow[r, "pml\otimes pml"] \arrow[d, "mzip"'] & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[dd, "mzip"] \\
\mathcal{N}[K](\mathcal{D}(X) \times \mathcal{D}(Y)) \arrow[d, "\mathcal{N}(\otimes)"'] & \\
\mathcal{N}[K](\mathcal{D}(X \times Y)) \arrow[r, "pml"] & \mathcal{N}[K](X \times Y)
\end{tikzcd}$$

Proof. The result follows from a big diagram chase in which the *mzip* operations on the left and on the right are expanded, according to (3.1).

$$\begin{tikzcd}
\mathcal{N}[K](\mathcal{D}(X)) \times \mathcal{N}[K](\mathcal{D}(Y)) \arrow[r, "pml\otimes pml"] \arrow[d, "arr\otimes arr"'] \arrow[ddd, bend right=60, "mzip"'] & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[d, "arr\otimes arr"] \arrow[ddd, bend left=60, "mzip"] \\
\mathcal{D}(X)^K \times \mathcal{D}(Y)^K \arrow[r, "\bigotimes\otimes\bigotimes"] \arrow[d, "zip"'] & X^K \times Y^K \arrow[d, "zip"] \\
(\mathcal{D}(X) \times \mathcal{D}(Y))^K \arrow[r, "\otimes^K"] \arrow[d, "acc"'] & \mathcal{D}(X \times Y)^K \arrow[r, "\bigotimes"] \arrow[dl, bend left, "acc"] & (X \times Y)^K \arrow[d, "acc"] \\
\mathcal{N}[K](\mathcal{D}(X) \times \mathcal{D}(Y)) \arrow[d, "\mathcal{N}(\otimes)"'] & & \\
\mathcal{N}[K](\mathcal{D}(X \times Y)) \arrow[rr, "pml"] & & \mathcal{N}[K](X \times Y)
\end{tikzcd}$$

The upper rectangle commutes by Proposition 3.7.2 (2) and the middle on by Lemma 2.3.5 (2). The lower-left subdiagram commutes by naturality of *acc* and the lower-right one via the definition of *pml* in (3.35). ❑

Theorem 3.8.9 *The extended functor* $\mathcal{N}[K]\colon \mathbf{Chan} \to \mathbf{Chan}$ *commutes with multizip: for channels* $f\colon X \rightsquigarrow U$ *and* $g\colon Y \rightsquigarrow V$ *one has:*

$$mzip \circ \big(\mathcal{N}[K](f) \otimes \mathcal{N}[K](g)\big) = \mathcal{N}[K](f \otimes g) \circ mzip. \tag{3.43}$$

Diagrammatically this amounts to:

$$\begin{tikzcd}
\mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[r, "mzip"] \arrow[d, "\mathcal{N}[K](f)\otimes\mathcal{N}[K](g)"'] & \mathcal{N}[K](X \times Y) \arrow[d, "\mathcal{N}[K](f\otimes g)"] \\
\mathcal{N}[K](U) \times \mathcal{N}[K](V) \arrow[r, "mzip"] & \mathcal{N}[K](U \times V)
\end{tikzcd}$$

In combination with the unit and associativity of Proposition 3.1.3 (2) and (5) this means that the extended functor $\mathcal{N}[K]\colon \mathbf{Chan} \to \mathbf{Chan}$ *is a monoidal functor, via* mzip.

Proof. We reason as follows.

$$\begin{aligned}
& mzip \circ \big(\mathcal{N}[K](f) \otimes \mathcal{N}[K](g)\big) \\
&= mzip \circ (pml \otimes pml) \circ (\mathcal{N}(f) \times \mathcal{N}(g)) \\
&= pml \circ \mathcal{N}(\otimes) \circ mzip \circ (\mathcal{N}(f) \times \mathcal{N}(g)) && \text{by Lemma 3.8.8} \\
&= pml \circ \mathcal{N}(\otimes) \circ \mathcal{N}(f \times g) \circ mzip && \text{by Proposition 3.1.3 (1)} \\
&= pml \circ \mathcal{N}(f \otimes g) \circ mzip \\
&= \mathcal{N}[K](f \otimes g) \circ mzip.
\end{aligned}$$

❑

For this result we really need the multizip operation *mzip*. One may think that one can use tensors $\otimes$ instead, but the tensor-version of Lemma 3.8.8 does not hold, see Exercise 3.8.8 below.

Earlier we have seen that accumulation, arrangement, and draw-delete are natural with respect to functions. We can now show that they are, more generally, natural with respect to channels. This takes the following form.

Lemma 3.8.10 *Arrangement and accumulation, and draw-delete are natural transformation in the situations:*

$$\begin{tikzcd} \mathbf{Chan} \arrow[r, bend left=60, "\mathcal{N}[K]"] \arrow[r, "(-)^K"'] \arrow[r, bend right=60, "\mathcal{N}[K]"'] \arrow[r, phantom, shift left=3, "\Downarrow arr"] \arrow[r, phantom, shift right=3, "\Downarrow acc"] & \mathbf{Chan} \end{tikzcd} \qquad \begin{tikzcd} \mathbf{Chan} \arrow[r, bend left=60, "\mathcal{N}[K+1]"] \arrow[r, bend right=60, "\mathcal{N}[K]"'] \arrow[r, phantom, "\Downarrow DD"] & \mathbf{Chan} \end{tikzcd}$$

Proposition 3.1.3 (4) and Lemma 3.2.8 say that these arr and DD are monoidal, as natural transformations.

Proof. Let $c\colon X \rightarrowtriangle Y$ be a channel. Then:

$$\begin{aligned} arr \circ \mathcal{N}[K](c) &= flat \circ \mathcal{D}(arr) \circ pml \circ \mathcal{N}(c) \\ &= flat \circ \mathcal{D}(\bigotimes) \circ arr \circ \mathcal{N}(c) && \text{by Proposition 3.7.2 (2)} \\ &= flat \circ \mathcal{D}(\bigotimes) \circ \mathcal{D}(c \times \cdots \times c) \circ arr && \text{by naturality of } arr \\ &= flat \circ \mathcal{D}(c \otimes \cdots \otimes c) \circ arr \\ &= c^K \circ arr. \end{aligned}$$

For accumulation the required equality $acc \circ c^K = \mathcal{N}[K](c) \circ acc$ can be obtained via a diagram chase:

$$\begin{tikzcd} X^K \arrow[r, "acc"] \arrow[d, "c\times\cdots\times c"'] & \mathcal{N}[K](X) \arrow[d, "\mathcal{N}(c)"] \\ \mathcal{D}(Y)^K \arrow[r, "acc"] \arrow[d, "\bigotimes"'] & \mathcal{N}[K]\mathcal{D}(Y) \arrow[d, "pml"] \\ \mathcal{D}(Y^K) \arrow[r, "\mathcal{D}(acc)"] & \mathcal{D}\mathcal{N}[K](Y) \end{tikzcd}$$

The upper part is ordinary naturality of *acc* and the lower is the formulation (3.35) of *pml*.

For naturality of draw-delete we use the equation $DD \circ pml = pml \circ DD$ from Proposition 3.8.4, together with naturality of *DD* in:

$$\begin{aligned} DD \circ \mathcal{N}[K+1](c) = flat \circ \mathcal{D}(DD) \circ pml \circ \mathcal{N}(c) &= flat \circ \mathcal{D}(pml) \circ DD \circ \mathcal{N}(c) \\ &= flat \circ \mathcal{D}(pml) \circ \mathcal{D}(\mathcal{N}(c)) \circ DD \\ &= flat \circ \mathcal{D}(\mathcal{N}[K](c)) \circ DD \\ &= \mathcal{N}[K](c) \circ DD. \end{aligned}$$

❑

Not only *acc*, *arr* and *DD* are natural with respect to channels, but also multinomial and hypergeometric maps. For the multinomial case we need to use the extension of the distribution functor $\mathcal{D}\colon \mathbf{Sets} \to \mathbf{Sets}$ to a functor $\overline{\mathcal{D}}\colon \mathbf{Chan} \to \mathbf{Chan}$. We have already seen this (general) construction in Exercise 1.12.8. We recall that $\overline{\mathcal{D}}(X) = \mathcal{D}(X)$ and $\overline{\mathcal{D}}(c) = c_*\colon \mathcal{D}(X) \to \mathcal{D}(Y)$ for a channel $c\colon X \rightarrowtriangle Y$. This extension $\overline{\mathcal{D}}$ is monoidal by a promotion of the usual tensor $\otimes\colon \mathcal{D}(X) \times \mathcal{D}(Y) \to \mathcal{D}(X \times Y)$ to a deterministic channel $\langle\otimes\rangle\colon \overline{\mathcal{D}}(X) \times \overline{\mathcal{D}}(Y) \rightarrowtriangle \overline{\mathcal{D}}(X \times Y)$.

Theorem 3.8.11 *The multinomial and hypergeometric channels are natural with respect to extended functors:*

$$\begin{tikzcd} \mathbf{Chan} \arrow[r, bend left, "\overline{\mathcal{D}}"] \arrow[r, bend right, "\mathcal{N}[K]"'] \arrow[r, phantom, "\Downarrow mn[K]"] & \mathbf{Chan} \end{tikzcd} \qquad \begin{tikzcd} \mathbf{Chan} \arrow[r, bend left, "\mathcal{N}[L]"] \arrow[r, bend right, "\mathcal{N}[K]"'] \arrow[r, phantom, "\Downarrow hg[K]"] & \mathbf{Chan} \end{tikzcd}$$

where $L \geq K$. These natural transformations are monoidal since they commute with mzip.

Concretely, this naturality of multinomial and hypergeometric channels means that for a channel $c \colon X \rightsquigarrow Y$ one has:

$$\begin{aligned} \mathcal{N}[K](c)_*(mn[K](\omega)) &= mn[K](c_*(\omega)) \\ \mathcal{N}[K](c)_*(hg[K](\upsilon)) &= hg[K]_*(\mathcal{N}[L](c)(\upsilon)), \end{aligned} \tag{3.44}$$

for urns $\omega \in \mathcal{D}(X)$ and $\upsilon \in \mathcal{N}[L](X)$.

Proof. We can write the multinomial channel as $mn[K] = acc \circ iid[K]$, see Theorem 2.6.7. Lemma 3.8.10 tells that accumulation is a natural transformation $acc \colon (-)^K \Rightarrow \mathcal{N}[K]$ between extended functors. Similarly, $iid[K]$ forms a natural transformation $\overline{\mathcal{D}} \Rightarrow (-)^K$ by Exercise 2.4.10 (4). The composition $mn[K]$ is monoidal by Corollary 3.3.6.

The hypergeometric $hg[K]$ channel is an iteration of draw-delete's, see Theorem 3.5.1. It is thus natural with respect to channels, since draw-delete is, see Lemma 3.8.10. By the same argument this hypergeometric natural transformation is monoidal, see Lemma 3.2.8. ❑

The Pólya channel from Section 3.6 does not fit in this picture since it is not natural w.r.t. channels. Intuitively, this can be explained from the fact that Pólya involves copying, and copying does not commute with channels, as we have seen early on in Exercises 2.5.1. Pólya channels are natural w.r.t. functions, see Proposition 3.6.2 (1), and indeed, functions do commute with copying, see Exercise 2.5.2.

3.8.1 Sampling of pushforward

We recall from Subsection 2.2.1 the programming language notation $x \leftarrow \omega$ for sampling an arbitrary element $x \in supp(\omega)$ from the (support of the) distribution $\omega \in \mathcal{D}(X)$, proportional to the probabilities in ω. The sampling mechanism works 'correctly', in the sense that if we accumulate a sufficiently long list of sampled elements to a multiset, then the resulting distribution obtained via frequentist learning approximates ω. As we have argued, the multinomial distribution is the mathematical counterpart of such sampling, and the correctness is expressed via the following diagram, repeated from Theorem 3.3.4:

$$\begin{tikzcd} \mathcal{D}(X) \arrow[r, "mn[K]"] \arrow[dr, bend right, "sam"'] & \mathcal{N}[K](X) \arrow[d, "Flrn"] \\ & X \end{tikzcd} \tag{3.45}$$

In this situation we sample K elements from the distribution ω, where the each multiset of K elements comes with its own multinomial probability. In the above composition $Flrn \circ mn[K]$, these probabilities are taken into account. Recall that the sample channel $sam \colon \mathcal{D}(X) \rightsquigarrow X$ is the identity function $id \colon \mathcal{D}(X) \to \mathcal{D}(X)$. This works for any number K.

In the code fragment (2.35) we have seen a two-step sampling $x \leftarrow \omega, y \leftarrow c(x)$, for a distribution $\omega \in \mathcal{D}(X)$ and a channel $c\colon X \rightsquigarrow Y$. We suggested that the resulting elements y from samples of the pushforward distribution $c_*(\omega)$ on Y. We now ask ourselves: is there a corresponding form of correctness?

We can use a multinomial distribution $mn[K](\omega)$ to sample K-sized multisets $\varphi \in \mathcal{N}[K](X)$ from ω. Then for each element x in φ we like to apply the channel c, giving a distribution $c(x) \in \mathcal{D}(Y)$. Then we can sample again from $c(x)$, say L-many elements, where both K and L are arbitrary numbers. In this set-up we use two constructions:

- the pointwise multinomial $mn[L](c) \coloneqq mn[L] \circ c\colon X \to \mathcal{D}(\mathcal{N}[L](Y))$, as already mentioned in Definition 2.6.6;
- the extension of the multiomial functor to channels, as in: $\mathcal{N}[K](c) \coloneqq pml \circ \mathcal{N}(c)\colon \mathcal{N}[K](X) \to \mathcal{D}(\mathcal{N}[K](Y))$, from Corollary 3.8.7.

We combine these points in the following composite:

$$\mathcal{D}(X) \xrightarrow{mn[K]} \mathcal{N}[K](X) \xrightarrow{\mathcal{N}[K](mn[L](c))} \mathcal{N}[L](\mathcal{N}[K](Y)) \xrightarrow{flat^{\mathcal{N}}} \mathcal{N}[L \cdot K](Y) \tag{3.46}$$

It first samples from a distribution $\omega \in \mathcal{D}(X)$, then applies $mn[L](c)$ inside the resulting samples, and then flattens the multisets of multisets to multisets. The correctness claim for sequential sampling is that frequentist learning of this map (3.46) is pushforward along the channel c. That is the context of the next result.

Proposition 3.8.12 *Sequential sampling is correct, in the sense that the following diagram commutes.*

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "mn[K]"] \arrow[d, "sam"'] & \mathcal{N}[K](X) \arrow[r, "\mathcal{N}[K](mn[L](c))"] & \mathcal{N}[K](\mathcal{N}[L](X)) \arrow[r, "flat^{\mathcal{N}}"] & \mathcal{N}[K \cdot L](Y) \arrow[d, "Flrn"] \\
X \arrow[rrr, "c"] & & & Y
\end{tikzcd}$$

Proof. We break the above rectangle up in three subdiagrams:

$$\begin{tikzcd}
& & \mathcal{N}[K](\mathcal{N}[L](X)) \arrow[rr, bend left, "flat^{\mathcal{N}}"] & & \\
\mathcal{D}(X) \arrow[r, "mn[K]"] \arrow[dd, "sam"'] & \mathcal{N}[K](X) \arrow[ur, bend left, "\mathcal{N}[K](mn[L](c))"] \arrow[dr, bend left, "mn[L](\mathcal{N}[K](c))"] \arrow[ddl, bend left, "Flrn"] & & & \mathcal{N}[K \cdot L](Y) \arrow[dd, "Flrn"] \\
& & \mathcal{N}[L](\mathcal{N}[K](X)) \arrow[urr, bend right, "flat^{\mathcal{N}}"'] & & \\
X \arrow[rrr, "c"] & & & & Y
\end{tikzcd}$$

The triangle on the left is (3.45). The diamond at the top commutes by Theorem 3.8.3 (2):

$$\begin{aligned}
flat^{\mathcal{N}} \circ \mathcal{N}[K](mn[L](c)) &= \mathcal{D}(flat^{\mathcal{N}}) \circ pml \circ \mathcal{N}(mn[L] \circ c) \\
&= \mathcal{D}(flat^{\mathcal{N}}) \circ mn[L] \circ pml \circ \mathcal{N}(c) \\
&= \mathcal{D}(flat^{\mathcal{N}}) \circ mn[L](\mathcal{N}[K](c)) \\
&= flat^{\mathcal{N}} \circ mn[L](\mathcal{N}[K](c)).
\end{aligned}$$

Next, the rectangle at the bottom commutes:

$$
\begin{aligned}
& \mathit{Flrn} \circledcirc \mathit{flat}^{\mathcal{N}} \circledcirc \mathit{mn}[L](\mathcal{N}[K](c)) \\
&= \mathit{flat} \circ \mathcal{D}(\mathit{Flrn}) \circ \mathcal{D}(\mathit{flat}^{\mathcal{N}}) \circ \mathit{mn}[L] \circ \mathit{pml} \circ \mathcal{N}(c) \\
&= \mathit{flat} \circ \mathcal{D}(\mathit{Flrn}) \circ \mathit{pml} \circ \mathcal{N}(c) && \text{by Theorem 3.3.8 (2)} \\
&= \mathit{flat} \circ \mathit{Flrn} \circ \mathcal{N}(c) && \text{by Theorem 3.8.1} \\
&= \mathit{flat} \circ \mathcal{D}(c) \circ \mathit{Flrn} && \text{by naturality} \\
&= c \circledcirc \mathit{Flrn}.
\end{aligned}
$$

❏

There is also 'anchestral' sampling of joint states $\tau \in \mathcal{D}(X \times Y)$ for which we have a graph representation $\tau = \langle \mathit{id}, c \rangle_*(\omega)$. In general, such a graph form can be obtained via 'disintegration', see Section 7.2, but here we assume it as given. This anchestral sampling works as follows.

```
x ← ω
y ← c(x)
return (x,y)
```
(3.47)

Proposition 3.8.13 *Ancestral sampling is correct, in the sense that the following diagram commutes.*

$$
\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "\mathit{mn}[K]"] \arrow[d, "\mathit{sam}"'] & \mathcal{N}[K](X) \arrow[r, "\mathcal{N}[K](\mathit{mzip} \circledcirc \langle \mathit{mn}[L](\mathit{id}), \mathit{mn}[L](c) \rangle)"] & \mathcal{N}[K](\mathcal{N}[L](X \times Y)) \arrow[r, "\mathit{flat}^{\mathcal{N}}"] & \mathcal{N}[K \cdot L](X \times Y) \arrow[d, "\mathit{Flrn}"] \\
X \arrow[rrr, "\langle \mathit{id}, c \rangle"] & & & X \times Y
\end{tikzcd}
$$

Proof. This follows from Proposition 3.8.12 once we show that the following diagram commutes:

$$
\begin{tikzcd}
X \arrow[r, "\langle \mathit{mn}[L](\mathit{id}), \mathit{mn}[L](c) \rangle"] \arrow[dr, bend right, "\mathit{mn}[L](\langle \mathit{id}, c \rangle)"'] & \mathcal{N}[L](X) \times \mathcal{N}[L](Y) \arrow[d, "\mathit{mzip}"] \\
& \mathcal{N}[L](X \times Y)
\end{tikzcd}
$$

This requires an application of Corollary 3.3.6 in:

$$
\begin{tikzcd}
X \arrow[r, "\langle \mathit{id}, c \rangle"] \arrow[dr, bend right, "\langle \mathit{id}, c \rangle"'] & \mathcal{D}(X) \times \mathcal{D}(Y) \arrow[r, "\mathit{mn}[L] \otimes \mathit{mn}[L]"] \arrow[d, "\otimes"'] & \mathcal{N}[L](X) \times \mathcal{N}[L](Y) \arrow[d, "\mathit{mzip}"] \\
& \mathcal{D}(X \times Y) \arrow[r, "\mathit{mn}[L]"] & \mathcal{N}[L](X \times Y)
\end{tikzcd}
$$

❏

Exercises

3.8.1 Consider the two distributions ω, ρ in (3.37) and check yourself the following equation, which is an instance of Theorem 3.8.1.

$$
\mathit{Flrn}_*\Big(\mathit{pml}(2|\omega\rangle + 1|\rho\rangle)\Big) = \tfrac{17}{36}|a\rangle + \tfrac{19}{36}|b\rangle. = (\mathit{flat} \circ \mathit{Flrn})(2|\omega\rangle + 1|\rho\rangle).
$$

3.8.2 Check that the following diagram does *not* commute.

$$\begin{tikzcd}
 & \mathcal{N}[K](\mathcal{D}(X)) \arrow[d, "pml"] \\
\mathcal{N}[K]^2(X) \arrow[ur, bend left, "\mathcal{N}(Flrn)"] \arrow[r, bend right, "Flrn"'] & \mathcal{D}(\mathcal{N}[K](X))
\end{tikzcd}$$

Consider for instance the multiset of multisets $2\big|1|a\rangle + 2|b\rangle\big\rangle + 1\big|3|a\rangle\big\rangle$.

3.8.3 Check that the construction of Corollary 3.8.7 indeed yields a functor $\mathcal{N}[K]\colon \mathbf{Chan} \to \mathbf{Chan}$, *i.e.* that identities and composition are preserved.

3.8.4 Consider the channel $c\colon \{1,2,3\} \multimap \{a,b\}$ given by:

$$c(1) \;=\; \tfrac{1}{6}|a\rangle + \tfrac{5}{6}|b\rangle \qquad c(2) \;=\; \tfrac{2}{3}|a\rangle + \tfrac{1}{3}|b\rangle \qquad c(3) \;=\; \tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle.$$

Show that the channel $\mathcal{N}[K](c)\colon \mathcal{N}[6](\{1,2,3\}) \multimap \mathcal{N}[6](\{a,b\})$ satisfies:

$$\begin{aligned}
&\mathcal{N}[K](c)\Big(1|1\rangle + 2|2\rangle + 3|3\rangle\Big) \\
&= \tfrac{1}{108}\big|6|a\rangle\big\rangle + \tfrac{1}{12}\big|5|a\rangle + 1|b\rangle\big\rangle + \tfrac{35}{144}\big|4|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{3}\big|3|a\rangle + 3|b\rangle\big\rangle \\
&\qquad + \tfrac{17}{72}\big|2|a\rangle + 4|b\rangle\big\rangle + \tfrac{1}{12}\big|1|a\rangle + 5|b\rangle\big\rangle + \tfrac{5}{432}\big|6|b\rangle\big\rangle.
\end{aligned}$$

3.8.5 Consider a distribution $\omega \in \mathcal{D}(X)$ as a channel $\mathbf{1} \multimap X$. Recall the extension of the functors $(-)^K$ and $\mathcal{N}[K]$ to the category of channels, see Corollary 3.8.7 and show that:

$$\Big(\mathbf{1} \overset{\omega}{\multimap} X\Big)^K \;=\; \mathbf{1} \overset{\omega^K}{\multimap} X^K \qquad\qquad \mathcal{N}[K]\Big(\mathbf{1} \overset{\omega}{\multimap} X\Big) \;=\; \mathbf{1} \xrightarrow{mn[K](\omega)} \mathcal{N}[K](X)$$

3.8.6 Show that frequentist learning is a natural transformation in:

$$\begin{tikzcd}
\mathbf{Chan} \arrow[r, bend left, "\mathcal{N}[K+1]"] \arrow[r, bend right, "id"'] \arrow[r, phantom, "\Downarrow Flrn"] & \mathbf{Chan}
\end{tikzcd}$$

3.8.7 Show that the extended functors $\mathcal{N}[K]\colon \mathbf{Chan} \to \mathbf{Chan}$ commute with sums of multisets: for a channel $f\colon X \multimap Y$,

$$\begin{tikzcd}
\mathcal{N}[K](X) \times \mathcal{N}[L](Y) \arrow[r, "+"] \arrow[d, "\mathcal{N}[K](f)\otimes\mathcal{N}[L](f)"'] & \mathcal{N}[K+L](X) \arrow[d, "\mathcal{N}[K+L](f)"] \\
\mathcal{N}[K](Y) \times \mathcal{N}[L](Y) \arrow[r, "+"] & \mathcal{N}[K+L](Y)
\end{tikzcd}$$

3.8.8 The parallel multinomial law *pml* does *not* commute with tensors (of multisets and distributions), as in the following diagram.

$$\begin{tikzcd}
\mathcal{N}[K](\mathcal{D}(X)) \times \mathcal{N}[L](\mathcal{D}(Y)) \arrow[r, "pml\otimes pml"] \arrow[d, "\otimes"] & \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \arrow[dd, "\otimes"] \\
\mathcal{N}[K\cdot L](\mathcal{D}(X) \times \mathcal{D}(Y)) \arrow[d, "\mathcal{N}[K\cdot L](\otimes)"'] \arrow[r, phantom, "\neq"] & \\
\mathcal{N}[K\cdot L](\mathcal{D}(X \times Y)) \arrow[r, "pml"] & \mathcal{N}[K\cdot L](X \times Y)
\end{tikzcd}$$

Take for instance $X = \{a, b\}$, $Y = \{0, 1\}$ with $K = 2$, $L = 1$ with

$$\text{distributions:} \quad \begin{cases} \omega = \frac{3}{4}|a\rangle + \frac{1}{4}|b\rangle \\ \rho = \frac{2}{3}|0\rangle + \frac{1}{3}|1\rangle \end{cases} \qquad \text{and multisets:} \quad \begin{cases} \varphi = 2|\omega\rangle \\ \psi = 1|\rho\rangle \end{cases}$$

1 Calculate:

$$\begin{aligned}(\otimes \circ (pml \otimes pml))(\varphi, \psi) &= \tfrac{3}{8}\big|2|a, 0\rangle\big\rangle + \tfrac{1}{4}\big|1|a, 0\rangle + 1|b, 0\rangle\big\rangle + \tfrac{1}{24}\big|2|b, 0\rangle\big\rangle \\ &\quad + \tfrac{3}{16}\big|2|a, 1\rangle\big\rangle + \tfrac{1}{8}\big|1|a, 1\rangle + 1|b, 1\rangle\big\rangle + \tfrac{1}{48}\big|2|b, 1\rangle\big\rangle.\end{aligned}$$

2 And also:

$$\begin{aligned}(pml \circ \mathcal{M}[K{\cdot}L](\otimes) \circ \otimes)(\varphi, \psi) &= \tfrac{1}{4}\big|2|a, 0\rangle\big\rangle + \tfrac{1}{16}\big|2|a, 1\rangle\big\rangle + \tfrac{1}{36}\big|2|b, 0\rangle\big\rangle + \tfrac{1}{144}\big|2|b, 1\rangle\big\rangle \\ &\quad + \tfrac{1}{4}\big|1|a, 0\rangle + 1|a, 1\rangle\big\rangle + \tfrac{1}{6}\big|1|a, 0\rangle + 1|b, 0\rangle\big\rangle \\ &\quad + \tfrac{1}{12}\big|1|a, 0\rangle + 1|b, 1\rangle\big\rangle + \tfrac{1}{12}\big|1|a, 1\rangle + 1|b, 0\rangle\big\rangle \\ &\quad + \tfrac{1}{24}\big|1|a, 1\rangle + 1|b, 1\rangle\big\rangle + \tfrac{1}{36}\big|1|b, 0\rangle + 1|b, 1\rangle\big\rangle.\end{aligned}$$

3.8.9 Let $\Omega \in \mathcal{D}^2(X)$ be a distribution of distributions. Argue that the following two program fragments yield the same outcome.

```
x ← flat(Ω)
return x
```

```
ω ← Ω
x ← ω
return x
```

Hint: Recall from Exercise 2.4.6 that $flat(\Omega) = sam_*(\Omega)$.

3.9 Parallel multinomials as law of monads

There is one thing we still wish to do in relation to the parallel multinomial law. We have described it as a map $pml[K]\colon \mathcal{N}[K](\mathcal{D}(X)) \to \mathcal{D}(\mathcal{N}[K](X))$ in a restricted manner, namely restricted to multisets of size K. The same holds for the tensor map $\bigotimes[K]\colon \mathcal{D}(X)^K \to \mathcal{D}(X^K)$. What if we drop this size restriction?

The formulation (3.39) already describes pml as a map $\mathcal{N}\mathcal{D}(X) \to \mathcal{D}\mathcal{N}(X)$, without size restrictions. It forms a natural transformation $\mathcal{N}\mathcal{D} \Rightarrow \mathcal{D}\mathcal{N}$ by Exercise 3.7.3. There is also such a formulation for the tensor $\bigotimes[K]$.

Definition 3.9.1 For a set X we define the *parallel states* map $pst\colon \mathcal{L}(\mathcal{D}(X)) \to \mathcal{D}(\mathcal{L}(X))$ as:

$$pst([\omega_1, \ldots, \omega_K]) := \bigotimes[K](\vec{\omega}). = \omega_1 \otimes \cdots \otimes \omega_K. \tag{3.48}$$

By Lemma 2.3.5 (1) this yields a natural transformation $pst\colon \mathcal{L}\mathcal{D} \Rightarrow \mathcal{D}\mathcal{L}$.

We collect some basic results.

Proposition 3.9.2 *Let X be an arbitrary set.*

1 Parallel multinomial and parallel states commute with accumulation in:

$$\begin{tikzcd} \mathcal{L}\mathcal{D}(X) \arrow[r, "pst"] \arrow[d, "acc"'] & \mathcal{D}\mathcal{L}(X) \arrow[d, "\mathcal{D}(acc)"] \\ \mathcal{N}\mathcal{D}(X) \arrow[r, "pml"] & \mathcal{D}\mathcal{N}(X) \end{tikzcd}$$

2 *The parallel states law* pst *commutes in the following way with the unit and flatten operations of the list monad* $\mathcal{L}$.

$$\begin{tikzcd} \mathcal{D}(X) \arrow[d, "unit"'] \arrow[dr, bend left, "\mathcal{D}(unit)"] & \\ \mathcal{L}\mathcal{D}(X) \arrow[r, "pml"] & \mathcal{D}\mathcal{L}(X) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{L}^2\mathcal{D}(X) \arrow[r, "\mathcal{L}(pml)"] \arrow[d, "flat"'] & \mathcal{L}\mathcal{D}\mathcal{L}(X) \arrow[r, "pml"] & \mathcal{D}\mathcal{L}^2(X) \arrow[d, "\mathcal{D}(flat)"] \\ \mathcal{L}\mathcal{D}(X) \arrow[rr, "pml"] & & \mathcal{D}\mathcal{L}(X) \end{tikzcd}$$

3 *The parallel multinomial law* $\mathrm{pml}\colon \mathcal{N}\mathcal{D}(X) \to \mathcal{D}\mathcal{N}(X)$ *commutes in the following way with the unit and flatten operations of the (natural) multiset monad* $\mathcal{N}$.

$$\begin{tikzcd} \mathcal{D}(X) \arrow[d, "unit"'] \arrow[dr, bend left, "\mathcal{D}(unit)"] & \\ \mathcal{N}\mathcal{D}(X) \arrow[r, "pml"] & \mathcal{D}\mathcal{N}(X) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{N}^2\mathcal{D}(X) \arrow[r, "\mathcal{N}(pml)"] \arrow[d, "flat"'] & \mathcal{N}\mathcal{D}\mathcal{N}(X) \arrow[r, "pml"] & \mathcal{D}\mathcal{N}^2(X) \arrow[d, "\mathcal{D}(flat)"] \\ \mathcal{N}\mathcal{D}(X) \arrow[rr, "pml"] & & \mathcal{D}\mathcal{N}(X) \end{tikzcd}$$

4 *The parallel multinomial law* $\mathrm{pml}\colon \mathcal{N}\mathcal{D}(X) \to \mathcal{D}\mathcal{N}(X)$ *is a map of monoids.*

The analogous property for $\mathrm{pst}\colon \mathcal{L}\mathcal{D}(X) \to \mathcal{D}\mathcal{L}(X)$ *also holds, but should be formulated slightly differently, since* $\mathcal{L}(X)$ *is not a* commutative *monoid, so Proposition 2.7.2 does not apply. For lists* $\ell_1, \ell_2 \in \mathcal{L}\mathcal{D}(X)$ *one does have:*

$$pst(\ell_1 +\!\!+ \ell_2) = \mathcal{D}(+\!\!+)\big(pst(\ell_1) \otimes pst(\ell_2)\big).$$

Proof. 1 For a sequence $\vec{\omega} \in \mathcal{L}\mathcal{D}(X)$, say of length $K \in \mathbb{N}$, with $acc(\vec{\omega}) = \sum_{1 \leq i \leq m} n_i|\omega_i\rangle$, we have, by Exercise 3.3.10 (3),

$$\begin{aligned} (\mathcal{D}(acc) \circ pst)(\vec{\omega}) = \mathcal{D}(acc)\big(\omega_1^{n_1} \otimes \cdots \omega_m^{n_m}\big) &= mn[n_1](\omega_1) + \cdots + mn[n_m](\omega_m) \\ &\overset{(3.36)}{=} pml(\textstyle\sum_i n_i|\omega_i\rangle) = (pml \circ acc)(\vec{\omega}). \end{aligned}$$

2 The equation for the units is easy. For $\omega \in \mathcal{D}(X)$ we have:

$$(pst \circ unit)(\omega) = pst([\omega]) = \sum_{x \in X} \omega(x)\big|[x]\rangle = \mathcal{D}(unit)(\omega).$$

Let lists of distributions $L_1, \ldots, L_K \in \mathcal{L}\mathcal{D}(X)$ be given with $\|L_i\| = N_i$. We write $L := flat([L_1, \ldots, L_K]) = L_1 +\!\!+ \cdots +\!\!+ L_K$, see Exercise 1.4.5, so that $\|L\| = \sum_i N_i$. Then:

$$\begin{aligned} &(\mathcal{D}(flat) \circ pst \circ \mathcal{L}(pst))(\vec{L}) \\ &= (\mathcal{D}(flat) \circ pst)([pst(L_1), \ldots, pst(L_K)]) \\ &= \sum_{\ell_1 \in X^{N_1}, \ldots, \ell_K \in X^{N_K}} (pst(L_1) \otimes \cdots \otimes pst(L_K))(\ell_1, \ldots, \ell_K)\,\big|\,flat([\ell_1, \ldots, \ell_K])\rangle \\ &= \sum_{\ell_1 \in X^{N_1}, \ldots, \ell_K \in X^{N_K}} \prod_{1 \leq i \leq K} pst(L_i)(\ell_i)\,\big|\,\ell_1 +\!\!+ \cdots +\!\!+ \ell_K\rangle \\ &= \sum_{\ell_1 \in X^{N_1}, \ldots, \ell_K \in X^{N_K}} \prod_{1 \leq i \leq K} \prod_{1 \leq j \leq N_i} L_{ij}(\ell_{ij})\,\big|\,\ell_1 +\!\!+ \cdots +\!\!+ \ell_K\rangle \\ &= \sum_{\ell \in X^{N_1 + \cdots + N_K}} \prod_{1 \leq k \leq \sum_i N_i} L_k(\ell_k)\,\big|\,\ell\rangle = pst(L) = pst(flat([L_1, \ldots, L_K])) = (pst \circ flat)(\vec{L}). \end{aligned}$$

3 First, for $\omega \in \mathcal{D}(X)$ we have, by Exercise 3.7.1,

$$(pml \circ unit)(\omega) = pml(1|\omega\rangle) = mn[1](\omega) = \sum_{x \in X} \omega(x)\big|1|x\rangle\rangle = \mathcal{D}(unit)(\omega).$$

For flatten we have to do a bit more work. We recall, for a commutative monoid M, the $\mathcal{N}$-algebra $sum\colon \mathcal{N}(\mathcal{D}(M)) \to \mathcal{D}(M)$ given by convolution. We also recall that by Proposition 1.6.6 the following two diagrams commute, where $f\colon M_1 \to M_2$ is a map of (commutative) monoids.

$$\begin{tikzcd} \mathcal{N}^2\mathcal{D}(M) \arrow[r, "\mathcal{N}(sum)"] \arrow[d, "flat"'] & \mathcal{N}\mathcal{D}(M) \arrow[d, "sum"] \\ \mathcal{N}\mathcal{D}(M) \arrow[r, "sum"] & \mathcal{D}(M) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{N}\mathcal{D}(M_1) \arrow[r, "\mathcal{N}\mathcal{D}(f)"] \arrow[d, "sum"'] & \mathcal{N}\mathcal{D}(M_2) \arrow[d, "sum"] \\ \mathcal{D}(M_1) \arrow[r, "\mathcal{D}(f)"] & \mathcal{D}(M_2) \end{tikzcd} \tag{3.49}$$

We use the formulation (3.39) namely $pml = sum \circ \mathcal{N}\mathcal{D}(unit)$. Then:

$$\begin{aligned} & \mathcal{D}(flat) \circ pml \circ \mathcal{N}(pml) \\ &= \mathcal{D}(flat) \circ sum \circ \mathcal{N}\mathcal{D}(unit) \circ \mathcal{N}(pml) && \text{by (3.39)} \\ &= sum \circ \mathcal{N}\mathcal{D}(flat) \circ \mathcal{N}\mathcal{D}(unit) \circ \mathcal{N}(pml) && \text{by (3.49), on the right} \\ &= sum \circ \mathcal{N}(pml) && \text{by a flatten-unit law} \\ &= sum \circ \mathcal{N}(sum \circ \mathcal{N}\mathcal{D}(unit)) && \text{by (3.39) again} \\ &= sum \circ flat \circ \mathcal{N}\mathcal{N}\mathcal{D}(unit) && \text{by (3.49), on the left} \\ &= sum \circ \mathcal{N}\mathcal{D}(unit) \circ flat && \text{by naturality} \\ &= pml \circ flat && \text{once again by (3.39).} \end{aligned}$$

4 In Proposition 3.7.5 (2) we have already seen that the parallel multinomial law $pml\colon \mathcal{N}\mathcal{D}(X) \to \mathcal{D}\mathcal{N}(X)$ is a map of monoids. Here we give an alternative, more abstract proof, based on Proposition 1.6.6. This amounts to showing that the outer rectangle commutes in:

$$\begin{tikzcd} \mathcal{N}^2\mathcal{D}(X) \arrow[r, "\mathcal{N}^2\mathcal{D}(unit)"] \arrow[d, "flat"'] \arrow[rr, bend left, "\mathcal{N}(pml)"] & \mathcal{N}^2\mathcal{D}\mathcal{N}(X) \arrow[r, "\mathcal{N}(sum)"] \arrow[d, "flat"'] & \mathcal{N}\mathcal{D}\mathcal{N}(X) \arrow[d, "sum"] \\ \mathcal{N}\mathcal{D}(X) \arrow[r, "\mathcal{N}\mathcal{D}(unit)"] \arrow[rr, bend right, "pml"'] & \mathcal{N}\mathcal{D}\mathcal{N}(X) \arrow[r, "sum"] & \mathcal{D}\mathcal{N}(X) \end{tikzcd}$$

The rectangle on the left commutes by naturality of *flat*; the one on the right is an instance of the rectangle on the left in (3.49).

For the parallel state map *pst*, let $\vec{\omega}, \vec{\rho} \in \mathcal{L}\mathcal{D}(X)$ be lists of state, say with $\|\vec{\omega}\| = K$ and $\|\vec{\rho}\| = L$. Then:

$$\begin{aligned} \mathcal{D}(+\!\!+)\Big(pst(\vec{\omega}) \otimes pst(\vec{\rho})\Big) &= \mathcal{D}(+\!\!+)\Big(\bigotimes[K](\vec{\omega}) \otimes \bigotimes[L](\vec{\rho})\Big) \\ &= \sum_{\vec{x}\in X^K} \sum_{\vec{y}\in X^L} \bigotimes[K](\vec{\omega})(\vec{x}) \cdot \bigotimes[L](\vec{\rho})(\vec{y}) \,\big|\, \vec{x} +\!\!+ \vec{y}\rangle \\ &= \sum_{\vec{z}\in X^{K+L}} \bigotimes[K+L](\vec{\omega} +\!\!+ \vec{\rho})(\vec{z}) \,\big|\, \vec{z}\rangle \\ &= \bigotimes[K+L](\vec{\omega} +\!\!+ \vec{\rho}) \;=\; pst(\vec{\omega} +\!\!+ \vec{\rho}). \end{aligned}$$

❑

Theorem 3.9.3 *The two composite functors*

$$\mathbf{Sets} \xrightarrow{\mathcal{D}\mathcal{L}} \mathbf{Sets} \qquad\qquad \mathbf{Sets} \xrightarrow{\mathcal{D}\mathcal{N}} \mathbf{Sets}$$

are both monads. Their unit and flatten maps are given as follows, via pst *and* pml*.*

- *For $\mathcal{DL}$,*

$$\mathit{unit}^{\mathcal{DL}} := \left(\begin{tikzcd} & \mathcal{D}(X) \arrow[dr, "\mathcal{D}(\mathit{unit})"] & \\ X \arrow[ur, "\mathit{unit}"] \arrow[dr, "\mathit{unit}"'] & & \mathcal{DL}(X) \\ & \mathcal{L}(X) \arrow[ur, "\mathcal{L}(\mathit{unit})"'] & \end{tikzcd} \right)$$

$$\mathit{flat}^{\mathcal{DL}} := \left(\begin{tikzcd} & & & \mathcal{DL}^2(X) \arrow[dr, "\mathcal{D}(\mathit{flat})"] & \\ \mathcal{DLDL}(X) \arrow[r, "\mathcal{D}(\mathit{pst})"] & \mathcal{D}^2\mathcal{L}^2(X) \arrow[ur, "\mathit{flat}"] \arrow[dr, "\mathcal{D}^2(\mathit{flat})"'] & & & \mathcal{DL}(X) \\ & & \mathcal{D}^2\mathcal{L}(X) \arrow[urr, "\mathit{flat}"'] & & \end{tikzcd} \right)$$

- *For $\mathcal{DN}$,*

$$\mathit{unit}^{\mathcal{DN}} := \left(\begin{tikzcd} & \mathcal{D}(X) \arrow[dr, "\mathcal{D}(\mathit{unit})"] & \\ X \arrow[ur, "\mathit{unit}"] \arrow[dr, "\mathit{unit}"'] & & \mathcal{DN}(X) \\ & \mathcal{N}(X) \arrow[ur, "\mathcal{N}(\mathit{unit})"'] & \end{tikzcd} \right)$$

$$\mathit{flat}^{\mathcal{DN}} := \left(\begin{tikzcd} & & & \mathcal{DN}^2(X) \arrow[dr, "\mathcal{D}(\mathit{flat})"] & \\ \mathcal{DNDN}(X) \arrow[r, "\mathcal{D}(\mathit{pml})"] & \mathcal{D}^2\mathcal{N}^2(X) \arrow[ur, "\mathit{flat}"] \arrow[dr, "\mathcal{D}^2(\mathit{flat})"'] & & & \mathcal{DN}(X) \\ & & \mathcal{D}^2\mathcal{N}(X) \arrow[urr, "\mathit{flat}"'] & & \end{tikzcd} \right)$$

Proof. This is a standard result in category theory, originally by Beck, see *e.g.* Beck (1969); Barr and Wells (1985); Jacobs (2016b), and see also Exercise 3.9.4 below. The two diamonds in the above descriptions of unit and flatten commute by naturality. The monad laws for $\mathcal{DL}$ and $\mathcal{DN}$ hold by items (2) and (3) in Proposition 3.9.2. ❑

In the remainder of this section we concentrate on the monad $\mathcal{DN}\colon \mathbf{Sets} \to \mathbf{Sets}$. The aim is to show that there is a map of monads $\mathcal{DN} \Rightarrow \mathcal{M}$, where $\mathcal{M}\colon \mathbf{Sets} \to \mathbf{Sets}$ is the multiset monad, with non-negative numbers as multiplicities — and not just natural numbers, as for natural multisets, in $\mathcal{N}$. This map of monads is introduced in Dash and Staton (2020), under the name 'intensity'. Therefor we shall write it as $\mathit{ity}\colon \mathcal{DN} \Rightarrow \mathcal{M}$.

There are obvious inclusions $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ and $\mathcal{N}(X) \hookrightarrow \mathcal{M}(X)$ that we used before. We now need to formalise the situation and so we shall use explicit names for these inclusions, as natural transformations, namely:

$$\mathcal{D} \overset{\sigma}{\Longrightarrow} \mathcal{M} \qquad\qquad \mathcal{N} \overset{\tau}{\Longrightarrow} \mathcal{M}.$$

These σ and τ are maps of monads. They are used implicitly for instance in the equation $\mathit{flat}(\mathit{pml}(\Psi)) = \mathit{flat}(\Psi)$ in Proposition 3.8.2. As a commuting diagram (of natural transformations), it looks as follows.

$$\begin{tikzcd} \mathcal{ND} \arrow[r, "\mathcal{N}(\sigma)"] \arrow[d, "\mathit{pml}"'] & \mathcal{NM} \arrow[r, "\tau"] & \mathcal{MM} \arrow[dr, "\mathit{flat}"] & \\ \mathcal{DN} \arrow[r, "\sigma"] & \mathcal{MN} \arrow[r, "\mathcal{M}(\tau)"] & \mathcal{MM} \arrow[r, "\mathit{flat}"'] & \mathcal{M} \end{tikzcd} \tag{3.50}$$

$$
\begin{aligned}
ity \circ flat^{\mathcal{DN}} &\overset{(3.51)}{=} flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(\tau) \circ \mathcal{D}(flat^{\mathcal{N}}) \circ flat^{\mathcal{D}} \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{DM}(\tau) \circ \mathcal{D}(\tau) \circ flat^{\mathcal{D}} \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \mathcal{M}^2(\tau) \circ \mathcal{M}(\tau) \circ \sigma \circ flat^{\mathcal{D}} \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \mathcal{M}^2(\tau) \circ \mathcal{M}(\tau) \circ flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \mathcal{M}^3(\tau) \circ \mathcal{M}^2(\tau) \circ \sigma \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \mathcal{M}^3(\tau) \circ \sigma \circ \mathcal{DM}(\tau) \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \mathcal{M}^2(\tau) \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \sigma \circ \mathcal{DM}(\tau) \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \sigma \circ \mathcal{DM}(\tau) \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{DM}(\tau) \circ \mathcal{D}(\sigma) \circ \mathcal{D}(pml) \\
&\overset{(3.50)}{=} flat^{\mathcal{M}} \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \sigma \circ \mathcal{DM}(\tau) \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{D}(\tau) \circ \mathcal{DN}(\sigma) \\
&= flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{DM}^2(\tau) \circ \mathcal{D}(\tau) \circ \mathcal{DN}(\sigma) \\
&= flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(flat^{\mathcal{M}}) \circ \mathcal{DM}(flat^{\mathcal{M}}) \circ \mathcal{DM}^2(\tau) \circ \mathcal{D}(\tau) \circ \mathcal{DN}(\sigma) \\
&= flat^{\mathcal{M}} \circ \mathcal{M}(flat^{\mathcal{M}}) \circ \sigma \circ \mathcal{DM}(flat^{\mathcal{M}}) \circ \mathcal{DM}^2(\tau) \circ \mathcal{D}(\tau) \circ \mathcal{DN}(\sigma) \\
&= flat^{\mathcal{M}} \circ flat^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(\tau) \circ \mathcal{DN}(flat^{\mathcal{M}}) \circ \mathcal{DNM}(\tau) \circ \mathcal{DN}(\sigma) \\
&\overset{(3.51)}{=} flat^{\mathcal{M}} \circ ity \circ \mathcal{DN}(ity).
\end{aligned}
$$

Figure 3.1 Equational proof that the intensity natural transformation *ity* from (3.51) commutes with flattens, as part of the proof of Theorem 3.9.4.

Theorem 3.9.4 (From Dash and Staton (2020)) *The* intensity *natural transformation* $ity \colon \mathcal{DN} \Rightarrow \mathcal{M}$ *has components:*

$$
ity \;:=\; \left(\begin{tikzcd}
 & \mathcal{MN}(X) \arrow[dr, "\mathcal{M}(\tau)"] & & \\
\mathcal{DN}(X) \arrow[ur, "\sigma"] \arrow[dr, "\mathcal{D}(\tau)"'] & & \mathcal{MM}(X) \arrow[r, "flat"] & \mathcal{M}(X) \\
 & \mathcal{DM}(X) \arrow[ur, "\sigma"'] & &
\end{tikzcd} \right) \tag{3.51}
$$

This intensity ity *is a map of monads. It thus induces a functor* $\mathbf{Chan}(\mathcal{DN}) \to \mathbf{Chan}(\mathcal{M})$ *between the associated categories of channels.*

This definition 3.51 looks impressive, but for practical purposes we can just write $ity(\Psi) = flat(\Psi)$, leaving inclusions σ, τ implicit. However, in the proof below we will be very precise and make these inclusions explicit.

Proof. Commutation of intensity with units is easy. For clarity we write $unit^{\mathcal{M}}$ for the unit of the monad $\mathcal{M}$ and $unit^{\mathcal{DN}}$ for the unit of $\mathcal{DN}$, see Theorem 3.9.3. Then:

$$
\begin{aligned}
ity \circ unit^{\mathcal{DN}} &\overset{(3.51)}{=} flat^{\mathcal{M}} \circ \mathcal{M}(\tau) \circ \sigma \circ \mathcal{D}(unit^{\mathcal{N}}) \circ unit^{\mathcal{D}} \\
&= flat^{\mathcal{M}} \circ \mathcal{M}(\tau) \circ \mathcal{M}(unit^{\mathcal{N}}) \circ \sigma \circ unit^{\mathcal{D}} \\
&= flat^{\mathcal{M}} \circ \mathcal{M}(unit^{\mathcal{M}}) \circ unit^{\mathcal{M}} && \text{since } \tau, \sigma \text{ are maps of monads} \\
&= unit^{\mathcal{M}}.
\end{aligned}
$$

Commutation with flatten maps is more laborious and involves a long calculation. Figure 3.1 provides all required equational steps, using the standard equations for (maps of) monads. ❑

There is more to say about this situation.

Proposition 3.9.5 *For each set X, the intensity map is a homomorphism of monoids, so that we*

can rephrase Diagram (3.50) *as a triangle of monoid homomorphisms:*

$$\begin{tikzcd} (\mathcal{ND}(X), +, 0) \arrow[r, "pml"] \arrow[dr, bend right] & (\mathcal{DN}(X), +, 0) \arrow[d, "ity"] \\ & (\mathcal{M}(X), +, 0). \end{tikzcd}$$

Proof. We use the convolution monoid structure on $\mathcal{DN}(X)$, defined in Proposition 2.7.2, in the form of a map $sum\colon \mathcal{NDN}(X) \to \mathcal{DN}(X)$, as above. The zero element in $\mathcal{DN}(X)$ is $1|\mathbf{0}\rangle$, where $\mathbf{0} \in \mathcal{N}(X)$ is the empty multiset. It satisfies, for $x \in X$,

$$ity(1|\mathbf{0}\rangle)(x) \;=\; flat(1|\mathbf{0}\rangle)(x) \;=\; 1 \cdot \mathbf{0}(x) \;=\; 0.$$

Hence $ity(1|\mathbf{0}\rangle) = \mathbf{0}$.

For distributions $\omega, \rho \in \mathcal{DN}(X)$ we have $ity(\omega + \rho) = ity(\omega) + ity(\rho)$ since for each $x \in X$,

$$\begin{aligned} ity(\omega+\rho)(x) \;&=\; flat(\omega+\rho)(x) \;=\; \sum_{\varphi \in \mathcal{N}(X)} (\omega+\rho)(\varphi) \cdot \varphi(x) \\ &\overset{(2.60)}{=} \sum_{\varphi \in \mathcal{N}(X)} \mathcal{D}(+)(\omega \otimes \rho)(\varphi) \cdot \varphi(x) \\ &= \sum_{\psi, \chi \in \mathcal{N}(X)} \omega(\psi) \cdot \rho(\chi) \cdot (\psi + \chi)(x) \\ &= \sum_{\psi, \chi \in \mathcal{N}(X)} \omega(\psi) \cdot \rho(\chi) \cdot (\psi(x) + \chi(x)) \\ &= \sum_{\psi \in \mathcal{N}(X)} \omega(\psi) \cdot \Big(\sum_{\chi \in \mathcal{N}(X)} \rho(\chi) \Big) \cdot \psi(x) \;+\; \sum_{\chi \in \mathcal{N}(X)} \Big(\sum_{\psi \in \mathcal{N}(X)} \omega(\psi) \Big) \cdot \rho(\chi) \cdot \chi(x) \\ &= flat(\omega)(x) + flat(\rho)(x) \;=\; \Big(ity(\omega) + ity(\rho) \Big)(x). \end{aligned}$$

❑

Exercises

3.9.1 Check that the intensity natural transformation $ity\colon \mathcal{DN} \Rightarrow \mathcal{M}$ defined in (3.51) restricts to $\mathcal{DN}[K] \Rightarrow \mathcal{M}[K]$, for each $K \in \mathbb{N}$.

3.9.2 Check that the 'mean' results for multinomial, hypergeometric and Pólya distributions from Proposition 3.3.7 and from Lemmas 3.5.5 (2) and 3.6.1 (2) can be reformulated in terms of intensity as:

$$\begin{aligned} ity\Big(mn[K](\omega)\Big) \;&=\; K \cdot \omega \\ ity\Big(hg[K](\psi)\Big) \;&=\; K \cdot Flrn(\psi) \\ ity\Big(pl[K](\psi)\Big) \;&=\; K \cdot Flrn(\psi). \end{aligned}$$

3.9.3 In Corollary 3.8.7 we have seen the extended functor $\mathcal{N}[K]\colon \mathbf{Chan} \to \mathbf{Chan}$. The flatten operation $flat\colon \mathcal{NN} \Rightarrow \mathcal{N}$ for (natural) multisets, from Subsection 1.6.2, restricts to $\mathcal{N}[K]\mathcal{N}[L] \Rightarrow \mathcal{N}[K \cdot L]$, making $\mathcal{N}[K]\colon \mathbf{Sets} \to \mathbf{Sets}$ into what is called a *graded* monad, see *e.g.* Milius et al. (2015); Fujii et al. (2016). Also the extended functor $\mathcal{N}[K]\colon \mathbf{Chan} \to \mathbf{Chan}$ is such a graded monad, essentially by Proposition 3.9.2.

The aim of this exercise is to show that these (extended) flattens do not commute with

multizip. This means that the following diagram of channels does *not* commute.

$$\begin{tikzcd}
\mathcal{N}[K]\mathcal{N}[L](X) \times \mathcal{N}[K]\mathcal{N}[L](Y) \arrow[r, "flat \otimes flat"] \arrow[d, "mzip"'] & \mathcal{N}[K \cdot L](X) \times \mathcal{N}[K \cdot L](Y) \arrow[dd, "mzip"] \\
\mathcal{N}[K](\mathcal{N}[L](X) \times \mathcal{N}[L](Y)) \arrow[d, "\mathcal{N}[K](mzip)"'] \arrow[r, phantom, "\neq"] & \\
\mathcal{N}[K]\mathcal{N}[L](X \times Y) \arrow[r, "flat"] & \mathcal{N}[K \cdot L](X \times Y)
\end{tikzcd}$$

Elaborating a counterexample is quite initimidating, so we proceed step by step. We take as spaces:

$$X = \{a, b\} \qquad \text{and} \qquad Y = \{0, 1\}.$$

We use $K = 2$ and $L = 3$ for the multisets of multisets:

$$\begin{aligned}
\Phi &= 1\big|2|a\rangle + 1|b\rangle\big\rangle + 1\big|1|a\rangle + 2|b\rangle\big\rangle \in \mathcal{N}[2]\mathcal{N}[3](X) \\
\Psi &= 1\big|2|0\rangle + 1|1\rangle\big\rangle + 1\big|3|1\rangle\big\rangle \in \mathcal{N}[2]\mathcal{N}[3](Y).
\end{aligned}$$

1 Check that going east-south in the above diagram yields:

$$\begin{aligned}
&mzip(flat(\Phi), flat(\Psi)) \\
&= mzip\Big(3|a\rangle + 3|b\rangle,\ 2|0\rangle + 4|1\rangle\Big) \\
&= \tfrac{1}{5}\big|3|a,1\rangle + 2|b,0\rangle + 1|b,1\rangle\big\rangle + \tfrac{3}{5}\big|1|a,0\rangle + 2|a,1\rangle + 1|b,0\rangle + 2|b,1\rangle\big\rangle \\
&\qquad + \tfrac{1}{5}\big|2|a,0\rangle + 1|a,1\rangle + 3|b,1\rangle\big\rangle.
\end{aligned}$$

2 The other path, south-east, will be done in several steps. Write $\Phi = 1|\varphi_1\rangle + 1|\varphi_2\rangle$ and $\Psi = 1|\psi_1\rangle + 1|\psi_2\rangle$ where:

$$\begin{cases} \varphi_1 = 2|a\rangle + 1|b\rangle \\ \varphi_2 = 1|a\rangle + 2|b\rangle \end{cases} \qquad \begin{cases} \psi_1 = 2|0\rangle + 1|1\rangle \\ \psi_2 = 3|1\rangle. \end{cases}$$

Show then that:

$$mzip(\Phi, \Psi) = \tfrac{1}{2}\big|1|\varphi_1, \psi_1\rangle + 1|\varphi_2, \psi_2\rangle\big\rangle + \tfrac{1}{2}\big|1|\varphi_1, \psi_2\rangle + 1|\varphi_2, \psi_1\rangle\big\rangle.$$

3 Show next:

$$\begin{aligned}
mzip(\varphi_1, \psi_1) &= \tfrac{2}{3}\big|1|a,0\rangle + 1|a,1\rangle + 1|b,0\rangle\big\rangle + \tfrac{1}{3}\big|2|a,0\rangle + 1|b,1\rangle\big\rangle \\
mzip(\varphi_1, \psi_2) &= 1\big|2|a,1\rangle + 1|b,1\rangle\big\rangle \\
mzip(\varphi_2, \psi_1) &= \tfrac{1}{3}\big|1|a,1\rangle + 2|b,0\rangle\big\rangle + \tfrac{2}{3}\big|1|a,0\rangle + 1|b,0\rangle + 1|b,1\rangle\big\rangle \\
mzip(\varphi_2, \psi_2) &= 1\big|1|a,1\rangle + 2|b,1\rangle\big\rangle.
\end{aligned}$$

4 Show now that:

$$\begin{aligned}
&pml\Big(1|mzip(\varphi_1, \psi_1)\rangle + 1|mzip(\varphi_2, \psi_2)\rangle\Big) \\
&= \tfrac{2}{3}\big|1|1|a,0\rangle + 1|a,1\rangle + 1|b,0\rangle\big\rangle + 1|1|a,1\rangle + 2|b,1\rangle\big\rangle\big\rangle \\
&\qquad + \tfrac{1}{3}\big|1|2|a,0\rangle + 1|b,1\rangle\big\rangle + 1|1|a,1\rangle + 2|b,1\rangle\big\rangle\big\rangle \\
&pml\Big(1|mzip(\varphi_1, \psi_2)\rangle + 1|mzip(\varphi_2, \psi_1)\rangle\Big) \\
&= \tfrac{1}{3}|1|1|a,1\rangle + 2|b,0\rangle\big\rangle + 1|2|a,1\rangle + 1|b,1\rangle\big\rangle\big\rangle + \\
&\qquad + \tfrac{2}{3}|1|1|a,0\rangle + 1|b,0\rangle + 1|b,1\rangle\big\rangle + 1|2|a,1\rangle + 1|b,1\rangle\big\rangle\big\rangle.
\end{aligned}$$

5 Finally, check that the south-east past yields:

$$\begin{aligned}\big(\mathit{flat} \circ \mathcal{N}[2](\mathit{mzip}) \circ \mathit{mzip}\big)(\Phi, \Psi) \;=\; & \tfrac{2}{3}\big|1|a,0\rangle + 2|a,1\rangle + 1|b,0\rangle + 2|b,1\rangle\big\rangle \\ & + \tfrac{1}{6}\big|2|a,0\rangle + 3|b,1\rangle + 1|a,1\rangle\big\rangle \\ & + \tfrac{1}{6}\big|3|a,1\rangle + 2|b,0\rangle + 1|b,1\rangle\big\rangle.\end{aligned}$$

This differs from what we get in the first item, via the east-south route.

3.9.4 Let $S = (S, \mathit{unit}^S, \mathit{flat}^S)$ and $T = (T, \mathit{unit}^T, \mathit{flat}^T)$ be two monads on the same category $\mathbb{C}$, with a distributive law of monads $\lambda\colon ST \Rightarrow TS$ between them. This means that λ commutes with the unit and flatten operations of both S and T.

1 Write down in diagrams what these commutation requirements mean for λ; for the original source, see Beck (1969); alternatively, see *e.g.* Barr and Wells (1985); Jacobs (2016b).

2 Use λ to define new unit and flatten operations $\mathit{unit}^{TS} : \mathit{id} \Rightarrow TS$ and $\mathit{flat}^{TS} : TSTS \Rightarrow TS$ as in Theorem 3.9.3 and show that they make the composite TS into a monad.

3.9.5 We have seen that parallel multinomial *pml* forms a distributive law $\mathcal{N}\mathcal{D} \Rightarrow \mathcal{D}\mathcal{N}$. There is also a distributive law $\mathcal{N}\mathcal{P} \Rightarrow \mathcal{P}\mathcal{N}$, with the powerset instead of the distribution monad, see *e.g.* (Zwart, 2020, Ex. 5.29) (referring to Manes and Mulry (2007)). Its components $\lambda_X\colon \mathcal{N}(\mathcal{P}(X)) \to \mathcal{P}(\mathcal{N}(X))$ given by:

$$\lambda\Big(\textstyle\sum_i n_i|U_i\rangle\Big) := \Big\{\textstyle\sum_i n_i|x_i\rangle \;\Big|\; \forall i.\, x_i \in U_i\Big\}.$$

1 Check that λ is *not* the free extension, as in Proposition 1.6.5, of the map $\mathcal{P}(\mathit{unit}^{\mathcal{N}})\colon \mathcal{P}(X) \to \mathcal{P}(\mathcal{N}(X))$.

2 Describe the monad $\mathcal{P}\mathcal{N}$ on **Sets** resulting from the distributive law λ.

3 Describe also the resulting 'lifting' of $\mathcal{N}$ to the Kleisli category $\mathcal{K}\ell(\mathcal{P})$.

3.10 Discrete Poisson point processes

In the previous section we have seen the composite monads $\mathcal{D}\mathcal{L}$ and $\mathcal{D}\mathcal{N}$. As illustration of how these monads can be used, we introduce (discrete) Poisson process as infinite mixtures of K-sized iid and K-sized multinomial distributions, where the numbers K have a Poisson distribution. This gives composite distributions on lists and on (natural) multisets of arbitrary size. We actually stretch things a bit, because we are not using $\mathcal{D}\mathcal{L}$ and $\mathcal{D}\mathcal{N}$ but the infinite version $\mathcal{D}_\infty\mathcal{L}$ and $\mathcal{D}_\infty\mathcal{N}$. This does not change things fundamentally.

Definition 3.10.1 For an arbitrary set X and a rate $\lambda \in \mathbb{R}_{\geq 0}$ we define the *Poisson-iid* and *Poisson-multinomial* maps:

$$\mathcal{D}(X) \xrightarrow{\mathit{Piid}[\lambda]} \mathcal{D}_\infty(\mathcal{L}(X)) \qquad\qquad \mathcal{D}(X) \xrightarrow{\mathit{Pmn}[\lambda]} \mathcal{D}_\infty(\mathcal{N}(X)).$$

On a distribution $\omega \in \mathcal{D}(X)$ they are given as follows.

1 The Poisson-idd is:

$$\begin{aligned} Piid[\lambda](\omega) &:= \sum_{\ell \in \mathcal{L}(supp(\omega))} e^{-\lambda} \cdot \frac{\lambda^{\|\ell\|}}{\|\ell\|!} \cdot \prod_{1 \leq i \leq \|\ell\|} \omega(\ell_i) \,\big|\ell\rangle \\ &= \sum_{K \in \mathbb{N}} \sum_{\vec{x} \in X^K} e^{-\lambda} \cdot \frac{\lambda^K}{K!} \cdot \prod_{1 \leq i \leq K} \omega(x_i) \,\big|\vec{x}\rangle \\ &= \sum_{K \in \mathbb{N}} \sum_{\vec{x} \in X^K} pois[\lambda](K) \cdot iid[K](\omega)(\vec{x}) \,\big|\vec{x}\rangle \;=\; iid[-](\omega)_*\Big(pois[\lambda]\Big). \end{aligned} \tag{3.52}$$

The latter formulation describes the Poisson-iid as an infinite Poisson-mixture of iid distributions. Thus, it is clear that it indeed forms a probability distribution, with probabilities adding up to one.

2 Similarly, the Poisson-multinomial is:

$$\begin{aligned} Pmn[\lambda](\omega) &:= \sum_{\varphi \in \mathcal{N}(supp(\omega))} e^{-\lambda} \cdot \frac{\lambda^{\|\varphi\|}}{\varphi\llparenthesis\rrparenthesis} \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \,\big|\varphi\rangle \\ &= \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} e^{-\lambda} \cdot \frac{\lambda^K}{K!} \cdot (\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \,\big|\varphi\rangle \\ &= \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} pois[\lambda](K) \cdot mn[K](\omega)(\varphi) \,\big|\varphi\rangle \;=\; mn[-](\omega)_*\Big(pois[\lambda]\Big). \end{aligned} \tag{3.53}$$

Example 3.10.2 We consider a hospital emergency department where patients arrive with a rate of 5 per hour. We have a simplified situation where only four types of diseases show up, labeled as a, b, c, d; they are collected in a set $X = \{a, b, c, d\}$. The probabilities of the diseases that appear are given by a distribution $\omega \in \mathcal{D}(X)$, of the form:

$$\omega \;=\; \tfrac{1}{6}|a\rangle + \tfrac{1}{8}|b\rangle + \tfrac{3}{8}|c\rangle + \tfrac{1}{3}|d\rangle.$$

The Poisson point process distribution $Pmn[5](\omega) \in \mathcal{D}_\infty(\mathcal{N}(\{a, b, c, d\}))$ now gives for each multiset of diseases the arrival probability. We elaborate this distribution for the first few sizes.

$$\begin{aligned} &Pmn[5](\omega) \\ &= pois[5](0) \cdot mn[0](\omega) + pois[5](1) \cdot mn[1](\omega) + pois[5](2) \cdot mn[2](\omega) \;+\; \cdots \\ &= e^{-5}\Big(\tfrac{5^0}{0!} \cdot mn[0](\omega) + \tfrac{5^1}{1!} \cdot mn[1](\omega) + \tfrac{5^2}{2!} \cdot mn[2](\omega) + \;\cdots\; \Big) \\ &= e^{-5}\Big(1\big|\mathbf{0}\rangle + \tfrac{5}{6}\big|1|a\rangle\rangle + \tfrac{5}{8}\big|1|b\rangle\rangle + \tfrac{15}{8}\big|1|c\rangle\rangle + \tfrac{5}{3}\big|1|d\rangle\rangle + \tfrac{25}{72}\big|2|a\rangle\rangle + \tfrac{25}{48}\big|1|a\rangle + 1|b\rangle\rangle \\ &\qquad + \tfrac{25}{128}\big|2|b\rangle\rangle + \tfrac{25}{16}\big|1|a\rangle + 1|c\rangle\rangle\tfrac{75}{64}\big|1|b\rangle + 1|c\rangle\rangle + \tfrac{225}{128}\big|2|c\rangle\rangle + \tfrac{25}{18}\big|1|a\rangle + 1|d\rangle\rangle \\ &\qquad + \tfrac{25}{24}\big|1|b\rangle + 1|d\rangle\rangle + \tfrac{25}{8}\big|1|c\rangle + 1|d\rangle\rangle + \tfrac{25}{18}\big|2|d\rangle\rangle + \;\cdots\; \Big) \\ &\approx 0.0067\big|\mathbf{0}\rangle + 0.0056\big|1|a\rangle\rangle + 0.0042\big|1|b\rangle\rangle + 0.013\big|1|c\rangle\rangle + 0.0112\big|1|d\rangle\rangle + 0.0023\big|2|a\rangle\rangle \\ &\qquad + 0.0035\big|1|a\rangle + 1|b\rangle\rangle + 0.0013\big|2|b\rangle\rangle + 0.011\big|1|a\rangle + 1|c\rangle\rangle + 0.0079\big|1|b\rangle + 1|c\rangle\rangle \\ &\qquad + 0.012\big|2|c\rangle\rangle + 0.0094\big|1|a\rangle + 1|d\rangle\rangle + 0.007\big|1|b\rangle + 1|d\rangle\rangle + 0.021\big|1|c\rangle + 1|d\rangle\rangle \\ &\qquad + 0.0094\big|2|d\rangle\rangle \;+\; \cdots \end{aligned}$$

We first show that the Poisson-iid and Poisson-multinomial distributions can als be expressed via tensor products.

Lemma 3.10.3 *Consider rate $\lambda \in \mathbb{R}_{\geq 0}$ and a distribution $\omega \in \mathcal{D}(X)$. For a natural multiset $\varphi \in \mathcal{N}(X)$ we can write the Poisson-multinomial as a tensor of the form:*

$$Pmn[\lambda](\omega)(\varphi) \;=\; \prod_{x \in supp(\omega)} pois[\omega(x) \cdot \lambda](\varphi(x))$$

When the set X is finite and when we identify $\mathcal{N}(X)$ with $\mathbb{N}^{|X|}$, as in Exercise 1.6.5, we can rewrite this equation as:

$$Pmn[\lambda](\omega) \;=\; \bigotimes_{x \in X} pois[\omega(x) \cdot \lambda] \;\in\; \mathcal{D}_\infty(\mathbb{N}^{|X|}) \;\cong\; \mathcal{D}_\infty(\mathcal{N}(X)).$$

Proof. Using that the probabilities $\omega(x)$, for $x \in supp(\omega)$, add up to one, we get:

$$\begin{aligned} Pmn[\lambda](\omega)(\varphi) &\overset{(3.53)}{=} e^{-\lambda} \cdot \frac{\lambda^{\|\varphi\|}}{\varphi\!\underset{\circ}{!}} \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \\ &= e^{-\lambda} \cdot \prod_{x \in X} \frac{(\omega(x) \cdot \lambda)^{\varphi(x)}}{\varphi(x)!} \\ &= \prod_{x \in X} e^{-\omega(x) \cdot \lambda} \cdot \frac{(\omega(x) \cdot \lambda)^{\varphi(x)}}{\varphi(x)!} \;=\; \prod_{x \in X} pois[\omega(x) \cdot \lambda](\varphi(x)). \end{aligned}$$

❑

The next result collects some basic properties of Poisson-iid and Poisson-multinomial distributions.

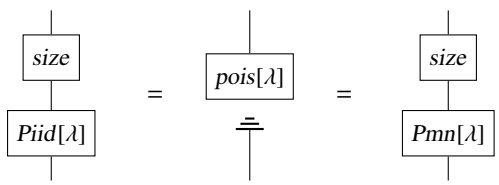
Lemma 3.10.4 *Consider the Poisson-iid and Poisson-multinomial functions $Piid[\lambda] \colon \mathcal{D}(X) \to \mathcal{D}_\infty(\mathcal{L}(X))$ and $Pmn[\lambda] \colon \mathcal{D}(X) \to \mathcal{D}_\infty(\mathcal{N}(X))$ introduced in Definition 3.10.1. They satisfy the following properties.*

1 *The are both natural in X.*

2 *Using the two size functions $size \colon \mathcal{L}(X) \to \mathbb{N}$ and $size \colon \mathcal{N}(X) \to \mathbb{N}$ satisfy:*

$$\mathcal{D}(size)\Big(Piid[\lambda](\omega)\Big) \;=\; pois[\lambda] \;=\; \mathcal{D}(size)\Big(Pmn[\lambda](\omega)\Big).$$

Equivalently, as string diagrams:



3 *Pushforward with uniform projection and frequentist learning yields the original distribution:*

$$unpr_*(Piid[\lambda](\omega)) \;=\; \omega \;=\; Flrn_*(Pmn[\lambda](\omega)).$$

4 *The intensity natural transformation ity from (3.51) satisfies:*

$$ity\Big(Pmn[\lambda](\omega)\Big) \;=\; \lambda \cdot \omega.$$

Proof. 1 We do the multinomial case and use that multinomials $mn[K]\colon \mathcal{D}(X) \to \mathcal{D}(\mathcal{N}[K](X))$ are natural in X. Thus, for a function $f\colon X \to Y$,

$$\begin{aligned}
(Pmn[\lambda] \circ \mathcal{D}(f))(\omega) &= \sum_{K\in\mathbb{N}} \sum_{\psi\in\mathcal{N}[K](Y)} pois[\lambda](K) \cdot mn[K]\big(\mathcal{D}(f)(\omega)\big)(\psi) \,\big|\psi\rangle \\
&= \sum_{K\in\mathbb{N}} \sum_{\psi\in\mathcal{N}[K](Y)} pois[\lambda](K) \cdot \mathcal{D}\mathcal{N}(f)\big(mn[K](\omega)\big)(\psi) \,\big|\psi\rangle \\
&= \sum_{K\in\mathbb{N}} \sum_{\psi\in\mathcal{N}[K](Y)} pois[\lambda](K) \cdot \left(\sum_{\varphi\in\mathcal{N}(f)^{-1}(\psi)} mn[K](\omega)(\varphi)\right) \big|\psi\rangle \\
&= \sum_{K\in\mathbb{N}} \sum_{\varphi\in\mathcal{N}[K](X)} pois[\lambda](K) \cdot mn[K](\omega)(\varphi) \,\big|\mathcal{N}(f)(\varphi)\rangle \\
&= (\mathcal{D}_\infty\mathcal{N}(f) \circ Pmn[\lambda])(\omega).
\end{aligned}$$

2 We now do the iid case:

$$\begin{aligned}
\mathcal{D}(size)(Piid[\lambda](\omega)) &= \sum_{K\in\mathbb{N}} \sum_{\vec{x}\in X^K} pois[\lambda](K) \cdot iid[K](\omega)(\vec{x}) \,\big|size(\vec{x})\rangle \\
&= \sum_{K\in\mathbb{N}} pois[\lambda](K) \cdot \left(\sum_{\vec{x}\in X^K} iid[K](\omega)(\vec{x})\right) \big|K\rangle \;=\; pois[\lambda].
\end{aligned}$$

3 For the multinomial case we use the equation $Flrn_*(mn[K](\omega)) = \omega$ from Theorem 3.3.4. For $x \in X$,

$$\begin{aligned}
Flrn_*(Pmn[\lambda](\omega))(x) &= \sum_{K\in\mathbb{N}} \sum_{\varphi\in\mathcal{N}[K](X)} pois[\lambda](K) \cdot mn[K](\omega)(\varphi) \cdot Flrn(\varphi)(x) \\
&= \sum_{K\in\mathbb{N}} pois[\lambda](K) \cdot \sum_{\varphi\in\mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot Flrn(\varphi)(x) \\
&= \sum_{K\in\mathbb{N}} pois[\lambda](K) \cdot Flrn_*(mn[K](\omega))(x) \\
&= \sum_{K\in\mathbb{N}} pois[\lambda](K) \cdot \omega(x) \;=\; \omega(x).
\end{aligned}$$

For the iid case we similarly use $unpr_*(iid[K](\omega)) = \omega$, see Exercise 2.4.9.

4 For $x \in X$ we get:

$$\begin{aligned}
ity\big(Pmn[\lambda](\omega)\big)(x) &= flat\big(Pmn[\lambda](\omega)\big)(x) \\
&= \sum_{K\in\mathbb{N}} \sum_{\varphi\in\mathcal{N}[K](X)} pois[\lambda](K) \cdot mn[K](\omega)(\varphi) \cdot \varphi(x) \\
&= \sum_{K\in\mathbb{N}} pois[\lambda](K) \cdot \sum_{\varphi\in\mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x) \\
&= \sum_{K\in\mathbb{N}} pois[\lambda](K) \cdot K \cdot \omega(x) \qquad \text{by Lemma 3.3.3} \\
&= \omega(x) \cdot \sum_{K\geq 1} e^{-\lambda} \cdot \frac{\lambda^K}{(K-1)!} \;=\; \lambda \cdot \omega(x) \cdot \sum_{K\geq 0} e^{-\lambda} \cdot \frac{\lambda^K}{K!} \;=\; \lambda \cdot \omega(x). \qquad \square
\end{aligned}$$

Next we use concatenation of lists and sums of natural multisets in a convolution result for Poisson-iid/multinomial distributions. It builds on closure under convolution for iid and for multinomials, see Lemma 2.7.8 and Exercise 3.3.10 (or Theorem 3.5.4, using marginalisation).

Proposition 3.10.5 *Poisson-iid and Poisson-multinomial distributions are closed under convolution, as expressed by the following two diagrams.*

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "\Delta"] \arrow[drr, bend right, "Piid[\lambda_1+\lambda_2]"'] & \mathcal{D}(X) \times \mathcal{D}(X) \arrow[r, "Piid[\lambda_1]\otimes Piid[\lambda_2]"] & \mathcal{D}(\mathcal{L}(X) \times \mathcal{L}(X)) \arrow[d, "\mathcal{D}(+)"] \\
& & \mathcal{D}(\mathcal{L}(X))
\end{tikzcd}$$

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "\Delta"] \arrow[drr, bend right, "Pmn[\lambda_1+\lambda_2]"'] & \mathcal{D}(X) \times \mathcal{D}(X) \arrow[r, "Pmn[\lambda_1]\otimes Pmn[\lambda_2]"] & \mathcal{D}(\mathcal{N}(X) \times \mathcal{N}(X)) \arrow[d, "\mathcal{D}(+)"] \\
& & \mathcal{D}(\mathcal{N}(X))
\end{tikzcd}$$

In the latter case one can alternatively say that for each distribution $\omega \in \mathcal{D}(X)$, the channel $Pmn[-](\omega)\colon \mathbb{R}_{\geq 0} \to \mathcal{D}_\infty(\mathcal{N}(X))$ is a map of (commutative) monoids.

Proof. We do the multinomial case both in abstract form and more concretely. First, by Exercise 3.3.10 we know that $mn[-](\omega)\colon \mathbb{N} \to \mathcal{D}(\mathcal{N}(X))$ is a map of monoids. By Exercise 2.7.10 the pushforward function $mn[-](\omega)_*\colon \mathcal{D}_\infty(\mathbb{N}) \to \mathcal{D}(\mathcal{N}(X))$ is then also a map of monoids. By Proposition 2.7.9, we can conclude that $Pmn[-](\omega)$ preserves the monoid structure, as composite of two monoid maps:

$$Pmn[-](\omega) \;=\; \Big(\mathbb{R}_{\geq 0} \xrightarrow{pois[-]} \mathcal{D}_\infty(\mathbb{N}) \xrightarrow{mn[-](\omega)_*} \mathcal{D}(\mathcal{N}(X))\Big).$$

The concrete argument uses closure of multinomials under convolution, from Exercise 3.3.10, and reasons as in the proof of Proposition 2.7.9.

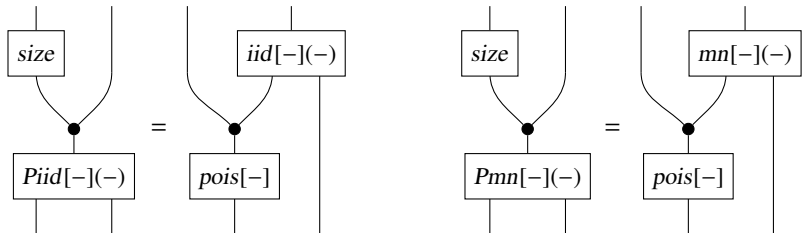
$$\begin{aligned}
&\mathcal{D}(+)\Big(Pmn[\lambda_1](\omega) \otimes Pmn[\lambda_2](\omega)\Big) \\
&= \sum_{K_1,K_2\in\mathbb{N}} \sum_{\varphi_1\in\mathcal{N}[K_1](X)} \sum_{\varphi_2\in\mathcal{N}[K_2](X)} e^{-\lambda_1}\cdot\frac{\lambda_1^{K_1}}{K_1!}\cdot e^{-\lambda_2}\cdot\frac{\lambda_2^{K_2}}{K_2!} \\
&\qquad\qquad \cdot mn[K_1](\omega)(\varphi_1)\cdot mn[K_2](\omega)(\varphi_2)\,\big|\varphi_1+\varphi_2\big\rangle \\
&= \sum_{L\in\mathbb{N}} \sum_{K\leq L} e^{-(\lambda_1+\lambda_2)}\cdot\frac{\lambda_1^K}{K!}\cdot\frac{\lambda_2^{L-K}}{(L-K)!} \\
&\qquad \cdot \sum_{\varphi\in\mathcal{N}[K](X)} \sum_{\psi\in\mathcal{N}[L-K](X)} mn[K](\omega)(\varphi)\cdot mn[L-K](\omega)(\psi)\,\big|\varphi+\psi\big\rangle \\
&= \sum_{L\in\mathbb{N}} \frac{e^{-(\lambda_1+\lambda_2)}}{L!}\cdot\left(\sum_{K\leq L}\binom{L}{K}\cdot\lambda_1^K\cdot\lambda_2^{L-K}\right)\cdot\mathcal{D}(+)\Big(mn[K](\omega)\otimes mn[L-K](\omega)\Big) \\
&= \sum_{L\in\mathbb{N}} \frac{e^{-(\lambda_1+\lambda_2)}}{L!}\cdot(\lambda_1+\lambda_2)^L\cdot mn[L](\omega) \;=\; Pmn[\lambda_1+\lambda_2](\omega).
\end{aligned}$$

❑

For the next result we shall now use the Poisson-iid / multinomial and the iid / multinomial maps as binary functions:

$$\begin{array}{ll}
\mathbb{R}_{>0}\times\mathcal{D}(X) \xrightarrow{Piid[-](-)} \mathcal{D}_\infty(\mathcal{L}(X)) & \qquad \mathbb{N}\times\mathcal{D}(X) \xrightarrow{iid[-](-)} \mathcal{D}_\infty(\mathcal{L}(X)) \\
\mathbb{R}_{>0}\times\mathcal{D}(X) \xrightarrow{Pmn[-](-)} \mathcal{D}_\infty(\mathcal{N}(X)) & \qquad \mathbb{N}\times\mathcal{D}(X) \xrightarrow{mn[-](-)} \mathcal{D}_\infty(\mathcal{N}(X))
\end{array}$$

Theorem 3.10.6 *The two equations below between string diagrams hold.*



By marginalising out the left part we obtain a string-diagrammatic description of the Poisson-iid / multinomial channels. Marginalising out the right parts gives the equations in Lemma 3.10.4 (2).

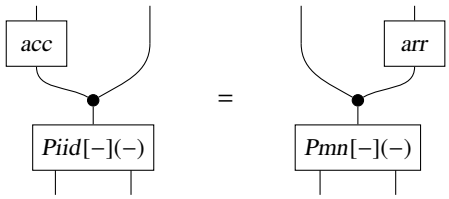
Proof. For $\lambda \in \mathbb{R}_{>0}$ and $\omega \in \mathcal{D}(X)$,

$$\begin{aligned}
\langle size, id\rangle_*(Piid[\lambda](\omega)) &= \sum_{K\in\mathbb{N}} \sum_{\vec{x}\in X^K} pois[\lambda](K)\cdot iid[K](\omega)(\vec{x}) \,\Big|\, K, \vec{x}\Big\rangle \\
&= \langle id, iid[-](\omega)\rangle_*(pois[\lambda]). \\
\langle size, id\rangle_*(Pmn[\lambda](\omega)) &= \sum_{K\in\mathbb{N}} \sum_{\varphi\in\mathcal{N}[K](X)} pois[\lambda](K)\cdot mn[K](\omega)(\varphi) \,\Big|\, K, \varphi\Big\rangle \\
&= \langle id, mn[-](\omega)\rangle_*(pois[\lambda]).
\end{aligned}$$

❑

Recall from Theorem 3.3.1 (2) the diagrammatic (sufficient-statistic) relationship between iid and multinomial via accumulation and arrangement. This also works for the Poisson versions.

Theorem 3.10.7 *Poisson-iid and Poisson-multinomial channels are related as follows.*



Proof. We reason essentially as in the proof of Theorem 3.3.1.

$$\begin{aligned}
\langle id, arr\rangle_*(Pmn[\lambda](\omega)) &= \sum_{K\in\mathbb{N}} \sum_{\varphi\in\mathcal{N}[K](X)} \sum_{\vec{x}\in X^K} arr(\varphi)(\vec{x})\cdot Pmn[\lambda](\omega)(\vec{x}) \,\big|\, \varphi, \vec{x}\rangle \\
&= \sum_{K\in\mathbb{N}} \sum_{\varphi\in\mathcal{N}[K](X)} \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)}\cdot pois[\lambda](K)\cdot mn[K](\omega)(\varphi) \,\big|\, \varphi, \vec{x}\rangle \\
&= \sum_{K\in\mathbb{N}} \sum_{\varphi\in\mathcal{N}[K](X)} \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)}\cdot pois[\lambda](K)\cdot (\varphi)\cdot \prod_{y\in X} \omega(y)^{\varphi(y)} \,\big|\, \varphi, \vec{x}\rangle \\
&= \sum_{K\in\mathbb{N}} \sum_{\varphi\in\mathcal{N}[K](X)} \sum_{\vec{x}\in acc^{-1}(\varphi)} pois[\lambda](K)\cdot \prod_{1\le i\le K} \omega(x_i) \,\big|\, \varphi, \vec{x}\rangle \\
&= \sum_{K\in\mathbb{N}} \sum_{\varphi\in\mathcal{N}[K](X)} \sum_{\vec{x}\in X^K} \langle acc\rangle(\vec{x})(\varphi)\cdot pois[\lambda](K)\cdot iid[K](\omega)(\vec{x}) \,\big|\, \varphi, \vec{x}\rangle \\
&= \langle acc, id\rangle_*(Piid[\lambda](\omega)).
\end{aligned}$$

❑

We finish with a characterisation of Poisson-multinomial distributions in terms of elementary properties — in the style of the characterisation of multinomial distributions in Exercise 3.3.4. The analogous characterisation for Poisson-iid is left as an exercise below.

Definition 3.10.8 Let X be a finite set with a distribution $\Omega \in \mathcal{D}_\infty(\mathcal{N}(X))$. We call this Ω a *counting distribution of multisets* if it satisfies the following four requirements.

(a) $0 < \Omega(\mathbf{0}) < 1$, where $\mathbf{0} \in \mathcal{N}(X)$ is the empty multiset;

(b) $\sum_{x \in X} \Omega(1|x\rangle) = -\Omega(\mathbf{0}) \cdot \ln(\Omega(\mathbf{0}))$;

(c) $\Omega(n|x\rangle) = \dfrac{\Omega(1|x\rangle)^n}{n! \cdot \Omega(\mathbf{0})^{n-1}}$, for each $x \in X$ and $n > 1$;

(d) $\Omega(\varphi + \psi) = \dfrac{\Omega(\varphi) \cdot \Omega(\psi)}{\Omega(\mathbf{0})}$, for multisets $\varphi, \psi \in \mathcal{N}(X)$ with disjoint support: $supp(\varphi) \cap supp(\psi) = \emptyset$.

Theorem 3.10.9 *Let X be a finite set.*

1 *For a distribution $\omega \in \mathcal{D}(X)$ and a rate $\lambda > 0$, the Poisson-multinomial $Pmn[\lambda](\omega) \in \mathcal{D}_\infty(\mathcal{N}(X))$ is a counting distribution of multisets: it satisfies requirements (a) – (d) in Definition 3.10.8.*

2 *If $\Omega \in \mathcal{D}_\infty(\mathcal{N}(X))$ is a counting distribution of multisets, then it is of the form $\Omega = Pmn[\lambda](\omega)$ for a unique rate $\lambda > 0$ and distribution $\omega \in \mathcal{D}(X)$.*

Proof. 1 Let $\lambda \in \mathbb{R}_{>0}$ and $\omega \in \mathcal{D}(X)$ be given and abbreviate $\Omega := Pmn[\lambda](\omega) \in \mathcal{D}_\infty(\mathcal{N}(X))$. We check the four requirements in Definition 3.10.8.

(a) The empty multiset $\mathbf{0} \in \mathcal{N}(X)$ satisfies:

$$\Omega(\mathbf{0}) = Pmn[\lambda](\omega)(\mathbf{0}) \overset{(3.53)}{=} e^\lambda \cdot \frac{\lambda^0}{0!} \cdot \prod_{x \in X} \omega(x)^0 = e^{-\lambda}.$$

Since $\lambda > 0$, where have $0 < e^{-\lambda} < 1$, as required.

(b) For an element $x \in X$, the singleton multiset $1|x\rangle \in \mathcal{N}(X)$ satisfies:

$$\Omega(1|x\rangle) = Pmn[\lambda](\omega)(1|x\rangle) \overset{(3.53)}{=} e^\lambda \cdot \frac{\lambda^1}{1!} \cdot \omega(x)^1 = e^{-\lambda} \cdot \lambda \cdot \omega(x).$$

Thus:

$$\sum_{x \in X} \Omega(1|x\rangle) = e^{-\lambda} \cdot \lambda \cdot \sum_{x \in X} \omega(x) = e^{-\lambda} \cdot \lambda = -\Omega(\mathbf{0}) \cdot \ln(\Omega(\mathbf{0})).$$

(c) Similarly, for $n > 1$,

$$\Omega(n|x\rangle) = Pmn[\lambda](\omega)(n|x\rangle) \overset{(3.53)}{=} e^{-\lambda} \cdot \frac{\lambda^n}{n!} \cdot \omega(x)^n = e^{-\lambda} \cdot \frac{\Omega(1|x\rangle)^n}{n! \cdot (e^{-\lambda})^n} = \frac{\Omega(1|x\rangle)^n}{n! \cdot \Omega(\mathbf{0})^{n-1}}.$$

(d) Let $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[L](X)$ have disjoint support. By Exercise 3.3.3:

$$\begin{aligned}
\Omega(\varphi + \psi) &= pois[\lambda](K+L) \cdot mn[K+L](\varphi + \psi) \\
&= e^{-\lambda} \cdot \frac{\lambda^{K+L}}{(K+L)!} \cdot \binom{K+L}{K} \cdot mn[K](\omega)(\varphi) \cdot mn[L](\omega)(\psi) \\
&= \frac{1}{e^{-\lambda}} \cdot e^{-\lambda} \cdot \frac{\lambda^K}{K!} \cdot mn[K](\omega)(\varphi) \cdot e^{-\lambda} \cdot \frac{\lambda^L}{L!} \cdot mn[L](\omega)(\psi) \\
&= \frac{Pmn[\lambda](\omega)(\varphi) \cdot Pmn[\lambda](\omega)(\psi)}{e^{-\lambda}} = \frac{\Omega(\varphi) \cdot \Omega(\psi)}{\Omega(\mathbf{0})}.
\end{aligned}$$

2 In the other direction, let $\Omega \in \mathcal{D}_\infty(\mathcal{N}(X))$ be a counting distribution of multisets, so that requirements (a) – (d) hold. Using $0 < \Omega(\mathbf{0}) < 1$ from (a), we can take $\lambda := -\ln(\Omega(\mathbf{0})) \in \mathbb{R}_{>0}$. We then define a distribution $\omega \in \mathcal{D}(X)$, by setting for $x \in X$,

$$\omega(x) := \frac{\Omega(1|x\rangle)}{-\Omega(\mathbf{0}) \cdot \ln(\Omega(\mathbf{0}))} = \frac{\Omega(1|x\rangle)}{\lambda \cdot e^{-\lambda}}.$$

This yields a distribution by (b):

$$\sum_{x\in X} \omega(x) = \sum_{x\in X} \frac{\Omega(1|x\rangle)}{-\Omega(\mathbf{0}) \cdot \ln(\Omega(\mathbf{0}))} = \frac{\sum_{x\in X} \Omega(1|x\rangle)}{-\Omega(\mathbf{0}) \cdot \ln(\Omega(\mathbf{0}))} \overset{(b)}{=} 1.$$

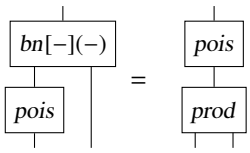
Now, for a multiset $\varphi \in \mathcal{N}(X)$,

$$\begin{aligned}
\Omega(\varphi) = \Omega\Big(\textstyle\sum_{x\in X} \varphi(x)|x\rangle\Big) &\overset{(d)}{=} \Omega(\mathbf{0}) \cdot \prod_{x\in X} \frac{\Omega(\varphi(x)|x\rangle)}{\Omega(\mathbf{0})} \\
&\overset{(c)}{=} e^{-\lambda} \cdot \prod_{x\in X} \frac{\Omega(1|x\rangle)^{\varphi(x)}}{\varphi(x)! \cdot \Omega(\mathbf{0})^{\varphi(x)}} \\
&= e^{-\lambda} \cdot \frac{1}{\prod_{x\in X} \varphi(x)!} \cdot \prod_{x\in X} \lambda^{\varphi(x)} \cdot \omega(x)^{\varphi(x)} \\
&= e^{-\lambda} \cdot \frac{1}{\varphi!} \cdot \lambda^{\sum_{x\in X} \varphi(x)} \cdot \prod_{x\in X} \omega(x)^{\varphi(x)} \overset{(3.53)}{=} Pmn[\lambda](\omega)(\varphi). \qquad \square
\end{aligned}$$

Exercises

3.10.1 Let $r \in [0, 1]$ and $\lambda \in \mathbb{R}_{\geq 0}$. Show that:

$$bn[-](r)_*(pois[\lambda]) = pois[r \cdot \lambda] \in \mathcal{D}_\infty(\mathbb{N}) \qquad i.e$$



This distribution is described in (Rice, 2006, §3.5.1) and in (Jaynes, 2003, §6.11) as an imperfect particle counter with successful detection rate r and expected number of particles λ per time unit. There is more about such counters in Section 7.6.

3.10.2 Check that the point process map $Pmn[\lambda] \colon \mathcal{D} \Rightarrow \mathcal{D}_\infty\mathcal{N}$ is not a map of monads, *e.g.* because units are not preserved.

3.10.3 Let X be a finite set. Call $\Omega \in \mathcal{D}_\infty(\mathcal{L}(X))$ a *counting distribution of lists* if:

(a) $0 < \Omega([]) < 1$, where $[] \in \mathcal{L}(X)$ is the empty list;

(b) $\displaystyle\sum_{x\in X} \Omega([x]) = -\Omega([]) \cdot \ln(\Omega([]))$;

(c) $\displaystyle\Omega(\ell_1 +\!\!+ \ell_2) = \frac{\Omega(\ell_1) \cdot \Omega(\ell_2)}{\Omega([]) \cdot \binom{\|\ell_1\|+\|\ell_2\|}{\|\ell_1\|}}$.

Show that these three points characterise Poisson-iid distributions.

4

Observables and validity

So far we have seen (discrete probability) distributions as formal convex sums $\sum_i r_i|x_i\rangle$ in $\mathcal{D}(X)$ and (probabilistic) channels $X \rightsquigarrow Y$, as functions $X \to \mathcal{D}(Y)$, describing probabilistic states and computations. This section develops the tools to reason about such distributions and channels, via what are called *observables*. They are functions from a set / sample space X to (a subset of) the real numbers $\mathbb{R}$ that associate some 'observable' numerical information with an element $x \in X$. The following table gives an overview of terminology and types, where X is a set (used as sample space).

name	**type**
observable / utility function	$X \to \mathbb{R}$
factor / potential function	$X \to \mathbb{R}_{\geq 0}$
(fuzzy) predicate / (soft / uncertain) evidence	$X \to [0, 1]$
sharp predicate / event	$X \to \{0, 1\}$

(4.1)

We shall use the term of 'observable' as generic expression for all the entries in this table. A function $X \to \mathbb{R}$ is thus the most general type of observable, and a sharp predicate $X \to \{0, 1\}$ is the most specific one. Predicates are the most appropriate observables for probabilistic logical reasoning. Often attention is restricted to subsets $U \subseteq X$ as predicates (or events Ross (2007)), but here, in this book, the fuzzy versions $X \to [0, 1]$ are the default. Such fuzzy predicates may also be called belief functions — or effects, in a quantum setting. A technical reason for using fuzzy, $[0, 1]$-valued predicates instead of sharp, $\{0, 1\}$-valued ones, is that these fuzzy predicates are closed under predicate transformation, and the sharp predicates are not, see below for details.

The commonly used notion of *random variable* can now be described as a *pair*, consisting of an observable together with a state (distribution), both on the same sample space. Often, this state is left implicit; it may be obvious in a particular context what it is. But leaving the state implicit may also be confusing, for instance when we deal with two random variables and we need to make explicit wether they involve different states, or share a state. Like elsewhere in this book, we like to be explicit about the state(s) that we are using.

This chapter starts with the definition of what can be seen as probabilistic truth $\omega \models p$, namely the validity of an observable p in a state ω. It is the expected value of p in ω. We shall see that many basic concepts can be defined in terms of validity, including mean, average, entropy,

distance, (co)variance. The algebraic, logical and categorical structure of the various observables in Table 4.1 will be investigated in Section 4.4.

In the previous chapter we have seen that a state ω on the domain X of a channel $c\colon X \multimap Y$ can be transformed into a state $c_*(\omega)$ on the channel's codomain Y. Analogous to such state transformation c_* there is also observable transformation c^*, acting in the opposite direction: for an observable q on the codomain Y of a channel $c\colon X \multimap Y$, there is a transformed observable $c^*(q)$ on the domain X. When c^* is applied to predicates, it is called predicate transformation. It is a basic operation in programming logics. These transformations in different directions are an aspect of the duality between states and predicates. These two transformations correspond to Schrödinger's (forward) Heisenberg's (backward) approach in quantum foundations, see Jacobs (2017c). At the end of this chapter, in Section 4.8, validity is used to give (dual) formulations of distances between states and between predicates. Via this distance function we can make important properties precise, like: each distribution can be approximated, with arbitrary precision, via frequentist learning of natural multisets. Technically, the 'fractional' distributions of the form $\mathit{Flrn}(\varphi)$, for natural multisets φ, form a dense subset of the (complete metric) space of all distributions.

As an aside: in this book we use 'distribution' and 'state' synonymously. In the current logical context there is a slight preference to use 'state' because the expressions predicate transformation and state transformation are well-established for reasoning about computations.

4.1 Validity

This section introduces the basic facts and terminology for observables, as described in Table 4.1 and defines their validity in a state / distribution. Recall that we write Y^X for the set of functions from X to Y. We will use notations:

- $\mathit{Obs}(X) := \mathbb{R}^X$ for the set of observables on a set X;
- $\mathit{Fact}(X) := (\mathbb{R}_{\geq 0})^X$ for the set of factors on X;
- $\mathit{Pred}(X) := [0,1]^X$ for the set of predicates on X;
- $\mathit{SPred}(X) := \{0,1\}^X$ for the set of sharp predicates (events) on X.

There are inclusions:

$$\mathit{SPred}(X) \subseteq \mathit{Pred}(X) \subseteq \mathit{Fact}(X) \subseteq \mathit{Obs}(X).$$

The first set $\mathit{SPred}(X) = \{0,1\}^X$ of sharp predicates can be identified with the powerset $\mathcal{P}(X)$ of subsets of X, see below. We first define some special observables.

Definition 4.1.1 Let X be an arbitrary set.

1 For a subset $U \subseteq X$ we write $\mathbf{1}_U\colon X \to \{0,1\}$ for the *characteristic function* of U, defined as:

$$\mathbf{1}_U(x) := \begin{cases} 1 & \text{for } x \in U \\ 0 & \text{otherwise.} \end{cases}$$

This characteristic function $\mathbf{1}_U$, with type $X \to [0,1]$, is the (sharp) predicate associated with the subset $U \subseteq X$. The mapping $U \mapsto \mathbf{1}_U$ forms the isomorphism $\mathcal{P}(X) \xrightarrow{\cong} \{0,1\}^X$ that we just mentioned.

2 We use special notation for two extreme cases $U = X$ and $U = \emptyset$, giving the *truth* predicate $\mathbf{1}\colon X \to [0,1]$ and the *falsity* predicate $\mathbf{0}\colon X \to [0,1]$ on X. Explicitly:

$$\mathbf{1} := \mathbf{1}_X \quad \text{and} \quad \mathbf{0} := \mathbf{1}_\emptyset \qquad \text{so that} \qquad \mathbf{1}(x) = 1 \quad \text{and} \quad \mathbf{0}(x) = 0,$$

for all $x \in X$.

3 For a singleton subset $\{x\}$ we simply write $\mathbf{1}_x$ for $\mathbf{1}_{\{x\}}$. Such functions $\mathbf{1}_x\colon X \to [0,1]$ are also called *point predicates*, where the element $x \in X$ is seen as a point.

Recall that a state of the form $\mathit{unit}(x) = 1|x\rangle$ is called a point state or a point distribution.

4 There is a (sharp) equality predicate $\mathit{Eq}\colon X \times X \to [0,1]$ defined in the obvious way as:

$$\mathit{Eq}(x, x') := \begin{cases} 1 & \text{if } x = x' \\ 0 & \text{otherwise.} \end{cases}$$

5 If the set X can be mapped into $\mathbb{R}$ in an obvious manner, we write this inclusion function as $\mathit{incl}_X\colon X \hookrightarrow \mathbb{R}$ and may consider it as an observable on X. This applies for instance if $X = \boldsymbol{n} = \{0, 1, \ldots, n-1\}$.

We now come to the basic definition of validity.

Definition 4.1.2 Let X be a set with an observable $p\colon X \to \mathbb{R}$ on X.

1 For a distribution / state $\omega = \sum_i r_i|x_i\rangle$ on X we define its validity $\omega \models p$ as:

$$\omega \models p := \sum\nolimits_i r_i \cdot p(x_i) = \sum_{x \in X} \omega(x) \cdot p(x). \tag{4.2}$$

2 If X is a finite set, with size $|X| \in \mathbb{N}$, one can define the *average* $\mathit{avg}(p)$ of the observable p as its validity in the uniform state unif_X on X, *i.e.*,

$$\mathit{avg}(p) := \mathit{unif}_X \models p = \sum_{x \in X} \frac{p(x)}{|X|}.$$

Notice that validity $\omega \models p$ involves a finite sum, since a distribution $\omega \in \mathcal{D}(X)$ has by definition a finite support. The sample space X may well be infinite. Validity is often called *expected value* and written for instance as:

$$\omega \models p = \mathop{\mathbb{E}}_{x \sim \omega} p(x).$$

This $\mathbb{E}$-expression uses the sampling notation $\sim$ as in (2.17). We prefer to use the simpler notation $\omega \models p$ in which we do not have to work with bound variables, like x in $x \sim \omega$ and $p(x)$.

The even simpler notation $\mathbb{E}[p]$ for $\omega \models p$ frequently occurs, where the state / distribution ω is left implicit. This is incovenient, especially in situations where we change the state, for instance via state transformation or via updating. That's why we avoid the common $\mathbb{E}$-notation.

The validity $\omega \models p$ is non-negative (is in $\mathbb{R}_{\geq 0}$) if p is a factor and lies in the unit interval $[0,1]$ if p is a predicate (whether sharp or not). The fact that the multiplicities r_i in a distribution $\omega = \sum_i r_i|x_i\rangle$ add up to one means that the validity $\omega \models \mathbf{1}$ of truth is one. Notice that for a point predicate one has $\omega \models \mathbf{1}_x = \omega(x)$ and similarly, for a point state, $1|x\rangle \models p = p(x)$.

For a fixed state $\omega \in \mathcal{D}(X)$ we can view $\omega \models (-)$ as a function $\mathit{Pred}(X) \to [0,1]$ that assigns a likelihood to a belief function. Thus, validity $\omega \models (-)$ becomes a predicate on predicates — or a factor on factors, or an observable on observables. This is at the heart of the Bayesian interpretation of probability, see Section 6.10 for more details.

As an aside, we typically do not write brackets in equations like $(\omega \models p) = a$, but use the convention that $\models$ has higher precedence than $=$, so that the equation can simply be written as $\omega \models p = a$.

Example 4.1.3 1 Let $\mathit{flip}(\frac{3}{10}) = \frac{3}{10}|1\rangle + \frac{7}{10}|0\rangle$ be a biased coin. Suppose there is a game where you can throw the coin and win €100 if head (1) comes up, but you lose €50 if the outcome is tail (0). Is it a good idea to play the game?

The possible gain can be formalised as an observable $v\colon \{0,1\} \to \mathbb{R}$ with $v(0) = -50$ and $v(1) = 100$. Such a map v is often called a utility function. We get an anwer to the above question by computing the validity:

$$\begin{aligned} \mathit{flip}(\tfrac{3}{10}) \models v \;=\; \sum_{x \in \{0,1\}} \mathit{flip}(\tfrac{3}{10})(x) \cdot v(x) \;&=\; \mathit{flip}(\tfrac{3}{10})(0) \cdot v(0) + \mathit{flip}(\tfrac{3}{10})(1) \cdot v(1) \\ &=\; \tfrac{7}{10} \cdot -50 + \tfrac{3}{10} \cdot 100 \;=\; -35 + 30 \;=\; -5. \end{aligned}$$

Hence it is wiser not to play.

2 We consider a slightly different situation involving a lottery with a 4% chance of winning, as given by the distribution $\rho = \frac{1}{25}|W\rangle + \frac{24}{25}|L\rangle$. The prize for the winner is €500. You own a lottery ticket and are given the option: play the lottery, or sell the ticket for €15.

The expected gain of playing is $\rho \models v = \frac{1}{25} \cdot 500 =$ €20, via the utility function $v(W) = 500$ and $v(L) = 0$. Hence it is better to play.

Now let's assume that a poor and a rich person are in the situation as just sketched, where the poor person's owns €10 and the rich person owns €1000. Intuitively, their risk assesments are different. This can be accomodated by using a different utility function $u_o\colon \{W, L\} \to \mathbb{R}$, of the form:

$$u_o(W) \;=\; \ln(o + 500) \qquad\qquad u_o(L) \;=\; \ln(o).$$

The parameter o represents the owned amount. The natural logarithm ln is used in this utility function. It is a monotonely increasing function, with positive first derivative, but with negative second derivative. This captures a decreasing increase. It corresponds to the idea that for a rich person an increase of wealth means less than to a poor person.

With this utility function u_o we now get the following situation.

	poor ($o = 10$)	rich ($o = 1000$)
sell	$\ln(10 + 15) = 3.22$	$\ln(1000 + 15) = 6.922$
play	$\rho \models u_{10} = 2.46$	$\rho \models u_{1000} = 6.923$

Thus, for a poor person it is wiser to sell, because this gives a certain increase of wealth. This certain increase is relatively more important than for a rich individual.

3 Suppose that we claim that in a throw of a (fair) dice the outcome is even. How likely is this claim? We formalise it as a (sharp) predicate $e\colon \mathit{pips} \to [0,1]$ with $e(1) = e(3) = e(5) = 0$ and $e(2) = e(4) = e(6) = 1$. Then, as expected:

$$\mathit{dice} \models e \;=\; \sum_{x \in \mathit{pips}} \mathit{dice}(x) \cdot e(x) \;=\; \tfrac{1}{6} \cdot 0 + \tfrac{1}{6} \cdot 1 + \tfrac{1}{6} \cdot 0 + \tfrac{1}{6} \cdot 1 + \tfrac{1}{6} \cdot 0 + \tfrac{1}{6} \cdot 1 \;=\; \tfrac{1}{2}.$$

We now consider a non-sharp 'evenish' predicate $p\colon \mathit{pips} \to [0,1]$. It expresses that the even

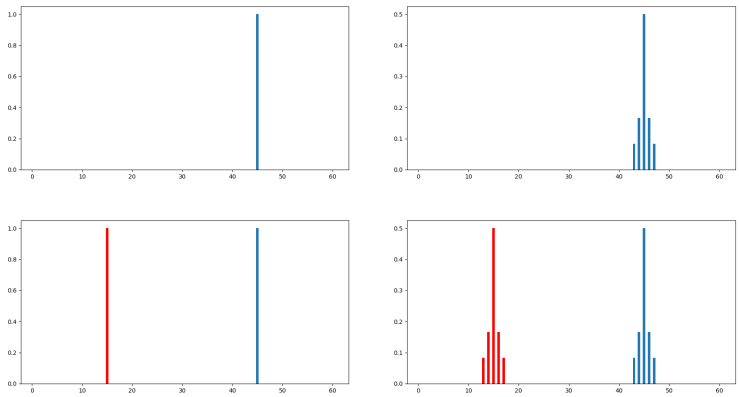


Figure 4.1 Bus arrival times, in minutes within one hour, of one bus at the top, and two buses at the bottom. On the left the buses are punctual, but on the right the arrival times of the buses are given by a distribution. Does that affect the expected waiting time? See Example 4.1.4 for details.

pips are more likely than the odd ones, where the precise likelihoods are determined by:

$$\begin{array}{lll} p(1) = \frac{1}{10} & p(2) = \frac{9}{10} & p(3) = \frac{3}{10} \\ p(4) = \frac{8}{10} & p(5) = \frac{2}{10} & p(6) = \frac{7}{10}. \end{array}$$

This new evenish claim p happens to be equally probable as the even claim e, since:

$$dice \models p = \frac{1}{6} \cdot \frac{1}{10} + \frac{1}{6} \cdot \frac{9}{10} + \frac{1}{6} \cdot \frac{3}{10} + \frac{1}{6} \cdot \frac{8}{10} + \frac{1}{6} \cdot \frac{2}{10} + \frac{1}{6} \cdot \frac{7}{10} = \frac{1+9+3+8+2+7}{60} = \frac{30}{60} = \frac{1}{2}.$$

We include another, extended example with factors defined on joint distributions.

Example 4.1.4 We look at the expected time that one has to wait for a bus, in different scenarios. We consider a time frame of one hour, chopped up in 60 minutes, in a set $H = \{1, 2, \ldots, 60\}$. The minute of an indivudual's arrival at a particular bus stop is given by a uniform distribution $unif_H = \sum_{i \in H} \frac{1}{60}|i\rangle$ on H.

1 We first look at the scenario of a single bus arriving at the bus stop, exactly at the 45-th minute. We assume that the bus halts there for one minute, and then leaves again, so when a passenger arrives at minute 45, the waiting time is zero. If the passenger arrives at minute 44, the waiting time is 1, *etc.* For arrivals after minute 45, there is no more bus — since we consider one hour only — and the time waiting for the bus is considered to be 0. This is an interpretation, for instance corresponding to the passenger's choice to return home and to not travel at all.

We consider the combination of the arrivals of the passenger and of the bus via a joint distribution $unif_H \otimes 1|45\rangle$ on $H \times H$. Explicitly, it is $\frac{1}{60}|1, 45\rangle + \cdots + \frac{1}{60}|60, 45\rangle \in \mathcal{D}(H \times H)$.

The bus waiting time is captured via factor $p\colon H \times H \to \mathbb{N}$, given by:

$$p(i, b) := \begin{cases} b - i & \text{if } i \leq b \\ 0 & \text{otherwise.} \end{cases}$$

In the first case the passenger arrives at minute $i \leq b$ before the arrival minute b of the bus.

The waiting time is thus $b - i$ minutes. Now we can compute the expected arrival time, via the sum formula of Proposition 1.2.6 (1).

$$\begin{aligned} \mathit{unif}_H \otimes 1|45\rangle \models p \;&=\; \sum_{(i,b)\in H\times H} \frac{p(i,b)}{60} \;=\; \sum_{1\leq i\leq 45} \frac{45-i}{60} \\ &=\; \sum_{1\leq i\leq 44} \frac{i}{60} \;=\; \frac{44\cdot 45}{2\cdot 60} \;=\; \frac{33}{2} \;=\; 16.5. \end{aligned}$$

Thus, the expected waiting time is 16.5 minutes. It is not 22.5, that is half of 45, because we have defined the waiting time to be zero when there is no more bus, that is, after 45 minutes.

2 The question that we are interested in is the following. What happens with the expected waiting time when the bus does not arrive precisely at minute 45, but *around* minute 45? We model the bus arrival time now via the distribution:

$$\nu \;:=\; \tfrac{1}{12}|43\rangle + \tfrac{1}{6}|44\rangle + \tfrac{1}{2}|45\rangle + \tfrac{1}{6}|46\rangle + \tfrac{1}{12}|47\rangle.$$

The two distributions for the punctual bus and for the non-punctual bus are in Figure 4.1, at the top.

Our new joint distribution is $\mathit{unif}_H \otimes \nu \in \mathcal{D}(H\times H)$. What is the expected waiting time now, as given by the validity of p now? It turns out be only slightly higher than for the punctual bus. It is computed as:

$$\begin{aligned} \mathit{unif}_H \otimes \nu \models p \;&=\; \frac{1}{12}\left(\sum_{1\leq i\leq 43} \frac{43-i}{60}\right) + \frac{1}{6}\left(\sum_{1\leq i\leq 44} \frac{44-i}{60}\right) + \frac{1}{2}\left(\sum_{1\leq i\leq 45} \frac{45-i}{60}\right) \\ &\qquad + \frac{1}{6}\left(\sum_{1\leq i\leq 46} \frac{46-i}{60}\right) + \frac{1}{12}\left(\sum_{1\leq i\leq 47} \frac{47-i}{60}\right) \\ &=\; \frac{42\cdot 43}{12\cdot 2\cdot 60} + \frac{43\cdot 44}{6\cdot 2\cdot 60} + \frac{44\cdot 45}{2\cdot 2\cdot 60} + \frac{45\cdot 46}{6\cdot 2\cdot 60} + \frac{46\cdot 47}{12\cdot 2\cdot 60} \\ &=\; \frac{1981}{120} \;=\; 16.508333\cdots \end{aligned}$$

Why the difference? One may have thought that the differences in waiting times cancel each other out. An intuitive explanation for the increase in expected waiting time is that the passenger is more likely to arrive in a period in which is the bus is late, than in a period where the bus is early.

3 Now suppose there are two buses, both punctual, at the 15-th and 45-th minute. We then use as joint distribution $\mathit{unif}_H \otimes 1|15\rangle \otimes 1|45\rangle$ on $H\times H\times H$. The arrival time factor $q\colon H\times H\times H \to \mathbb{N}$ now is:

$$q(i,b_1,b_2) \;:=\; \begin{cases} b_1 - i & \text{if } i \leq b_1 \text{ and } \big(b_2 < i \text{ or } b_1 \leq b_2\big) \\ b_2 - i & \text{if } i \leq b_2 \text{ and } \big(b_1 < i \text{ or } b_2 \leq b_1\big) \\ 0 & \text{otherwise.} \end{cases}$$

The first clause describes the situation where the passenger arrives at moment i, before the arrival of bus 1 (given by b_1), with bus 2 either already passed ($b_2 < i$) or arriving later than bus 1 ($b_1 \leq b_2$). The second clause is symmetric. We now get as expected waiting time:

$$\mathit{unif}_H \otimes 1|15\rangle \otimes 1|45\rangle \models q \;=\; \sum_{1\leq i\leq 15} \frac{15-i}{60} + \sum_{16\leq i\leq 45} \frac{45-i}{60} \;=\; \frac{14\cdot 15 + 29\cdot 30}{2\cdot 60} \;=\; 9.$$

Clearly, with two buses the waiting is less, as expected.

4 What if the two buses are both not punctual, with distributions ν_1 and ν_2 around 15 and 45, see Figure 4.1, at the bottom, and:

$$\begin{aligned} \nu_1 &:= \tfrac{1}{12}|13\rangle + \tfrac{1}{6}|14\rangle + \tfrac{1}{2}|15\rangle + \tfrac{1}{6}|16\rangle + \tfrac{1}{12}|17\rangle \\ \nu_2 &:= \tfrac{1}{12}|43\rangle + \tfrac{1}{6}|44\rangle + \tfrac{1}{2}|45\rangle + \tfrac{1}{6}|46\rangle + \tfrac{1}{12}|47\rangle. \end{aligned}$$

Again there is a (slightly) longer expected waiting time than with two punctual buses:

$$unif_H \otimes \nu_1 \otimes \nu_2 \models q \;=\; \frac{361}{40} \;=\; 9.025.$$

This concludes the bus illustration. Exercise 4.1.4 contains two follow-up questions.

The next result shows that validity of an observable in a pushforward distribution can also be expressed as validity in the original distribution, of a translated observable. This equality of validities will appear later in Proposition 4.5.3, in more general form, for channels, instead of for functions only.

Lemma 4.1.5 *Consider an arbitrary distribution $\omega \in \mathcal{D}(X)$, a function $f\colon X \to Y$ and an observable $q \in Obs(Y)$. Then:*

$$f_*(\omega) \models q \;=\; \omega \models q \circ f \tag{4.3}$$

The pushforward $f_*(\omega) \in \mathcal{D}(Y)$ involves state transformation with the deterministic channel $\langle f \rangle = unit \circ f$, and equals $\mathcal{D}(f)(\omega)$, see (2.36). This equation (4.3) is sometimes called 'the law of the unconscious statistician', see *e.g.* (Casella and Berger, 2002, §2.2)

Proof. Since:

$$\begin{aligned} f_*(\omega) \models q \;=\; \sum_{y \in Y} \mathcal{D}(f)(\omega)(y) \cdot p(y) \;&=\; \sum_{y \in Y} \left(\sum_{x \in f^{-1}(y)} \omega(x) \right) \cdot q(y) \\ &=\; \sum_{y \in Y} \sum_{x \in f^{-1}(y)} \omega(x) \cdot q(f(x)) \\ &=\; \sum_{x \in X} \omega(x) \cdot (q \circ f)(x) \;=\; \omega \models q \circ f. \end{aligned}$$

❑

Remark 4.1.6 What is the difference between a multiset and a factor, and between a distribution and a predicate? A multiset on X is a function $X \to \mathbb{R}_{\geq 0}$ with finite support, and thus a factor. In the same way a distribution on X is a function $X \to [0, 1]$ with finite support and probabilities adding up to one, and thus a predicate on X. Hence, there are inclusions:

$$\mathcal{M}(X) \subseteq Fact(X) = (\mathbb{R}_{\geq 0})^X \qquad \mathcal{D}(X) \subseteq Pred(X) = [0, 1]^X. \tag{4.4}$$

The first inclusion $\subseteq$ may actually be seen as an equality $=$ when the set X is a finite set. A predicate p on a finite set however, need not be a state, since its values $p(x)$ need not add up to one. Such a predicate can be normalised, if it is non-zero; then it is turned into a state.

We are however reluctant to identify states / distributions with certain predicates, since they belong to different universes and have quite different algebraic properties. For instance, the set of distributions is a convex set, closed under convex sums, whereas predicates are effect modules carrying a commutative monoid, see the next section. Keeping states and predicates apart is a matter of mathematicaly hygiene[1]. We have already seen state transformation c_* along a channel

[1] Moreover, in continuous probability there are no inclusions of distributions in predicates.

c; it preserves the convex structure. Later on in this chapter we shall also see predicate transformation c^* along a channel, in opposite direction; this operation preserves the effect module structure on predicates.

Predicates and factors can be obtained from distributions via affine maps. Recall from Exercise 2.1.18 that we write **Conv** for the category of convex sets and affine functions. The 'homset' of affine functions then gives isomorphisms:

$$\mathbf{Conv}\big(\mathcal{D}(X), [0,1]\big) \cong \mathit{Pred}(X) \qquad\qquad \mathbf{Conv}\big(\mathcal{D}(X), \mathbb{R}_{\geq 0}\big) \cong \mathit{Fact}(X). \tag{4.5}$$

These isomorphisms follow from the freeness of $\mathcal{D}(X)$, see Exercise 2.1.18. In Section 6.10 we shall see that, in the other direction, we can also obtain distributions from predicates, again by 'homming into $[0,1]$'.

Apart from these mathematical differences, states and predicates play entirely different roles and are understood in different ways: states play an onotologoical role and describe 'states of affairs'; predicates are epistemological in nature, and describe evidence (what we know).

If ever we do make use of the above inclusions (4.4), we shall make this usage explicit.

Exercises

4.1.1 Check that the average of the set $\boldsymbol{n+1} = \{0, 1, \ldots, n\}$, considered as random variable, is $\frac{n}{2}$.

4.1.2 Let $c\colon X \rightsquigarrow Y$ be channel with a distribution $\omega \in \mathcal{D}(X)$. Check that pushforward can be expressed via validity $\models$, namely as:

$$c_*(\omega)(y) \;=\; \omega \models c(-)(y) \qquad \text{where } c(-)(y)\colon X \to [0,1].$$

4.1.3 The birthday paradox tells that with at least 23 people in a room there is a more than 50% chance that at least to of them have the same birthday. This is called a 'paradox', because the number 23 looks surprisingly low.

Let us scale this down so that the problem becomes manageable. Suppose there are three people, each with their birthday in the same week.

1 Show that the probability that they all have different birthdays is $\frac{30}{49}$.
2 Conclude that the probability that at least two of the three have the same birthday is $\frac{19}{49}$.
3 Consider the set $\mathit{days} := \{1, 2, 3, 4, 5, 6, 7\}$. The set $\mathcal{N}[3](\mathit{days})$ of multisets of size three contains the possible combinations of the three birthdays. Define the sharp predicate $p\colon \mathcal{N}[3](\mathit{days}) \to \{0,1\}$ by $p(\varphi) = 1$ iff $\mathbf{(}\varphi\mathbf{)} \leq 3$. Check that p holds in those cases where at least two birthdays coincide — see also Exercise 1.7.12.
4 Show that the probability $\frac{19}{49}$ of at least two coinciding birthdays can also be obtained via validity, namely as:

$$\mathit{mn}[3](\mathit{unif}_{\mathit{days}}) \models p \;=\; \tfrac{19}{49}.$$

4.1.4 Consider the situation with one bus (per hour) in Example 4.1.4.

1 Compute the expected waiting time when the bus arrives at minute 60, that is, compute the validity $\mathit{unif}_H \otimes 1|60\rangle \models p$, for the waiting time factor p from Example 4.1.4.

2 Now assume that not only the passenger arrives uniformly random, but the bus too. Show that the resulting expected waiting time $unif_H \otimes unif_H \models p$ equals $\frac{3599}{360}$, which is almost 10 minutes.

Hint: Use Proposition 1.2.6 (1) and (2).

4.1.5 For a set X and a number $K \geq 1$, consider a sequence $\vec{x} = (x_1, \ldots, x_K) \in X^K$ and an observable $p \in Obs(X)$. Show that:

$$Flrn\Big(acc(\vec{x})\Big) \models p \;=\; \frac{1}{K} \cdot \sum_{1 \leq i \leq k} p(x_i).$$

4.1.6 Let a distribution $\omega \in \mathcal{D}(X)$ and a non-zero factor $p \colon X \to \mathbb{R}_{>0}$ be given. We write $\ln(p) \colon X \to \mathbb{R}$ for the observable given by $\ln(p)(x) = \ln(p(x))$. Show that Jensen's inequality 2.73 gives the following validity inequalities.

$$\ln(\omega \models p) \;\geq\; \omega \models \ln(p), \qquad \text{equivalently} \qquad \omega \models p \;\geq\; e^{\omega \models \ln(p)}.$$

The latter form is called log-approximation in Tull et al. (2023).

4.1.7 Find counterexamples demonstrating that, in general,

1 $(\omega_1 \models p) \leq (\omega_2 \models p) \;\not\Longrightarrow\; (\omega_1 \models \ln(p)) \leq (\omega_2 \models \ln(p))$;
2 $(\omega \models p_1) \leq (\omega \models p_2) \;\not\Longrightarrow\; (\omega \models \ln(p_1)) \leq (\omega \models \ln(p_2))$.

4.1.8 Let $\omega, \rho \in \mathcal{D}(X)$ and $c, d \colon X \rightarrowtail Y$.

1 Prove the following property — called the chain rule in (Murphy, 2023, §5.1).

$$D_{KL}\Big(\langle id, c\rangle_*(\omega),\, \langle id, d\rangle_*(\rho)\Big) \;=\; D_{KL}(\omega, \rho) \;+\; \omega \models D_{KL}\Big(c(-),\, d(-)\Big).$$

2 Prove a similar chain rule for entropy:

$$H\Big(\langle id, c\rangle_*(\omega)\Big) \;=\; H(\omega) \;+\; \Big(\omega \models H \circ c\Big).$$

The latter validity expression is often called conditional entropy.

4.1.9 Let $c \colon X \rightarrowtail Y$ be a channel with distributions $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$.

1 Prove for the differences of validities below:

$$\Big(\sigma \models D_{KL}(c(-), \tau)\Big) - \Big(\sigma \models D_{KL}(c(-), c_*(\sigma))\Big) \;=\; D_{KL}(c_*(\sigma), \tau).$$

2 Conclude that the subset:

$$\operatorname*{argmin}_{\tau \in \mathcal{D}(Y)} \sigma \models D_{KL}(c(-), \tau)$$

has the prediction $c_*(\sigma)$ as only element. This minimum validity $\sigma \models D_{KL}(c(-), c_*(\sigma))$ is the mutual information of the (joint) graph distribution $\langle id, c\rangle_*(\sigma)$, see Exercise 7.1.8.

3 Check that when the divergence arguments are swapped, there is an inequality of the form:

$$\sigma \models D_{KL}(\tau,\, c(-)) \;\geq\; D_{KL}(\tau,\, c_*(\sigma)).$$

Hint: The proof of Proposition 2.10.6 (1) may provide inspiration.

4.2 The mean of a distribution

The mean is defined for a distribution on the real numbers — or on a subset. It summarises the distribution in a very elementary way. Suppose that I do a study about a group of people and I have a distribution describing their ages. Suppose I tell you the mean age is 18. It tells you nothing about the distribution, but it does give you an idea. And this idea is quite different when I tell that the mean age is 65. The mean is sometimes called a 'summary statistic' since it summarises a distribution in a certain way.

This section studies the mean of a distribution. Technically it is defined as the validity of the identity observable. Although this is probably the most simple observable, it does lead to interesting results. For instance, it allows us to calculate the KL-divergence between two binomial or Poisson distributions.

Definition 4.2.1 Let X be a set (of numbers) with an inclusion function of the form $\mathit{incl} \colon X \hookrightarrow \mathbb{R}$.

1 The *mean* $\mathit{Mean}(\omega) \in \mathbb{R}$, also known expectation, of a distribution $\omega \in \mathcal{D}(X)$ is defined as the validity of incl, considered as an observable:

$$\mathit{Mean}(\omega) \coloneqq \omega \models \mathit{incl} = \sum_{x \in X} \omega(x) \cdot x.$$

When there is such an inclusion $\mathit{incl} \colon X \hookrightarrow \mathbb{R}$ we may as well simply consider ω as a distribution on the reals, that is, write $\omega \in \mathcal{D}(\mathbb{R})$, so that $\mathit{Mean}(\omega) = \omega \models \mathit{id}$, with identity function id as observable $\mathit{id} \colon \mathbb{R} \to \mathbb{R}$ on the reals $\mathbb{R}$.

2 A distribution $\omega \in \mathcal{D}(\mathbb{R})$ can be *centralised* via shifting to $\omega' \in \mathcal{D}(\mathbb{R})$, where:

$$\omega' = \mathit{shift}(-\mathit{Mean}(\omega), \omega) = \mathcal{D}\big((-) - \mathit{Mean}(\omega)\big)(\omega).$$

This uses the shift action from Example 2.1.4 (1) and Exercise 2.7.8. This ω' is called the *centralisation* of ω. It satisfies $\mathit{Mean}(\omega') = 0$, by construction. It is thus centered around 0.

We shall use the above definition of mean also for distributions $\omega \in \mathcal{D}_\infty(X)$, but this requires some care, see Remark 4.2.3 below.

The definition of mean may be used in more general situations when the set X is a convex set (or a cone or a vector space), equipped with scalar multiplication and addition. For instance, for $\omega \in \mathcal{D}(\mathbb{R}^N)$ one obtains, since scalar multiplication and sums are obtained component-wise:

$$\begin{aligned}
\mathit{Mean}(\omega) &\coloneqq \sum_{(x_1,\ldots,x_N) \in \mathit{supp}(\omega)} \omega(x_1,\ldots,x_N) \cdot \big(x_1,\ldots,x_N\big) \\
&= \sum_{(x_1,\ldots,x_N) \in \mathit{supp}(\omega)} \big(\omega(x_1,\ldots,x_N) \cdot x_1, \ldots, \omega(x_1,\ldots,x_N) \cdot x_N\big) \\
&= \left(\sum_{(x_1,\ldots,x_N) \in \mathit{supp}(\omega)} \omega(x_1,\ldots,x_N) \cdot x_1, \ldots, \sum_{(x_1,\ldots,x_N) \in \mathit{supp}(\omega)} \omega(x_1,\ldots,x_N) \cdot x_N\right) \\
&= \big(\omega \models \pi_1, \ldots, \omega \models \pi_N\big).
\end{aligned} \tag{4.6}$$

In the last line the projection functions $\pi_i \colon \mathbb{R}^N \to \mathbb{R}$ are used as observables.

Abstractly, the mean is a $\mathcal{D}$-algebra $\mathcal{D}(X) \to X$, for a convex set X, see Exercise 4.2.5.

Example 4.2.2 We look at three standard cases.

1 We write *pips* for the set $\{1,2,3,4,5,6\}$, considered as a subset of $\mathbb{R}$, via the map $pips \subseteq \mathbb{R}$. As distribution we use the (uniform) fair dice $dice = unif_{pips} = \frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$. We may write $dice \in \mathcal{D}(pips)$ or also $dice \in \mathcal{D}(\mathbb{R})$. The mean of this distribution is:

$$Mean(dice) = dice \models id = \sum_{i \in pips} \tfrac{1}{6} \cdot i = \tfrac{1}{6} + \tfrac{2}{6} + \tfrac{3}{6} + \tfrac{4}{6} + \tfrac{5}{6} + \tfrac{6}{6} = \tfrac{21}{6} = \tfrac{7}{2}.$$

This is the expected outcome when throwing a dice.

2 Recall the binomial distribution $bn[K](r)$ on the set $\{0,1,\ldots,K\}$ from Example 2.1.2 (2), for $r \in [0,1]$. There is an inclusion function $\{0,1,\ldots,K\} \hookrightarrow \mathbb{R}$ that allows us to compute the mean of the binomial distribution.

We can treat the binomial distribution as a special instance of the multinomial distribution, via the isomorphism $\{0,1,\ldots,K\} \cong \mathcal{N}[K](2)$, and then use Lemma 3.3.3. But one can also compute directly:

$$\begin{aligned} Mean\Big(bn[K](r)\Big) &= \sum_{0 \le i \le K} bn[K](r)(i) \cdot i = \sum_{1 \le i \le K} \binom{K}{i} \cdot r^i \cdot (1-r)^{K-i} \cdot i \\ &= \sum_{1 \le i \le K} \frac{K!}{(i-1)! \cdot (K-i)!} \cdot r^i \cdot (1-r)^{K-i} \\ &= K \cdot r \cdot \sum_{0 \le j \le K-1} \frac{(K-1)!}{j! \cdot ((K-1)-j)!} \cdot r^j \cdot (1-r)^{(K-1)-j} \\ &= K \cdot r \cdot \sum_{0 \le j \le K-1} bn[K-1](r)(j) = K \cdot r. \end{aligned}$$

In the two plots of binomial distributions $bn[10](\frac{1}{3})$ and $bn[10](\frac{3}{4})$ in Figure 2.2 (on page 87) one can see that the associated means $\frac{10}{3} = 3.333\ldots$ and $\frac{30}{4} = 7.5$ make sense.

Means for multivariate drawing will be considered in Section 4.6.

3 Recall from Example 2.1.8 (1) that the Poisson distribution $pois[\lambda] \in \mathcal{D}_\infty(\mathbb{N})$ has infinite support. Its rate λ is also called its mean since one has:

$$Mean\Big(pois[\lambda]\Big) = \lambda.$$

This follows from an easy argument, much like in the previous point:

$$\begin{aligned} Mean\Big(pois[\lambda]\Big) = Mean\left(\sum_{k \ge 0} e^{-\lambda} \cdot \frac{\lambda^k}{k!} |k\rangle\right) &= \sum_{k \ge 1} k \cdot e^{-\lambda} \cdot \frac{\lambda^k}{k!} \\ &= \sum_{k \ge 1} \lambda \cdot e^{-\lambda} \cdot \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \cdot \sum_{k \ge 0} e^{-\lambda} \cdot \frac{\lambda^k}{k!} = \lambda \cdot 1 = \lambda. \end{aligned}$$

Remark 4.2.3 In the above examples we have computed the mean of the infinite Poisson distribution. It gave a clear outcome, eventhough it involves an infinite sum. However, this sum need not always converge. We reproduce a well-known counterexample, using the 'Basel' sum formula (due to Euler) on the left below, giving rise to the distribution $\rho \in \mathcal{D}_\infty(\mathbb{N})$ on the right.

$$\sum_{k \ge 1} \frac{1}{k^2} = \frac{\pi^2}{6} \qquad\qquad \rho := \frac{6}{\pi^2} \cdot \sum_{k \ge 1} \frac{1}{k^2} |k\rangle.$$

The sum $\sum_k \rho(k) \cdot k = \frac{6}{\pi^2} \cdot \sum_k \frac{1}{k}$ is infinite since the sum of the harmonic series $\sum_k \frac{1}{k}$ is infinite. Hence the distribution $\rho \in \mathcal{D}_\infty(\mathbb{N})$ does not have a mean.

For the next result we recall from Proposition 2.7.2 that if M is a commutative monoid, so is the set $\mathcal{D}(M)$ of distributions on M, via convolution. This is used below where M is the additive monoid $\mathbb{R}$ of real numbers.

Lemma 4.2.4 *The mean function, for distributions on real numbers, is a homomorphism of (commutative) monoids of the form:*

$$(\mathcal{D}(\mathbb{R}), +, 0) \xrightarrow{\mathit{Mean}} (\mathbb{R}, +, 0).$$

It also commutes with scalar multiplication $s \cdot (-) \colon \mathbb{R} \to \mathbb{R}$, for $s \in \mathbb{R}$, in the sense that the following diagram commutes.

$$\begin{tikzcd} \mathcal{D}(\mathbb{R}) \arrow[r, "\mathit{Mean}"] \arrow[d, "s*(-) = \mathcal{D}(s\cdot(-))"'] & \mathbb{R} \arrow[d, "s\cdot(-)"] \\ \mathcal{D}(\mathbb{R}) \arrow[r, "\mathit{Mean}"] & \mathbb{R} \end{tikzcd}$$

Proof. Preservation of the zero element is easy, since:

$$\mathit{Mean}(1|0\rangle) \;=\; 1 \cdot 0 \;=\; 0.$$

Next, for $\omega, \rho \in \mathcal{D}(\mathbb{R})$,

$$\begin{aligned} \mathit{Mean}(\omega + \rho) \overset{(2.60)}{=} \mathcal{D}(+)(\omega \otimes \rho) \models \mathit{id} \;&=\; \sum_{z \in \mathbb{R}} \mathcal{D}(+)(\omega \otimes \rho)(z) \cdot z \\ &=\; \sum_{x,y \in \mathbb{R}} (\omega \otimes \rho)(x,y) \cdot (x+y) \\ &=\; \sum_{x,y \in \mathbb{R}} \omega(x) \cdot \rho(y) \cdot (x+y) \\ &=\; \left(\sum_{x,y \in \mathbb{R}} \omega(x) \cdot \rho(y) \cdot x\right) + \left(\sum_{x,y \in \mathbb{R}} \omega(x) \cdot \rho(y) \cdot y\right) \\ &=\; \left(\sum_{x \in \mathbb{R}} \omega(x) \cdot \left(\sum_{y \in \mathbb{R}} \rho(y)\right) \cdot x\right) + \left(\sum_{y \in \mathbb{R}} \left(\sum_{x \in \mathbb{R}} \omega(x)\right) \cdot \rho(y) \cdot y\right) \\ &=\; \left(\sum_{x \in \mathbb{R}} \omega(x) \cdot x\right) + \left(\sum_{y \in \mathbb{R}} \rho(y) \cdot y\right) \;=\; \mathit{Mean}(\omega) + \mathit{Mean}(\rho). \end{aligned}$$

Mean commutes with scalar multiplication since for $\omega \in \mathcal{D}(\mathbb{R})$ and $s \in \mathbb{R}$ one has:

$$\begin{aligned} \mathit{Mean}(s * \omega) \;=\; \mathcal{D}(s \cdot (-))(\omega) \models \mathit{id} \;&=\; \left(\sum_{x \in \mathbb{R}} \omega(x) \big| s \cdot x\rangle\right) \models \mathit{id} \\ &=\; \sum_{x \in \mathbb{R}} \omega(x) \cdot s \cdot x \;=\; s \cdot \sum_{x \in \mathbb{R}} \omega(x) \cdot x \;=\; s \cdot \mathit{Mean}(\omega). \end{aligned}$$ ❑

In Example 4.2.2 (2) we have explicitly calculated the mean of the binomial distribution. With the above result we can give an alternative proof, using that the binomial distribution is an iterated convolution of flip distributions, see Proposition 2.7.6. This new proof goes as follows.

$$\mathit{Mean}\Big(bn[K](r)\Big) \;=\; \mathit{Mean}\Big(K \cdot \mathit{flip}(r)\Big) \;=\; K \cdot \mathit{Mean}\Big(\mathit{flip}(r)\Big) \;=\; K \cdot r.$$

In Example 4.2.2 we have seen the means of the binomial and Poisson distributions. We now use these formulas to calculate the KL-divergence between such distributions. Moreover, we use them to obtain Pythagorean results — like in Theorem 3.4.4 for multinomial distributions. In both cases below the mean gives the best approximant parameter, in terms of divergence.

Theorem 4.2.5 *Fix a numbers $K \in \mathbb{N}$ and $r \in [0,1]$.*

1 For $s \in [0,1]$,

$$D_{KL}\Big(bn[K](r),\ bn[K](s)\Big) = K \cdot D_{KL}\Big(flip(r),\ flip(s)\Big).$$

2 For $\omega \in \mathcal{D}(\{0,1,\dots,K\})$ one has:

$$D_{KL}\Big(\omega,\ bn[K](r)\Big) = D_{KL}\Big(\omega,\ bn[K](^{Mean(\omega)}/_K)\Big) + D_{KL}\Big(bn[K](^{Mean(\omega)}/_K),\ bn[K](r)\Big).$$

3 As a result,

$$\frac{Mean(\omega)}{K} \in \operatorname*{argmin}_{r\in[0,1]} D_{KL}\Big(\omega, bn[K](r)\Big).$$

Proof. 1 We expand the definition (2.72) and use the mean formula from Example 4.2.2 (2) in:

$$\begin{aligned}
D_{KL}\Big(bn[K](r),\ bn[K](s)\Big) &= \sum_{0\le i\le K} bn[K](r)(i)\cdot \ln\left(\frac{bn[K](r)(i)}{bn[K](s)}\right)\\
&= \sum_{0\le i\le K} bn[K](r)(i)\cdot \ln\left(\frac{\binom{K}{i}\cdot r^i\cdot(1-r)^{K-i}}{\binom{K}{i}\cdot s^i\cdot(1-s)^{K-i}}\right)\\
&= \sum_{0\le i\le K} bn[K](r)(i)\cdot \left[i\cdot\ln\left(\tfrac{r}{s}\right) + (K-i)\cdot\ln\left(\tfrac{1-r}{1-s}\right)\right]\\
&= \sum_{0\le i\le K} bn[K](r)(i)\cdot i\ln\left(\tfrac{r}{s}\right) + \sum_{0\le i\le K} bn[K](r)(i)\cdot(K-i)\cdot\ln\left(\tfrac{1-r}{1-s}\right)\\
&= K\cdot r\cdot\ln\left(\tfrac{r}{s}\right) + (K-K\cdot r)\cdot\ln\left(\tfrac{1-r}{1-s}\right)\\
&= K\cdot\left[r\cdot\ln\left(\tfrac{r}{s}\right) + (1-r)\cdot\ln\left(\tfrac{1-r}{1-s}\right)\right] = K\cdot D_{KL}\Big(flip(r),\ flip(s)\Big).
\end{aligned}$$

2 We reason along the same lines and use this first point in:

$$\begin{aligned}
&D_{KL}\Big(\omega,\ bn[K](r)\Big) - D_{KL}\Big(\omega,\ bn[K](^{Mean(\omega)}/_K)\Big)\\
&= \sum_{0\le i\le K} \omega(i)\cdot\ln\left(\frac{bn[K](^{Mean(\omega)}/_K)(i)}{bn[K](r)(i)}\right)\\
&= \sum_{0\le i\le K} \omega(i)\cdot i\cdot\ln\left(\frac{^{Mean(\omega)}/_K}{r}\right) + \sum_{0\le i\le K} \omega(i)\cdot(K-i)\cdot\ln\left(\frac{1-{}^{Mean(\omega)}/_K}{1-r}\right)\\
&= K\cdot\left[{}^{Mean(\omega)}/_K\cdot\ln\left(\frac{^{Mean(\omega)}/_K}{r}\right) + (1-{}^{Mean(\omega)}/_K)\cdot\ln\left(\frac{1-{}^{Mean(\omega)}/_K}{1-r}\right)\right]\\
&= K\cdot D_{KL}\Big(flip(^{Mean(\omega)}/_K),\ flip(r)\Big)\\
&= D_{KL}\Big(bn[K](^{Mean(\omega)}/_K),\ bn[K](r)\Big).
\end{aligned}$$

3 Directly by the previous point. ❑

There is a similar result for Poisson distributions.

Theorem 4.2.6 *Let $\lambda \in \mathbb{R}_{\ge 0}$ be given.*

1 For $\lambda' \in \mathbb{R}_{\geq 0}$,

$$D_{KL}\big(pois[\lambda],\, pois[\lambda']\big) \;=\; (\lambda' - \lambda) + \lambda \cdot \ln\big(\tfrac{\lambda}{\lambda'}\big).$$

2 For an arbitrary infinite distribution $\omega \in \mathcal{D}_\infty(X)$,

$$D_{KL}\big(\omega,\, pois[\lambda]\big) \;=\; D_{KL}\big(\omega,\, pois[Mean(\omega)]\big) + D_{KL}\big(pois[Mean(\omega)],\, pois[\lambda]\big).$$

3 Thus:

$$Mean(\omega) \in \operatorname*{argmin}_{\lambda \in \mathbb{R}_{\geq 0}} D_{KL}\big(\omega,\, pois[\lambda]\big).$$

Proof. 1 We use the formula for the Poisson mean from Example 4.2.2 (3) in:

$$\begin{aligned}
D_{KL}\big(pois[\lambda],\, pois[\lambda']\big) &= \sum_{n \in \mathbb{N}} pois[\lambda](n) \cdot \ln\left(\frac{e^{-\lambda} \cdot \lambda^n/n!}{e^{-\lambda'} \cdot \lambda'^n/n!}\right) \\
&= \sum_{n \in \mathbb{N}} pois[\lambda](n) \cdot \ln\left(e^{\lambda' - \lambda} \cdot \frac{\lambda^n}{\lambda'^n}\right) \\
&= \sum_{n \in \mathbb{N}} pois[\lambda](n) \cdot (\lambda' - \lambda) \;+\; \sum_{n \in \mathbb{N}} pois[\lambda](n) \cdot n \cdot \ln\big(\tfrac{\lambda}{\lambda'}\big) \\
&= (\lambda' - \lambda) + \lambda \cdot \ln\big(\tfrac{\lambda}{\lambda'}\big).
\end{aligned}$$

2 By the previous point:

$$\begin{aligned}
D_{KL}\big(\omega,\, pois[\lambda]\big) - D_{KL}\big(\omega,\, pois[Mean(\omega)]\big) &= \sum_{n \in \mathbb{N}} \omega(n) \cdot \ln\left(\frac{e^{-Mean(\omega)} \cdot Mean(\omega)^n/n!}{e^{-\lambda} \cdot \lambda^n/n!}\right) \\
&= \sum_{n \in \mathbb{N}} \omega(n) \cdot \Big[(\lambda - Mean(\omega)) + n \cdot \ln\big(\tfrac{Mean(\omega)}{\lambda}\big)\Big] \\
&= (\lambda - Mean(\omega)) + Mean(\omega) \cdot \ln\big(\tfrac{Mean(\omega)}{\lambda}\big) \\
&= D_{KL}\big(pois[Mean(\omega)],\, pois[\lambda]\big).
\end{aligned}$$

3 This follows from the previous point. ❑

Formulas for the divergence between Poisson-iid and Poisson-multinomial distributions appear in Exercises 4.6.9 and 4.6.10.

We conclude with an illustration of a mean calculation in a product space, as in (4.6). It involves a 'bivariate' binomial distribution — which we will describe following Jacobs (2025d). We can throw K times two coins, each with their own bias, and record the numbers of heads, for both coins. This is simply a (tensor) product of the two binomial distributions, for both coins. Instead, we like to throw two entwined, dependent coins, in the form of a 'two-coin' joint distribution $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$, for $\mathbf{2} = \{0, 1\}$. We look at the outcomes after K throws, in each component separately. Recall from Proposition 2.7.6 that the binomial distribution can be described in convolution form via the addition function $sum \colon \mathbf{2}^K \to \{0, \ldots, K\}$ as:

$$bn[K](r) \;=\; K \cdot flip(r) \;=\; \mathcal{D}(sum)\big(flip(r)^K\big).$$

We can use this same formula for two-coins, but via a suitably adapted addition.

Definition 4.2.7 For a number $K \in \mathbb{N}$, define the 'marginal sums' function $sums \colon (\mathbf{2} \times \mathbf{2})^K \to \{0, \ldots, K\} \times \{0, \ldots, K\}$ as:

$$sums \;:=\; \langle sum \circ (\pi_1)^K, sum \circ (\pi_2)^K \rangle$$

More concretely, $sums((a_1, b_1), \ldots, (a_K, b_K)) = (a_1 + \cdots + a_K,\ b_1 + \cdots + b_K)$.

We then define the *bivariate binomial distribution* $bvbn[K](\gamma)$, for a two-coin $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$ as:

$$bvbn[K](\gamma) \coloneqq \mathcal{D}(sums)\big(\gamma^K\big) \in \mathcal{D}\big(\{0, 1, \ldots, K\} \times \{0, 1, \ldots, K\}\big).$$

It is easy to generalise this definition from 2-coins to N-coins $\gamma \in \mathcal{D}(\mathbf{2}^N)$, for $N > 2$, by using an N-ary addition function $sums\colon (\mathbf{2}^N)^K \to \{0, \ldots, K\}^N$.

We immediately introduce an alternative formulation of the bivariate binomial, via the multinomial distribution. This version is used in Jacobs (2025d) and may be a bit easier to compute.

Lemma 4.2.8 *There is a 'marginal heads' function heads function in a commuting diagram:*

$$\begin{tikzcd}
(\mathbf{2} \times \mathbf{2})^K \arrow[d, "acc"'] \arrow[rr, bend left, "sums"] & & \{0, 1, \ldots, K\} \times \{0, 1, \ldots, K\} \\
\mathcal{N}[K](\mathbf{2} \times \mathbf{2}) \arrow[rr, bend right, "heads"'] & & {}
\end{tikzcd} \tag{4.7}$$

It is defined on a multiset $\varphi \in \mathcal{N}[K](\mathbf{2} \times \mathbf{2})$ as:

$$\begin{aligned}
heads(\varphi) &\coloneqq (\,\varphi(1, 0) + \varphi(1, 1),\ \varphi(0, 1) + \varphi(1, 1)\,) \\
&= (\,\varphi[1, 0](1),\ \varphi[0, 1](1)\,) \\
&= (\,\mathcal{N}(\pi_1)(\varphi)(1),\ \mathcal{N}(\pi_2)(\varphi)(1)\,) \\
&= \langle ev_1 \circ \mathcal{N}(\pi_1),\ ev_1 \circ \mathcal{N}(\pi_2)\rangle(\varphi).
\end{aligned}$$

Then, for a two-coin distribution $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$, one gets the following alternative description of the bivariate binomial in terms of the multinomial:

$$bvbn[K](\gamma) = \mathcal{D}(heads)\big(mn[K](\gamma)\big). \tag{4.8}$$

Proof. We first show that the triangle (4.7) commutes, using the naturality of accumulation:

$$\begin{aligned}
heads \circ acc &= \langle ev_1 \circ \mathcal{N}(\pi_1),\ ev_1 \circ \mathcal{N}(\pi_2)\rangle \circ acc \\
&= \langle ev_1 \circ \mathcal{N}(\pi_1) \circ acc,\ ev_1 \circ \mathcal{N}(\pi_2) \circ acc\rangle \\
&= \langle ev_1 \circ acc \circ (\pi_1)^K,\ ev_1 \circ acc \circ (\pi_2)^K\rangle = \langle sum \circ (\pi_1)^K,\ sum \circ (\pi_2)^K\rangle = sums.
\end{aligned}$$

Equation (4.8) now follows easily:

$$\begin{aligned}
\mathcal{D}(heads)\big(mn[K](\gamma)\big) &\overset{(2.55)}{=} \mathcal{D}(heads)\big(\mathcal{D}(acc)(\gamma^K)\big) \\
&= \mathcal{D}(heads \circ acc)(\gamma^K) \overset{(4.7)}{=} \mathcal{D}(sums)(\gamma^K) = bvbn[K](\gamma)
\end{aligned}$$

❑

Consider the two-coin distribution $\gamma = \frac{3}{8}\big|0, 0\rangle + \frac{5}{12}\big|0, 1\rangle + \frac{1}{12}\big|1, 0\rangle + \frac{1}{8}\big|1, 1\rangle$. It is not hard to see that it is entwined, see Exercise 2.3.10. We first form the multinomial distribution:

$$\begin{aligned}
mn[2](\gamma) = \tfrac{9}{64}\Big|2|0, 0\rangle\Big\rangle &+ \tfrac{5}{16}\Big|1|0, 0\rangle + 1|0, 1\rangle\Big\rangle + \tfrac{25}{144}\Big|2|0, 1\rangle\Big\rangle + \tfrac{1}{16}\Big|1|0, 0\rangle + 1|1, 0\rangle\Big\rangle \\
&+ \tfrac{5}{72}\Big|1|0, 1\rangle + 1|1, 0\rangle\Big\rangle + \tfrac{1}{144}\Big|2|1, 0\rangle\Big\rangle + \tfrac{3}{32}\Big|1|0, 0\rangle + 1|1, 1\rangle\Big\rangle \\
&+ \tfrac{5}{48}\Big|1|0, 1\rangle + 1|1, 1\rangle\Big\rangle + \tfrac{1}{48}\Big|1|1, 0\rangle + 1|1, 1\rangle\Big\rangle + \tfrac{1}{64}\Big|2|1, 1\rangle\Big\rangle.
\end{aligned}$$

We now obtain the associated bivariate binomial distribution by applying the marginals heads

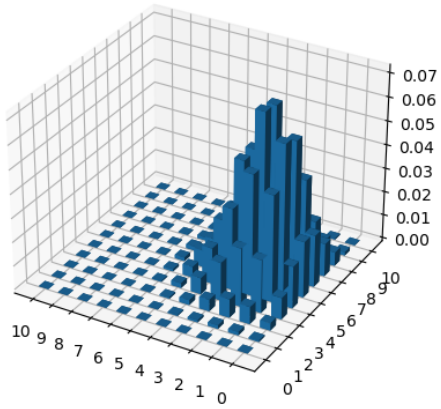


Figure 4.2 The bivariate binomial distribution obtained via 10 draws from the two-coin distribution $\gamma = \frac{3}{8}\big|0,0\rangle + \frac{5}{12}\big|0,1\rangle + \frac{1}{12}\big|1,0\rangle + \frac{1}{8}\big|1,1\rangle$.

function *heads* to the multisets inside the above big kets. By counting the total number of 1's on the left and on the right in these multisets one obtains:

$$\begin{aligned} bvbn[2](\gamma) \;=\; & \tfrac{9}{64}\big|0,0\rangle + \tfrac{5}{16}\big|0,1\rangle + \tfrac{25}{144}\big|0,2\rangle + \tfrac{1}{16}\big|1,0\rangle + \tfrac{47}{288}\big|1,1\rangle \\ & + \tfrac{5}{48}\big|1,2\rangle + \tfrac{1}{144}\big|2,0\rangle + \tfrac{1}{48}\big|2,1\rangle + \tfrac{1}{64}\big|2,2\rangle. \end{aligned}$$

The bivariate binomial for draw size $K = 10$, of the same γ, is plotted in Figure 4.2. One sees that its position is non-symmetric. Its mean is the pair $(\frac{5}{12}, \frac{13}{12})$. Below we shall see that this K times the mean of the two-coin distribution γ.

Next we collect some basic properties about the bivariate binomial distribution, in particular in relation to the ordinary binomial distribution.

Lemma 4.2.9 *Let a number $K \in \mathbb{N}$ and a two-coin distribution $\gamma \in \mathcal{D}(\mathbf{2}\times\mathbf{2})$ be given. We write $\gamma_1 := \gamma[1,0]$ and $\gamma_2 := \gamma[0,1]$ in $\mathcal{D}(\mathbf{2})$ for the two marginal coins. They are determined by the 'heads' probabilities $\gamma_1(1), \gamma_2(1) \in [0,1]$, via the isomorphism $\mathcal{D}(\mathbf{2}) \cong [0,1]$.*

1 *The marginals of the bivariate binomial distribution are the (ordinary) binomials of the marginals of the two-coin distribution:*

$$bvbn[K](\gamma)[1,0] \;=\; bn[K]\big(\gamma_1(1)\big) \qquad\qquad bvbn[K](\gamma)[0,1] \;=\; bn[K]\big(\gamma_2(1)\big).$$

2 *When γ is non-entwined, that is, when $\gamma = \gamma_1 \otimes \gamma_2$, then $bvbn[K](\gamma)$ is also non-entwined, with:*

$$bvbn[K]\big(\gamma_1 \otimes \gamma_2\big) \;=\; bn[K]\big(\gamma_1(1)\big) \otimes bn[K]\big(\gamma_2(1)\big).$$

This second result confirms that the notion of bivariate binomial is a well-defined extension of the ordinary binomial: if the two-coin distribution γ is a tensor product of two coins, the bivariate binomial is a tensor product of the two (marginal, ordinary) binomial distributions.

Proof. 1 We do the first equation

$$\begin{aligned}
bvbn[K](\gamma)[1,0] &= \mathcal{D}(\pi_1 \circ sums)\big(\gamma^K\big) \\
&= \mathcal{D}(sum \circ (\pi_1)^K)\big(\gamma^K\big) \\
&= \mathcal{D}(sum)\big(\mathcal{D}(\pi_1)(\gamma)^K\big) && \text{via Lemma 2.3.2 (3)} \\
&= K \cdot \gamma[1,0] && \text{reformulated as iterated convolution} \\
&= bn[K](\gamma[1,0](1)) && \text{see Proposition 2.7.6 (1).}
\end{aligned}$$

2 Now for $\gamma_1, \gamma_2 \in \mathcal{D}(\mathbf{2})$,

$$\begin{aligned}
bvbn[K]\big(\gamma_1 \otimes \gamma_2\big) &= \mathcal{D}(\langle sum \circ (\pi_1)^K, sum \circ (\pi_2)^K\rangle)\big((\gamma_1 \otimes \gamma_2)^K\big) \\
&= \mathcal{D}(sum \times sum)\big(\mathcal{D}(\langle (\pi_1)^K, (\pi_2)^K\rangle)\big((\gamma_1 \otimes \gamma_2)^K\big)\big) \\
&= \mathcal{D}(sum \times sum)\big(\gamma_1^K \otimes \gamma_2^K\big) \\
&= \mathcal{D}(sum)(\gamma_1^K) \otimes \mathcal{D}(sum)(\gamma_2^K) \\
&= \big(K \cdot \gamma_1\big) \otimes \big(K \cdot \gamma_2\big) \\
&= bn[K]\big(\gamma_1(1)\big) \otimes bn[K]\big(\gamma_2(1)\big).
\end{aligned}$$

❑

We can now describe the mean of the bivariate binomial analogously to the mean of the ordinary binomial, see Example 2.

Proposition 4.2.10 *The mean of the bivariate binomial is the can be expressed in terms of the draw size K and the two-coin distribution $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$, namely as:*

$$Mean\big(bvbn[K](\gamma)\big) = K \cdot Mean(\gamma).$$

Later on in Proposition 5.1.11 we shall see that the variance and covariance of a bivariate binomial and its underlying two-coin distribution also coincide.

Proof. We first elaborate what the mean of $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$ is, according to (4.6), in terms of γ's marginals $\gamma_1 = \gamma[1,0], \gamma_2 = \gamma[0,1] \in \mathcal{D}(\mathbf{2})$.

$$\begin{aligned}
Mean(\gamma) &\overset{(4.6)}{=} \left(\sum_{b_1,b_2 \in \mathbf{2}} \gamma(b_1,b_2) \cdot b_1, \sum_{b_1,b_2 \in \mathbf{2}} \gamma(b_1,b_2) \cdot b_2\right) \\
&= \left(\sum_{b_1 \in \mathbf{2}} \gamma_1(b_1) \cdot b_1, \sum_{b_2 \in \mathbf{2}} \gamma_2(b_2) \cdot b_2\right) = \big(\gamma_1(1), \gamma_2)(1)\big).
\end{aligned} \tag{$*$}$$

Now:

$$\begin{aligned}
Mean\Big(bvbn[K](\gamma)\Big) &\overset{(4.6)}{=} \left(\sum_{n_1,n_2\in\mathbb{N}} bvbn[K](\gamma)(n_1,n_2)\cdot n_1,\ \sum_{n_1,n_2\in\mathbb{N}} bvbn[K](\gamma)(n_1,n_2)\cdot n_2\right) \\
&= \left(\sum_{n_1\in\mathbb{N}} bvbn[K](\gamma)[1,0](n_1)\cdot n_1,\ \sum_{n_2\in\mathbb{N}} bvbn[K](\gamma)[1,0](n_2)\cdot n_2\right) \\
&= \left(\sum_{n_1\in\mathbb{N}} bn[K](\gamma_1(1))(n_1)\cdot n_1,\ \sum_{n_2\in\mathbb{N}} bn[K](\gamma_2(1))(n_2)\cdot n_2\right) \\
&= \Big(Mean(bn[K](\gamma_1(1)),\ Mean(bn[K](\gamma_2(1))\Big) \quad \text{by Lemma 4.2.9 (1)} \\
&= \Big(K\cdot\gamma_1(1),\ K\cdot\gamma_2(1)\Big) \\
&\overset{(*)}{=} K\cdot Mean(\gamma).
\end{aligned}$$

❑

This result about means can also be obtained via a suitable generalisation of Lemma 4.2.4, in the style of Exercise 2.7.11, using the convolution property of Exercise 4.2.6.

Exercises

4.2.1 Fix $K \in \mathbb{N}_{>0}$ and recall the negative binomial $nbn[K]\colon (0,1] \to \mathcal{D}_\infty(\mathbb{N})$ from Example 2.1.8 (3).

1 Prove that for $s \in (0,1]$ one has:

$$Mean\Big(nbn[K](s)\Big) = \frac{K}{s}.$$

2 Let $\psi \in \mathcal{N}(\mathbb{N})$ be a natural multiset of data with $K \leq \min(supp(\varphi))$, so that $supp(\varphi) \subseteq \{n \in \mathbb{N} \mid n \geq K\}$. Show, like in Examples 3.4.3 (2) and (3), that the probability s for which $nbn[K](s)$ best matches ψ is:

$$s = \frac{K}{Mean(Flrn(\psi))}.$$

4.2.2 Show that the three distributions (on the right-hand-side) in (2.58) all have $\frac{K}{N}$ as mean.

4.2.3 Recall Lemma 4.1.5 and check that for a distribution $\omega \in \mathcal{D}(X)$ and an observable $p\colon X \to \mathbb{R}$ one has:

$$Mean\Big(\mathcal{D}(p)(\omega)\Big) = \mathcal{D}(p)(\omega) \models id = p_*(\omega) \models id = \omega \models p.$$

4.2.4 Check yourself that the centralisation ω' in Definition 4.2.1 (2) satisfies $Mean(\omega') = 0$.

4.2.5 Consider the mean operation from Definition 4.2.1 as a function of the form $Mean\colon \mathcal{D}(\mathbb{R}) \to \mathbb{R}$. Show that it interacts in the following way with the unit and flatten maps for the distribution monad $\mathcal{D}$, see Section 2.1.

1 $Mean \circ unit = id$;

2 $Mean \circ flat = Mean \circ \mathcal{D}(Mean)$.

3 Let $p\colon X \to \mathbb{R}$ be an observable, giving a function $(-) \models p\colon \mathcal{D}(X) \to \mathbb{R}$, sending $\omega \in \mathcal{D}(X)$ to $\omega \models p$ in $\mathbb{R}$. Show that the following diagram commutes.

$$\begin{tikzcd} \mathcal{D}(\mathcal{D}(X)) \arrow[r, "\mathcal{D}((-)\models p)"] \arrow[d, "flat"'] & \mathcal{D}(\mathbb{R}) \arrow[d, "Mean"] \\ \mathcal{D}(X) \arrow[r, "(-)\models p"] & \mathbb{R} \end{tikzcd}$$

From the first two items we can conclude that *Mean* is an (Eilenberg-Moore) algebra of the distribution monad, see Subsection 1.12.4. This also works for the obvious restrictions $Mean\colon \mathcal{D}(\mathbb{R}_{\geq 0}) \to \mathbb{R}_{\geq 0}$ and $Mean\colon \mathcal{D}([0,1]) \to [0,1]$ of the mean function to non-negative reals and to the unit interval. The third item says that $(-) \models p\colon \mathcal{D}(X) \to \mathbb{R}$ is a homomorphism of algebras. It is the free extension of $p\colon X \to \mathbb{R}$.

4.2.6 Let $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$ be given.

1 Show that bivariate binomials are closed under convolution: for numbers $K, L \in \mathbb{N}$,

$$bvbn[K](\gamma) \;+\; bvbn[L](\gamma) \;=\; bvbn[K+L](\gamma).$$

Hint: Use the closure of multinomial distributions under convolution, from Exercise 3.3.10 and note that the commutative monoid involved is $\mathbb{N}^2$ with the component-wise sum $+ \cdot \mathbb{N}^2 \times \mathbb{N}^2 \to \mathbb{N}^2$, as in Exercise 2.7.11

2 Deduce that for $K \in \mathbb{N}$, the bivariate binomial is a K-fold convolution of the two-coin distribution γ:

$$bvbn[K](\gamma) \;=\; K \cdot \gamma \;=\; \gamma + \cdots + \gamma.$$

4.2.7 Fix a number $K \in \mathbb{N}$.

1 Show that for two numbers $n_1, n_2 \in \{0, \ldots, K\}$ one has:

$$\sum_{\varphi \in \mathcal{N}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi) = (n_1, n_2)} (\varphi) \;=\; \binom{K}{n_1} \cdot \binom{K}{n_2}.$$

2 Check that, as a result one can define a channel $heads[K]^\dagger\colon \{0, \ldots, K\}^2 \rightsquigarrow \mathcal{N}[K](\mathbf{2} \times \mathbf{2})$ by:

$$heads[K]^\dagger(n_1, n_2) \;:=\; \sum_{\varphi \in \mathcal{N}[K](\mathbf{2}\times\mathbf{2}),\, heads(\varphi) = (n_1, n_2)} \frac{(\varphi)}{\binom{K}{n_1} \cdot \binom{K}{n_2}} \,|\varphi\rangle.$$

4.2.8 Consider, for $N \geq 1$, the following N-ary generalisation of the marginal heads map from Lemma 4.2.8.

$$\begin{array}{ccc} \mathcal{N}[K](\mathbf{2}^N) & \xrightarrow{\quad heads \quad} & \{0, \ldots, K\}^N \\ \varphi & \longmapsto & (\,\mathcal{N}(\pi_1)(\varphi)(1), \ldots, \mathcal{N}(\pi_N)(\varphi)(1)\,) \end{array}$$

Prove that for for numbers $n_1, \ldots, n_N \in \{0, \ldots, K\}$ the following generalisation of Exercise 4.2.7.

$$\sum_{\varphi \in \mathcal{N}(\mathbf{2}^N),\, heads(\varphi) = \vec{n}} (\varphi) \;=\; \prod_{1 \leq i \leq N} \binom{K}{n_i}.$$

Hint: One can proceed via induction on $N \geq 1$ and use Lemma 3.1.6 (3) for the induction step.

4.3 Random variables

So far we have avoided discussing the concept of 'random variable'. It can be a confusing notion for people who start studying probability theory. One often encounters phrases like: let Y be a random variable with expectation $\mathbb{E}[Y]$. How should this be understood? For clarity, we define a random variable to be a *pair*, consisting of a distribution and an observable, with a shared underlying space.

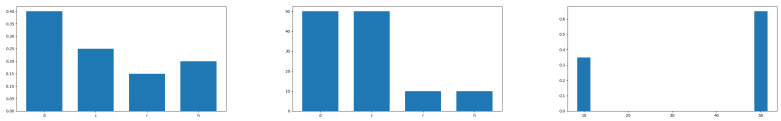
Definition 4.3.1 A *random variable* consists of a pair of a distribution $\omega \in \mathcal{D}(X)$ together with an observable $p\colon X \to \mathbb{R}$ on the same space X.

1 The *image distribution* associated with the random variable (ω, p) is the pushforward distribution $\mathcal{D}(p)(\omega) = \sum_{r\in\mathbb{R}} \big(\sum_{x\in p^{-1}(r)} \omega(x)\big)\big|r\rangle \in \mathcal{D}(\mathbb{R})$, obtained via the functoriality of $\mathcal{D}$. It is often written as $P[p = r]$, for $r \in \mathbb{R}$, see the pet illustration below.
2 The *expected value* $\mathbb{E}_\omega[p]$ of the random variable (ω, p), often simply written as $\mathbb{E}[p]$ with distribution ω implicit, is the validity $\omega \models p$. It is equal to the mean $\mathit{Mean}(\mathcal{D}(p)(\omega))$ of the image distribution, see Exercise 4.2.3.

The fact that a random variable is a *pair* of an observable *and* a distribution is not always so explicit in the literature. In Bailer-Jones (2017); Bishop (2006); Gelman et al. (2014) the notion of random variable is used without any definition. In (Rice, 2006, §2.4) the concept of a random variable is called "one of the fundamental ideas of probability theory". The entire second chapter of the book is devoted to this notion. It is described first as "essentially a random number" and then as "a function from Ω to the real numbers", and thus as an observable on the sample space Ω of outcomes, without making the distribution explicit. In (Casella and Berger, 2002, Def. 1.4.1) a random variable is only an observable, without any distribution. In (Murphy, 2023, 2.1.2) a random variable is also introduced as an observable: a function that maps an outcome to a number on the real line. But immediately after, the book talks about "the probability mass function (or pmf) for a random variable". This is where a distribution pops up. The third chapter of Blitzstein and Hwang (2019) is called "Random variable and their distributions", suggesting that distributions are part of the concept of a random variable. Random variables are described as "essential throughout the rest of the book, and throughout statistics". The connection between an observable and its distribution is left largely implicit in (Blitzstein and Hwang, 2019, Def. 3.1.1), stating: "Given an experiment with sample space S, a *random variable* (r.v.) is function from the sample space S to the real numbers $\mathbb{R}$." The experiment in this quote apparently refers to the distribution. In the literature one should thus be aware of what is called in (Kock, 1995, §16) the "notational confusion between a random variable and its distribution".

For emphasis, here we consider a random variable to be a *pair* (ω, p) of a distribution ω and an observable p, on the same space. We never leave the distribution ω implicit. If we drop anything, it is the observable p, but only when ω is a distribution on the reals $\mathbb{R}$, and p is the identity function $id : \mathbb{R} \to \mathbb{R}$, considered as observable on $\mathbb{R}$, the sample space of the distribution $\omega \in \mathcal{D}(\mathbb{R})$. Chapter 5 discusses variance, covariance and moments etc. for random variables.

Let's make things more concrete. Consider a space $\mathit{Pet} = \{d, c, r, h\}$ for d = dog, c = cat, r = rabbit and h = hamster. We look at the cost (*e.g.* of food) of a pet per month, via an observable $q\colon \mathit{Pet} \to \mathbb{R}$ given by $q(d) = q(c) = 50$ and $q(r) = q(h) = 10$. Let us assume that the distribution of pets in a certain neighbourhood is given by $\omega = \frac{2}{5}|D\rangle + \frac{1}{4}|C\rangle + \frac{3}{20}|R\rangle + \frac{1}{5}|H\rangle$. We can describe the situation via three plots:



The pet distribution ω is on the left and the cost observable q is in the middle. The plot on the right describes the image distribution $\mathcal{D}(q)(\omega) = \frac{7}{20}|10\rangle + \frac{13}{20}|50\rangle$, which is the distribution of pet costs. In traditional notation this image distribution $\mathcal{D}(q)(\omega)$ is described as:

$$P[q = 10] = \frac{7}{20} \qquad\qquad P[q = 50] = \frac{13}{20}.$$

Sometimes, a random variable is introduced as such an (image) distribution on the reals, with the underlying distribution ω and observable q left implicit. As this example illustrates, there may be much more structure around. We typically work directly with this structure — the distribution and observable — and not with the derived image distribution on the reals.

Alternatively, a random variable is sometimes described via a tilde $\sim$, as: $q \sim \omega$, like in phrases such as: "let $q \sim pois[\lambda]$ with ...". This means that $q\colon \mathbb{N} \to \mathbb{R}$ is an observable, on the underlying space $\mathbb{N}$ of the Poisson distribution $pois[\lambda] \in \mathcal{D}_\infty(\mathbb{N})$. The tilde notation thus properly describes a random variable as a pair of an observable with an associated distribution.

Example 4.3.2 We consider the expected value for the sum of two dices. In this situation we have an observable $S : pips \times pips \to \mathbb{R}$, on the sample space $pips = \{1, 2, 3, 4, 5, 6\}$, given by $S(i, j) = i + j$. It forms a random variable together with the product distribution $dice \otimes dice \in \mathcal{D}(pips \times pips)$. Recall, $dice = unif = \sum_{i \in pips} \frac{1}{6}|i\rangle$ is the uniform distribution $unif$ on $pips$. First, the distribution for the sum of the pips is the image distribution:

$$\begin{aligned} S_*(dice \otimes dice) &= \mathcal{D}(+)(dice \otimes dice) \\ &= \sum_{2 \leq n \leq 12} \left(\sum_{i,j,\, i+j=n} (dice \otimes dice)(i, j) \right) |n\rangle \\ &= \tfrac{1}{36}|2\rangle + \tfrac{1}{18}|3\rangle + \tfrac{1}{12}|4\rangle + \tfrac{1}{9}|5\rangle + \tfrac{5}{36}|6\rangle + \tfrac{1}{6}|7\rangle \\ &\qquad + \tfrac{5}{36}|8\rangle + \tfrac{1}{9}|9\rangle + \tfrac{1}{12}|10\rangle + \tfrac{1}{18}|11\rangle + \tfrac{1}{36}|12\rangle. \end{aligned} \tag{4.9}$$

This is actually the convolution sum of distributions $dice + dice$, as introduced in (2.60), see below.

The expected value of the random variable $(dice \otimes dice, S)$ is, according to Definition 4.3.1 (2),

$$\begin{aligned} Mean\big(\mathcal{D}(S)(dice \otimes dice)\big) &= dice \otimes dice \models S \\ &= \sum_{i,j \in pips} (dice \otimes dice)(i, j) \cdot S(i, j) \\ &= \sum_{i,j \in pips} \tfrac{1}{6} \cdot \tfrac{1}{6} \cdot (i + j) = \tfrac{1}{36} \cdot \sum_{i,j \in pips} i + j = \tfrac{252}{36} = 7. \end{aligned}$$

There is a more abstract way to look at this example, using Lemma 4.2.4. We have used $dice$ as a distribution on $pips = \{1, \ldots, 6\}$, but we can also see it as a distribution $dice \in \mathcal{D}(\mathbb{N})$ on the natural numbers. The sum of pips that we are interested in can then be described via a sum of distributions $dice + dice$, using the convolution $+$ from Proposition 2.7.2. Then, by Lemma 4.2.4, we also get:

$$Mean(dice + dice) = Mean(dice) + Mean(dice) = \tfrac{7}{2} + \tfrac{7}{2} = 7.$$

Sometimes one finds the expressions 'sum of random variables' in the literature. The next result makes this situation explicit.

Lemma 4.3.3 *Consider two sets X, Y with two random variables:*

$$\Big(\omega \in \mathcal{D}(X),\ X \xrightarrow{p} \mathbb{R}\Big) \qquad \Big(\rho \in \mathcal{D}(Y),\ Y \xrightarrow{q} \mathbb{R}\Big).$$

The sum of these two random variables is the new random variable with product space $X \times Y$ and with distribution-observable pair:

$$\Big(\omega \otimes \rho \in \mathcal{D}(X \times Y),\ X \times Y \xrightarrow{p \times q} \mathbb{R} \times \mathbb{R} \xrightarrow{+} \mathbb{R}\Big).$$

1 *The image of this sum of random variables is the convolution sum of the images of the two components.*
2 *The expected value of the sum is the sum of expected values.*

Proof. 1 Since:

$$\begin{aligned}
\mathcal{D}\Big(+ \circ (p \times q)\Big)(\omega \otimes \rho) &= \mathcal{D}(+)\Big(\mathcal{D}(p \times q)(\omega \otimes \rho)\Big) \\
&= \mathcal{D}(+)\Big(\mathcal{D}(p)(\omega) \otimes \mathcal{D}(q)(\rho)\Big) \quad \text{by Lemma 2.3.2 (3)} \\
&\overset{(2.60)}{=} \mathcal{D}(p)(\omega) + \mathcal{D}(q)(\rho).
\end{aligned}$$

2 One can compute this directly, but one can also use that the mean preserves sums, see Lemma 4.2.4:

$$\begin{aligned}
(\omega \otimes \rho) \models (+ \circ (p \times q)) &= \mathit{Mean}\Big(\mathcal{D}(+ \circ (p \times q))(\omega \otimes \rho)\Big) \\
&= \mathit{Mean}\Big(\mathcal{D}(p)(\omega) + \mathcal{D}(q)(\rho)\Big) \qquad \text{see the previous point} \\
&= \mathit{Mean}\Big(\mathcal{D}(p)(\omega)\Big) + \mathit{Mean}\Big(\mathcal{D}(q)(\rho)\Big) \\
&= (\omega \models p) + (\rho \models q).
\end{aligned}$$

❑

Exercises

4.3.1 In Example 4.3.2 we have seen that $\mathit{dice} \otimes \mathit{dice} \models S = 7$, for the observable $S : \mathit{pips} \times \mathit{pips} \to \mathbb{R}$ given by $S(x, y) = x + y$.

1 Now define $T : \mathit{pips}^3 \to \mathbb{R}$ by $T(x, y, z) = x+y+z$. Prove that $\mathit{dice} \otimes \mathit{dice} \otimes \mathit{dice} \models T = \frac{21}{2}$.
2 Can you generalise and show that summing on pips^n yields validity $\frac{7n}{2}$ in state dice^n?

4.3.2 Let $\Omega \in \mathcal{D}(\mathcal{D}(X))$ be a distribution of distributions, with an observable $p : X \to \mathbb{R}$. We turn p into a 'validity of p' observable $(-) \models p : \mathcal{D}(X) \to \mathbb{R}$ on $\mathcal{D}(X)$.

1 Show that:

$$\mathit{flat}(\Omega) \models p \ = \ \Omega \models \Big((-) \models p\Big).$$

2 Make the two random variables in this situation explicit.

4.3.3 The following result is a discrete formulation of what is called Fubini's theorem. For the more familiar continuous version, see Theorem **??**.

Let $\omega_1 \in \mathcal{D}(X_1)$ and $\omega_2 \in \mathcal{D}(X_2)$ be two distributions, with an observable (relation) $r : X_1 \times X_2 \to \mathbb{R}$.

1 Prove that:

$$\omega_1 \models r_1 \;=\; \omega_1 \otimes \omega_2 \models r \;=\; \omega_2 \models r_2,$$

where $r_1 \colon X_1 \to \mathbb{R}$ and $r_2 \colon X_2 \to \mathbb{R}$ are the observables defined by:

$$\begin{array}{rcl} r_1(x_1) & \coloneqq & \omega_2 \models r(x_1, -) \;=\; \sum_{x_2} \omega_2(x_2) \cdot r(x_1, x_2) \\ r_2(x_2) & \coloneqq & \omega_1 \models r(-, x_2) \;=\; \sum_{x_1} \omega_1(x_1) \cdot r(x_1, x_2). \end{array}$$

2 Make explicit what the (three) random variables in this situation are.

4.3.4 Let $(\omega \in \mathcal{D}(X), p \colon X \to \mathbb{R})$ be a random variable. For $s \in \mathbb{R}$ define the observable $s \cdot p$ by pointwise scalar multiplication: $(s \cdot p)(x) = s \cdot p(x)$. Prove, in the style of Lemma 4.3.3 that:

1 the image of $(\omega, s \cdot p)$ is $s * \mathcal{D}(p)(\omega)$, where $s * (-)$ is the scalar multiplication of distributions from (2.64);
2 The expected value of $(\omega, s \cdot p)$ is s times the expected value of (ω, p).
3 Generalise the situation to an arbitrary affine function $f \colon \mathbb{R} \to \mathbb{R}$.

4.4 The structure of observables

This section describes the algebraic structures that the various types of observables — in Table (4.1) — have, without going too much into mathematical details. We present the essentials and refer to the literature for details. We include the basic fact that mapping a set X to the set of observables on X is functorial, in a suitable sense. This will give rise to the notion of *weakening*. It plays the same (or dual) role for observables that marginalisation plays for states.

It turns out that our four types of observables are all (commutative) monoids, via multiplication / conjunction, but in different universes. The findings are summarised in the following table.

name	**type**	**monoid in**
observable	$X \to \mathbb{R}$	ordered vector spaces
factor	$X \to \mathbb{R}_{\geq 0}$	ordered cones
predicate	$X \to [0, 1]$	effect modules
sharp predicate	$X \to \{0, 1\}$	join semilattices

(4.10)

The least well-known structures in this table are effect modules. They will thus be described in greatest detail, in Subsection 4.4.3.

4.4.1 Observables

Observables are $\mathbb{R}$-valued functions on a set which, in the literature, are often written as capitals $X, Y, \ldots$. Here, these letters are typically used for sets and spaces. We shall use letters $p, q, \ldots$ for observables in general, and in particular for predicates. In some settings one allows observables $X \to \mathbb{R}^n$ to the n-dimensional space of real numbers. Whenever needed, we shall use such maps as n-ary tuples $\langle p_1, \ldots, p_n \rangle \colon X \to \mathbb{R}^n$ of observables $p_i \colon X \to \mathbb{R}$, see also Section 1.3.

Let us fix a set X, and consider the collection $Obs(X) = \mathbb{R}^X$ of observables on X. What structure does it have?

- Given two observables $p, q \in Obs(X)$, we can add them pointwise, giving $p + q \in Obs(X)$ via $(p + q)(x) = p(x) + q(x)$.
- Given an observable $p \in Obs(X)$ and a scalar $s \in \mathbb{R}$ we can form a 're-scaled' observable $s \cdot p \in Obs(X)$ via $(s \cdot p)(x) = s \cdot p(x)$. In this way we get $-p = (-1) \cdot p$ and $\mathbf{0} = 0 \cdot p$ for any observable p, where $\mathbf{0} = \mathbf{1}_\emptyset \in Obs(X)$ is the (always) zero observable from Definition 4.1.1 (2).
- For observables $p, q \in \mathbb{R}^X$ we have a partial order $p \leq q$ defined by: $p(x) \leq q(x)$ for all $x \in X$.

Together these operations of sum + and scalar multiplication $\cdot$ make $\mathbb{R}^X$ into a *vector space* over the real numbers, since + and $\cdot$ satisfy the appropriate axioms of vector spaces. Moreover, this is an *ordered* vector space by the third bullet.

One can restricts to *bounded* observables $p \colon X \to \mathbb{R}$ for which there is a bound $B \in \mathbb{R}_{>0}$ such that $-B \leq p(x) \leq B$ for all $x \in X$. The collection of such bounded observables forms an *order unit space* Alfsen and Shultz (2001); Jacobs et al. (2016); Nagel (1974); Paulsen and Tomforde (2009).

The set of observables $Obs(X) = \mathbb{R}^X$ also carries a commutative monoid structure $(\mathbf{1}, \&)$, where & is pointwise multiplication: $(p \,\&\, q)(x) = p(x) \cdot q(x)$. We prefer to write this operation as logical conjunction because that is what it is when restricted to predicates. Besides, having yet another operation that is written as dot $\cdot$ might be confusing. We will occasionally write p^n for $p \,\&\, \cdots \,\&\, p$ (n times).

4.4.2 Factors

We recall that a factor is a non-negative observables and that we write $Fact(X) = (\mathbb{R}_{\geq 0})^X$ for the set of factors on a set X. Updating a distribution makes sense for factors, and for predicates in particular, but not for observables, see Chapter 6. Here we concentrate on the mathematical structure of factors.

The set $Fact(X)$ looks like a vector space, except that there are no negatives. Using the order on observables introduced in the previous subsection, we can write:

$$Fact(X) = \{p \in Obs(X) \mid p \geq \mathbf{0}\}.$$

Factors can be added pointwise, with identity element $\mathbf{0} \in Fact(X)$, like random variables. But a factor $p \in Fact(X)$ cannot be re-scaled with an arbitrary real number, but only with a non-negative number $s \in \mathbb{R}_{\geq 0}$, giving $s \cdot p \in Fact(X)$. These structures are often called *cones*. The cone $Fact(X)$ is *positive*, since $p + q = \mathbf{0}$ implies $p = q = \mathbf{0}$. It is also *cancellative*: $p + r = q + r$ implies $p = q$.

The monoid $(\mathbf{1}, \&)$ on $Obs(X)$ restricts to $Fact(X)$, since $\mathbf{1} \geq \mathbf{0}$ and if $p, q \geq \mathbf{0}$ then also $p \,\&\, q \geq \mathbf{0}$.

4.4.3 Predicates

We first note that the set $Pred(X) = [0, 1]^X$ of predicates on a set X contains falsity $\mathbf{0}$ and truth $\mathbf{1}$, which are always 0 (resp. 1). There are some noteworthy differences between predicates on the one hand and observables and factors on the other hand.

- Predicates are not closed under addition, since the sum of two numbers in $[0, 1]$ may ly outside $[0, 1]$. Thus, addition of predicates is a *partial* operation, and is then written as $p \ovee q$. Thus: $p \ovee q$ is defined if $p(x) + q(x) \leq 1$ for all $x \in X$, and in that case $(p \ovee q)(x) = p(x) + q(x)$.

  This operation $\ovee$ is commutative and associative in a suitably partial sense. Moreover, it has $\mathbf{0}$ as identity element: $p \ovee \mathbf{0} = p = \mathbf{0} \ovee p$. This is structure $(Pred(X), \mathbf{0}, \ovee)$ is called a *partial commutative monoid*, see below for details.
- There is a 'negation' of predicates, written as $p^\perp$, and called *orthosupplement*. It is defined as $p^\perp = \mathbf{1} - p$, that is, as $p^\perp(x) = 1 - p(x)$. Then: $p \ovee p^\perp = \mathbf{1}$ and $p^{\perp\perp} = p$.
- Predicates are closed under scalar multiplication $s \cdot p$, but only if one restricts the scalar s to be in the unit interval $[0, 1]$. Such scalar multiplication interacts nicely with partial addition $\ovee$, in the sense that $s \cdot (p \ovee q) = (s \cdot p) \ovee (s \cdot q)$.

The combination of these items means that the set $Pred(X)$ carries the structure of an *effect module* Jacobs (2015), see also Dvurečenskij and Pulmannová (2000). These structures arose in mathematical physics Foulis and Bennett (1994) in order to axiomatise the structure of quantum predicates on Hilbert spaces and of effects in C^*-algebras and Von Neumann algebras Jacobs (2015).

The effect module $Pred(X)$ also carries a commutative monoid structure for conjunction, namely $(\mathbf{1}, \&)$. Indeed, when $p, q \in Pred(X)$, then also $p \,\&\, q \in Pred(X)$. We have $p \,\&\, \mathbf{0} = \mathbf{0}$ and $p \,\&\, (q_1 \ovee q_2) = (p \,\&\, q_1) \ovee (p \,\&\, q_2)$.

Since these effect structures are not so familiar, we include more formal descriptions.

Definition 4.4.1 1 A *partial commutative monoid* (PCM) consists of a set M with a zero element $0 \in M$ and a partial binary operation $\ovee \colon M \times M \to M$ satisfying the three requirements below. They involve the notation $x \perp y$ for: $x \ovee y$ is defined; in that case x, y are called orthogonal.

- Commutativity: $x \perp y$ implies $y \perp x$ and $x \ovee y = y \ovee x$;
- Associativity: $y \perp z$ and $x \perp (y \ovee z)$ implies $x \perp y$ and $(x \ovee y) \perp z$ and also $x \ovee (y \ovee z) = (x \ovee y) \ovee z$;
- Zero: $0 \perp x$ and $0 \ovee x = x$;

2 An *effect algebra* is a PCM $(E, 0, \ovee)$ with an orthosupplement. The latter is a total unary 'negation' operation $(-)^\perp \colon E \to E$ satisfying:

- $x^\perp \in E$ is the unique element in E with $x \ovee x^\perp = 1$, where $1 = 0^\perp$;
- $x \perp 1 \Rightarrow x = 0$.

  A homomorphism $E \to D$ of effect algebras is given by a function $f \colon E \to D$ between the underlying sets satisfying $f(1) = 1$, and if $x \perp x'$ in E then both $f(x) \perp f(x')$ in D and $f(x \ovee x') = f(x) \ovee f(x')$. Effect algebras and their homomorphisms form a category, denoted by **EA**.

3 An *effect module* is an effect algebra E with a scalar multiplication $s \cdot x$, for $s \in [0, 1]$ and $x \in E$, forming an action:

$$1 \cdot x = x \qquad (r \cdot s) \cdot x = r \cdot (s \cdot x),$$

and preserving sums (that exist) in both arguments:

$$\begin{array}{rcl} 0 \cdot x = 0 & \qquad & (r + s) \cdot x = r \cdot x \ovee s \cdot x \\ s \cdot 0 = 0 & & s \cdot (x \ovee y) = s \cdot x \ovee s \cdot y. \end{array}$$

We write **EMod** for the category of effect modules, where morphisms are maps of effect algebras that preserve scalar multiplication (*i.e.* are 'equivariant').

The following notion of 'test' comes from a quantum context and captures 'compatible' observations, see *e.g.* Davies and Lewis (1970); Ozawa (1984). It will be used occasionally later on, for instance in Exercise 6.1.6.

Definition 4.4.2 A *test* or, more explicitly, an *n-test* on a set X is an n-tuple of predicates $p_1, \dots, p_n \colon X \to [0,1]$ satisfying $p_1 \ovee \cdots \ovee p_n = \mathbf{1}$.

This notion of test can be formulated in an arbitrary effect algebra. Here we do it in $Pred(X)$ only.

Each predicate p forms a 2-test $p, p^\perp$. Exercise 4.5.3 asks to show that an n-test of predicates on X corresponds to a channel $X \twoheadrightarrow \boldsymbol{n}$. In particular, each predicate on X corresponds to a channel $X \twoheadrightarrow \mathbf{2}$, see Exercise 4.5.7. The probabilities $\omega(x_i) \in [0,1]$ of a distribution form a test on a singleton set 1.

The following easy observations give a normal form for predicates on a finite set.

Lemma 4.4.3 *Consider a finite set $X = \{x_1, \dots, x_n\}$.*

1 *The point predicates $\mathbf{1}_{x_1}, \dots, \mathbf{1}_{x_n}$ form a test on X.*
2 *A predicate $p \colon X \to [0,1]$ has a normal form $p = \bigovee_i r_i \cdot \mathbf{1}_{x_i}$, with scalars $r_i = p(x_i) \in [0,1]$.* ❑

On a set $X = \{a, b, c\}$ we can write a distribution and a predicate as:

$$\tfrac{1}{3}|a\rangle + \tfrac{1}{6}|b\rangle + \tfrac{1}{2}|c\rangle \qquad\qquad \tfrac{2}{3} \cdot \mathbf{1}_a \ovee \tfrac{5}{6} \cdot \mathbf{1}_b \ovee 1 \cdot \mathbf{1}_c.$$

Writing $\ovee$ looks a bit pedantic, so we often simply write $+$ instead. Recall that in a predicate the probabilities need not add up to one, see Remark 4.1.6.

A set of predicates $p_1, \dots, p_n$ on the same space, can be turned into a test via (pointwise) normalisation. Below we describe an alternative construction (also) called stick breaking. It has been described for numbers from the unit interval in Theorem 2.2.6, but it applies to (pointwise) predicates as well. It may be described even more generally, inside an arbitrary effect algebra with a commutative monoid structure for conjunction.

Lemma 4.4.4 *Let $p_1, \dots, p_n$ be arbitrary predicates, all on the same set, for $n \geq 1$. We can turn them via 'stick breaking' into an $n+1$-test $q_1, \dots, q_{n+1}$ via definitions:*

$$\begin{aligned} q_1 &\coloneqq p_1 \\ q_{i+1} &\coloneqq p_1^\perp \mathbin{\&} \cdots \mathbin{\&} p_i^\perp \mathbin{\&} p_{i+1} \qquad \textit{for } 0 < i < n \\ q_{n+1} &\coloneqq p_1^\perp \mathbin{\&} \cdots \mathbin{\&} p_n^\perp. \end{aligned}$$

❑

4.4.4 Sharp predicates

The structure of the set $SPred(X) = \{0,1\}^X = \mathbf{2}^X$ is most familiar to logicians: it is a Boolean algebra. The join $p \vee q$ of $p, q \in SPred(X)$ is the pointwise join $(p \vee q)(x) = p(x) \vee q(x)$, where the latter disjunction $\vee$ is the obvious one on $\{0,1\}$. One can see these disjunctions $(\mathbf{0}, \vee)$ as forming an additive structure with negation $(-)^\perp$. The partial sum $\ovee$ restricts from $Pred(X)$ to $SPred(X)$. For sharp predicates p, q the sum $p \ovee q$ is defined iff p, q are disjoint, as subsets. Conjunction $(\mathbf{1}, \&)$ forms an additional commutative monoid structure on $SPred(X)$.

Formally one can say that $SPred(X)$ also has scalar multiplication, with scalars from $\mathbf{2} = \{0, 1\}$, in such a way that $0 \cdot p = \mathbf{0}$ and $1 \cdot p = p$.

Exercise 4.4.11 below contains several alternative characterisations of sharpness for predicates.

The idea of a free structure as a 'minimal extension' has occurred earlier *e.g.* from Proposition 1.4.3 or 1.6.5. It also applies in the context of predicates, where fuzzy predicates are a free extension of sharp predicates. This is a fundamental insight that can be formulated in terms of effect algebras and modules, see (Jacobs et al., 2016, Prop. 33). Later on in Subsection **??** we will see that such a free extension also exists in continuous probability and forms the essence of integration (as validity).

Theorem 4.4.5 *1 For an arbitrary set X, the indicator function*

$$SPred(X) \cong \mathcal{P}(X) \xrightarrow{\mathbf{1}_{(-)}} Pred(X)$$

is a homomorphism of effect algebras.

2 Let X now be finite. This indicator map $\mathbf{1}_{(-)}$ *makes* $Pred(X)$ *into the* free effect module *on* $\mathcal{P}(X)$*, as described below: for each effect module E with a map of effect algebras* $f \colon \mathcal{P}(X) \to E$*, there is a unique map of effect modules* $\overline{f} \colon Pred(X) \to E$ *in:*

$$\begin{tikzcd} SPred(X) \cong \mathcal{P}(X) \arrow[r, "\mathbf{1}_{(-)}"] \arrow[dr, bend right, "f"'] & Pred(X) \arrow[d, dashed, "\overline{f},\ \textit{homomorphism}"] \\ & E \end{tikzcd}$$

Proof. 1 See Exercise 4.4.2.

2 We use that each predicate $p \in Pred(X)$ can be written in normal form as $p = \bigovee_{x \in X} p(x) \cdot \mathbf{1}_x$, see Lemma 4.4.3 (2). So we define $\overline{f}$ as:

$$\overline{f}(p) := \bigovee_{x \in X} p(x) \cdot f(\{x\}).$$

Obviously, $\overline{f}(\mathbf{0}) = \mathbf{0}$. Also, since f preserves $\ovee$ and top,

$$\overline{f}(\mathbf{1}) = \bigovee_{x \in X} 1 \cdot f(\{x\}) = f\left(\bigcup_{x \in X} \{x\}\right) = f(X) = 1.$$

If $p \perp q$, then $p(x) + q(x) \leq 1$ for each $x \in X$. Hence:

$$\begin{aligned} \overline{f}(p \ovee q) &= \bigovee_{x \in X} (p \ovee q)(x) \cdot f(\{x\}) = \bigovee_{x \in X} (p(x) + q(x)) \cdot f(\{x\}) \\ &= \bigovee_{x \in X} p(x) \cdot f(\{x\}) \ovee q(x) \cdot f(\{x\}) \\ &= \bigovee_{x \in X} p(x) \cdot f(\{x\}) \ovee \bigovee_{x \in X} q(x) \cdot f(\{x\}) = \overline{f}(p) \ovee \overline{f}(q). \end{aligned}$$

Scalar multiplication is also preserved:

$$\begin{aligned} \overline{f}(r \cdot p) &= \bigovee_{x \in X} (r \cdot p)(x) \cdot f(\{x\}) = \bigovee_{x \in X} r \cdot (p(x) \cdot f(\{x\})) \\ &= r \cdot \bigovee_{x \in X} p(x) \cdot f(\{x\}) = r \cdot \overline{f}(p). \end{aligned}$$

Finally, for uniqueness, let $g\colon Pred(X) \to E$ be a map of effect modules with $g(\mathbf{1}_U) = f(U)$ for each $U \in \mathcal{P}(X)$. Then $g = \overline{f}$ since:

$$\overline{f}(p) \;=\; \bigotimes_{x\in X} p(x)\cdot f(\{x\}) \;=\; \bigotimes_{x\in X} p(x)\cdot g(\mathbf{1}_x) \;=\; g\Big(\bigotimes_{x\in X} p(x)\cdot \mathbf{1}_x\Big) \;=\; g(p). \qquad \square$$

Now that we have seen observables, with factors and (sharp) predicates as special cases, we see that all these subsets of observables share the same multiplicative structure $(\mathbf{1}, \&)$ for conjunction, but that their additive structures and scalar multiplications differ. The additive structure is preserved under taking validity — as made explicit below — but not the multiplicative structure $(\mathbf{1}, \&)$.

Lemma 4.4.6 *Let $\omega \in \mathcal{D}(X)$ be a distribution on a set X. Operations on observables p, q on X satisfy:*

1. $\omega \models \mathbf{0} \;=\; 0$;
2. $\omega \models \mathbf{1} \;=\; 1$;
3. $\omega \models (p + q) \;=\; (\omega \models p) + (\omega \models q)$;
4. $\omega \models p^{\perp} \;=\; 1 - (\omega \models p)$;
5. $\omega \models (s \cdot p) \;=\; s \cdot (\omega \models p)$. $\square$

These preservation properties do not work for all logical operators. For instance, they fail for joins (disjunction $\vee$), see Exercise 4.4.12. This means that validity does not preserve the structure of MV-algebras. The abbreviation MV in MV-algebras stands for Many-Valued. These algebras have been introduced by Chang Chang (1958) for the study of infinite valued logic, called Łukasiewicz-logic, see Cignoli et al. (2000); Mundici (2011) for an overview. MV-algebras are special effect algebras, with a join $\vee$, that axiomatise the structure of the unit interval $[0, 1]$, and of fuzzy predicates in $[0, 1]^X$, in an alternative manner. We do not consider these MV-algebras here since their structure is not preserved by validity — or by predicate transformation, see Section 4.5. A join $\vee$ is an additional operation, like &, that is not preserved, like the linear (or effect module) structure, in the above result. In the terminology to be introduced in Section 4.5, joins are not preserved by observable transformation along an arbitrary channel, only along deterministic a channel (a function).

4.4.5 Parallel products and weakening

Earlier we have seen parallel products $\otimes$ of distributions and of channels. This product $\otimes$ can be defined for observables too, and is then often called parallel conjunction. The difference between parallel conjunction $\otimes$ and sequential conjunction & is that $\otimes$ acts on observables on *different* sets X, Y and yields an outcome on the product set $X \times Y$, whereas & works for observables on the same set Z, and produces a conjunction observable again on Z. These $\otimes$ and & are inter-definable, via transformation of observables, see Section 4.5 — in particular Exercise 4.5.9.

Definition 4.4.7 1 Let p be an observable on a set X, and q on Y. Then we define a new observable $p \otimes q$ on $X \times Y$ by:

$$(p \otimes q)(x, y) \;:=\; p(x) \cdot q(y).$$

2 Suppose we have an observable p on a set X and we like to use p on the product $X \times Y$. This can be done by taking $p \otimes \mathbf{1}$ instead, where $\mathbf{1}$ is the truth predicate. This $p \otimes \mathbf{1}$ is called a *weakening* of p. It satisfies $(p \otimes \mathbf{1})(x, y) = p(x)$ and is thus equal to $p \circ \pi_1 \colon X \times Y \to \mathbb{R}$.

More generally, consider a product $X_1 \times \cdots \times X_n$. For an observable p on the i-th set X_i, we weaken p to a predicate on the whole product $X_1 \times \cdots \times X_n$, namely:

$$\underbrace{\mathbf{1} \otimes \cdots \otimes \mathbf{1}}_{i-1 \text{ times}} \otimes p \otimes \underbrace{\mathbf{1} \otimes \cdots \otimes \mathbf{1}}_{n-i \text{ times}}.$$

Weakening is a structural operation in logic which makes it possible to use a predicate $p(x)$ depending on a single variable x in a larger context where one has for instance two variables x, y by ignoring the additional variable y. Weakening is usually not an explicit operation, except in settings like linear logic where one has to be careful about the use of resources. Here, we need weakening as an explicit operation in order to avoid type mismatches between observables and underlying sets.

Recall that marginalisation of states is an operation that moves a state to a smaller underlying set by projecting away. Weakening can be seen as a dual operation, moving an observable to a larger context. There is a close relationship between marginalisation and weakening via validity, using Lemma 4.1.5: for a state $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n)$ and an observable p on X_i we have:

$$\begin{aligned} \omega \models \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes p \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} &= \omega \models p \circ \pi_i \\ &= \mathcal{D}(\pi_i)(\omega) \models p \\ &= \omega[0, \ldots, 0, 1, 0, \ldots, 0] \models p, \end{aligned} \tag{4.11}$$

where the 1 in the latter marginalisation mask is at position i. Soon we shall see an alternative description of weakening in terms of predicate transformation. The above equation then appears as a special case of a more general result, namely of Proposition 4.5.3.

We mention some basic results about parallel products of observables. More such 'logical' results can be found in the exercises.

Lemma 4.4.8 *For states $\sigma \in \mathcal{D}(X)$, $\tau \in \mathcal{D}(Y)$ and observables $p \in Obs(X)$, $q \in Obs(Y)$ one has:*

$$(\sigma \otimes \tau \models p \otimes q) = (\sigma \models p) \cdot (\tau \models q).$$

Proof. Easy:

$$\begin{aligned} (\sigma \otimes \tau \models p \otimes q) &= \sum_{z \in X \times Y} (\sigma \otimes \tau)(z) \cdot (p \otimes q)(z) \\ &= \sum_{x \in X, y \in Y} (\sigma \otimes \tau)(x, y) \cdot (p \otimes q)(x, y) \\ &= \sum_{x \in X, y \in Y} \sigma(x) \cdot \tau(y) \cdot p(x) \cdot q(y) \\ &= \left(\sum_{x \in X} \sigma(x) \cdot p(x) \right) \cdot \left(\sum_{y \in Y} \tau(y) \cdot q(y) \right) = (\sigma \models p) \cdot (\tau \models q). \end{aligned}$$

❑

Lemma 4.4.9 *1 For observables $p_i, q_i \in Obs(X_i)$ one has:*

$$(p_1 \otimes \cdots \otimes p_n) \,\&\, (q_1 \otimes \cdots \otimes q_n) = (p_1 \,\&\, q_1) \otimes \cdots \otimes (p_n \,\&\, q_n).$$

2 *Parallel composition $p \otimes q$ of observables $p \in Obs(X)$ and $q \in Obs(Y)$ can be defined in terms of weakening and conjunction, namely as:*

$$p \otimes q \;=\; (p \otimes \mathbf{1}) \,\&\, (\mathbf{1} \otimes q) \;=\; (p \circ \pi_1) \,\&\, (q \circ \pi_2).$$

Proof. 1 For elements $x_i \in X_i$,

$$\begin{aligned}
&\big((p_1 \otimes \cdots \otimes p_n) \,\&\, (q_1 \otimes \cdots \otimes q_n)\big)(x_1, \ldots, x_n) \\
&= (p_1 \otimes \cdots \otimes p_n)(x_1, \ldots, x_n) \cdot (q_1 \otimes \cdots \otimes q_n)(x_1, \ldots, x_n) \\
&= (p_1(x_1) \cdot \ldots \cdot p_n(x_n)) \cdot (q_1(x_1) \cdot \ldots \cdot q_n(x_n)) \\
&= (p_1(x_1) \cdot q_1(x_1)) \cdot \ldots \cdot (p_n(x_n) \cdot q_n(x_n)) \\
&= (p_1 \,\&\, q_1)(x_1) \cdot \ldots \cdot (p_n \,\&\, q_n)(x_n) \;=\; \big((p_1 \,\&\, q_1) \otimes \cdots \otimes (p_n \,\&\, q_n)\big)(x_1, \ldots, x_n).
\end{aligned}$$

2 Directly from the previous item:

$$(p \otimes \mathbf{1}) \,\&\, (\mathbf{1} \otimes q) \;=\; (p \,\&\, \mathbf{1}) \otimes (\mathbf{1} \,\&\, q) \;=\; p \otimes q. \qquad \square$$

We conclude with some observations about proving equalities of states or predicates via validity.

Remark 4.4.10 For two states $\omega, \omega' \in \mathcal{D}(X_1 \times \cdots \times X_n)$ we have:

$$\begin{gathered}
\omega = \omega' \iff \omega \models p_1 \otimes \cdots \otimes p_n \;=\; \omega' \models p_1 \otimes \cdots \otimes p_n \\
\text{for all } p_i \in Pred(X_i).
\end{gathered} \tag{4.12}$$

The direction ($\Rightarrow$) is obvious. For ($\Leftarrow$) we use for an arbitrary $x_i \in X_i$ the corresponding tuple point predicate satisfies $\mathbf{1}_{(x_1,\ldots,x_n)} = \mathbf{1}_{x_1} \otimes \cdots \otimes \mathbf{1}_{x_n}$. Hence:

$$\begin{aligned}
\omega(x_1, \ldots, x_n) \;=\; \omega \models \mathbf{1}_{(x_1,\ldots,x_n)} \;&=\; \omega \models \mathbf{1}_{x_1} \otimes \cdots \otimes \mathbf{1}_{x_n} \\
&=\; \omega' \models \mathbf{1}_{x_1} \otimes \cdots \otimes \mathbf{1}_{x_n} \\
&=\; \omega' \models \mathbf{1}_{(x_1,\ldots,x_n)} \;=\; \omega'(x_1, \ldots, x_n).
\end{aligned}$$

Similarly, for predicates $p, p' \in Pred(X_1 \times \cdots \times X_n)$ one has:

$$\begin{gathered}
p = p' \iff \omega_1 \otimes \cdots \otimes \omega_n \models p \;=\; \omega_1 \otimes \cdots \otimes \omega_n \models p', \\
\text{for all } \omega_i \in \mathcal{D}(X_i).
\end{gathered} \tag{4.13}$$

Again, ($\Rightarrow$) is obvious and for ($\Leftarrow$) we use point states:

$$\begin{aligned}
p(x_1, \ldots, x_n) \;=\; 1|x_1, \ldots, x_n\rangle \models p \;&=\; 1|x_1\rangle \otimes \cdots \otimes 1|x_n\rangle \models p \\
&=\; 1|x_1\rangle \otimes \cdots \otimes 1|x_n\rangle \models p' \\
&=\; 1|x_1, \ldots, x_n\rangle \models p' \;=\; p'(x_1, \ldots, x_n).
\end{aligned}$$

Thus, proofs of equality of states, or of predicates, can be performed via validity $\models$. This may be convenient in situations where certain properties of validity can be used.

4.4.6 Functoriality

We like to conclude this section with some categorical observations. They are not immediately relevant for the sequel and may be skipped. We shall be using four (new) categories:

- **Vect**, with vector spaces (over the real numbers) as objects and linear maps as morphisms between them (preserving addition and scalar multiplication);
- **Cone**, with cones as objects and also with linear maps as morphisms, but this time preserving scalar multiplication with non-negative reals only;
- **EMod**, with effect modules as objects and homomorphisms of effect modules as maps between them (see Definition 4.4.1);
- **BA**, with Boolean algebras as objects and homomorphisms of Boolean algebras (preserving finite joins and negations, and then also finite meets).

Recall from Subsection 1.12.1 that we write $\mathbb{C}^{\mathrm{op}}$ for the opposite of category $\mathbb{C}$, with arrows reversed. This opposite is needed in the following result.

Proposition 4.4.11 *Taking particular observables on a set is functorial: there are functors:*

1. $Obs\colon \mathbf{Sets} \to \mathbf{Vect}^{\mathrm{op}}$;
2. $Fact\colon \mathbf{Sets} \to \mathbf{Cone}^{\mathrm{op}}$;
3. $Pred\colon \mathbf{Sets} \to \mathbf{EMod}^{\mathrm{op}}$;
4. $SPred\colon \mathbf{Sets} \to \mathbf{BA}^{\mathrm{op}}$.

On maps $f\colon X \to Y$ in **Sets** *these functors are all defined by the 'pre-compose with f' operation $q \mapsto q \circ f$. They thus reverse the direction of morphisms, which necessitates the use of opposite categories $(-)^{\mathrm{op}}$.*

The above functors $(-) \circ f$ all preserve the partial order on observables and also the commutative monoid structure $(\mathbf{1}, \&)$, since they are defined pointwise.

Proof. We consider the first instance of observables in some detail. The other cases are similar. For a set X we have have seen that $Obs(X) = \mathbb{R}^X$ is a vector space, and thus an object of the category **Vect**. Each function $f\colon X \to Y$ in **Sets** gives rise to a function $Obs(f)\colon Obs(Y) \to Obs(X)$ in the opposite direction. It maps an observable $q\colon Y \to \mathbb{R}$ on Y to the observable $q \circ f\colon X \to \mathbb{R}$ on X. It is not hard to see that this function $Obs(f) = (-) \circ f$ preserves the vector space structure. For instance, it preserves sums, since they are defined pointwise. We shall prove this in a precise, formal manner. First $Obs(f)(\mathbf{0})$ is the function that map $x \in X$ to:

$$Obs(f)(\mathbf{0})(x) = (\mathbf{0} \circ f)(x) = \mathbf{0}(f(x)) = 0.$$

Hence $Obs(f)(\mathbf{0})$ maps everything to 0 and is thus equal to the zero function itself: $Obs(f)(\mathbf{0}) = \mathbf{0}$. Next, addition $+$ is preserved since:

$$\begin{aligned} Obs(f)(p+q) &= (p+q) \circ f \\ &= [x \mapsto (p+q)(f(x))] \\ &= [x \mapsto p(f(x)) + q(f(x))] \\ &= [x \mapsto Obs(f)(p)(x) + Obs(f)(q)(x)] \\ &= Obs(f)(p) + Obs(f)(q). \end{aligned}$$

We leave preservation of scalar multiplication to the reader and conclude that $Obs(f)$ is a linear function, and thus a morphism $Obs(f)\colon Obs(Y) \to Obs(X)$ in **Vect**. Hence $Obs(f)$ is a morphism $Obs(X) \to Obs(Y)$ in the opposite category $\mathbf{Vect}^{\mathrm{op}}$. We still need to check that identity maps and

composition are preserved. We do the latter. For $f\colon X \to Y$ and $g\colon Y \to Z$ in **Sets** we have, for $r \in Obs(Z)$,

$$\begin{aligned} Obs(g \circ f)(r) \;=\; r \circ (g \circ f) &= (r \circ g) \circ f \\ &= Obs(g)(r) \circ f \\ &= Obs(f)(Obs(g)(r)) \\ &= (Obs(f) \circ Obs(g))(r) \;=\; (Obs(g) \circ^{\mathrm{op}} Obs(f))(r). \end{aligned}$$

This yields $Obs(g \circ f) = Obs(g) \circ^{\mathrm{op}} Obs(f)$, so that we get a functor of the form $Obs\colon \mathbf{Sets} \to \mathbf{Vect}^{\mathrm{op}}$. ❑

Notice that saying that we have a functor like $Pred\colon \mathbf{Sets} \to \mathbf{EMod}^{\mathrm{op}}$ contains remarkably much information, about the mathematical structure on objects $Pred(X)$, about preservation of this structure by maps $Pred(f) = (-) \circ f$, and about preservation of identity maps and composition by $Pred(-)$ on morphisms (see also Exercise 4.5.12). This makes the language of category theory both powerful and efficient.

Exercises

4.4.1 Consider the following question:

> An urn contains 10 balls of which 4 are red and 6 are blue. A second urn contains 16 red balls and an unknown number of blue balls. A single ball is drawn from each urn. The probability that both balls are the same color is $\frac{11}{25}$. Find the number of blue balls in the second urn.

1 Check that the givens can be expressed in terms of validity as:

$$\tfrac{11}{25} \;=\; Flrn(4|R\rangle{+}6|B\rangle) \otimes Flrn(16|R\rangle{+}x|B\rangle) \models (\mathbf{1}_R \otimes \mathbf{1}_R) \owedge (\mathbf{1}_B \otimes \mathbf{1}_B)$$

2 Prove, by solving the above equation, that there are 4 blue balls in the second urn.

4.4.2 1 Check that (sharp) indicator predicates $\mathbf{1}_E\colon X \to [0,1]$, for subsets $E \subseteq X$, satisfy:

- $\mathbf{1}_{E\cap D} = \mathbf{1}_E \mathbin{\&} \mathbf{1}_D$, and thus $\mathbf{1} \otimes \mathbf{1} = \mathbf{1}$, $\mathbf{1} \otimes \mathbf{0} = \mathbf{0} = \mathbf{0} \otimes \mathbf{1}$;
- $\mathbf{1}_{E\cup D} = \mathbf{1}_E \owedge \mathbf{1}_D$, if E, D are disjoint;
- $(\mathbf{1}_E)^{\perp} = \mathbf{1}_{\neg E}$, where $\neg E = X \setminus E = \{x \in X \mid x \notin E\}$ is the complement of E.

Formally, the function $\mathbf{1}_{(-)}\colon \mathcal{P}(X) \to Pred(X) = [0,1]^X$ is a homomorphism of effect algebras, see Theorem 4.4.5 (1).

2 Show that this indicator function is natural, in the sense that for $f\colon X \to Y$ the following diagram commutes.

$$\begin{tikzcd} \mathcal{P}(X) \arrow[r, "\mathbf{1}_{(-)}"] & Pred(X) \\ \mathcal{P}(Y) \arrow[u, "f^{-1}"] \arrow[r, "\mathbf{1}_{(-)}"] & Pred(Y) \arrow[u, "(-)\circ f"'] \end{tikzcd}$$

3 Now consider subsets $E \subseteq X$ and $D \subseteq Y$ of different sets X, Y together with their tensor product subset $E \otimes D \subseteq X \times Y$. Show that $\mathbf{1}_{E\otimes D} = \mathbf{1}_E \otimes \mathbf{1}_D$, with as special case $\mathbf{1}_{(x,y)} = \mathbf{1}_x \otimes \mathbf{1}_y$.

4.4.3 One may expect the following implication between inequalities of validities:

$$(\sigma \models p) \leq (\tau \models p) \implies (\sigma \models p \,\&\, q) \leq (\tau \models p \,\&\, q).$$

However, it *fails*. This exercise elaborates a counterexample. Take $X = \{a, b, c\}$ with states:

$$\sigma \;=\; \tfrac{19}{100}|a\rangle + \tfrac{47}{100}|b\rangle + \tfrac{17}{50}|c\rangle \qquad \tau \;=\; \tfrac{1}{5}|a\rangle + \tfrac{9}{20}|b\rangle + \tfrac{7}{20}|c\rangle$$

with predicates:

$$p \;=\; 1 \cdot \mathbf{1}_a + \tfrac{7}{10} \cdot \mathbf{1}_b + \tfrac{1}{2} \cdot \mathbf{1}_c \qquad q \;=\; \tfrac{1}{10} \cdot \mathbf{1}_a + \tfrac{1}{2} \cdot \mathbf{1}_b + \tfrac{1}{5} \cdot \mathbf{1}_c.$$

Now check consecutively that:

1 $\sigma \models p = \frac{689}{1000} < \frac{69}{100} = \tau \models p$.
2 $p \,\&\, q = \frac{1}{10} \cdot \mathbf{1}_a + \frac{7}{20} \cdot \mathbf{1}_b + \frac{1}{10} \cdot \mathbf{1}_c$.
3 $\sigma \models p \,\&\, q = \frac{87}{400} > \frac{85}{400} = \tau \models p \,\&\, q$.

Find a counterexample yourself in which the predicate q is sharp.

4.4.4 Consider a state $\sigma \in \mathcal{D}(X)$, a factor p on X, and a predicate q on X which is non-zero on the support of σ. Show that:

$$(\sigma \models \tfrac{p}{q}) \;\geq\; 1 \implies (\sigma \models p) \;\geq\; (\sigma \models q),$$

where $\frac{p}{q}(x) = \frac{p(x)}{q(x)}$.

4.4.5 Prove that the following items are equivalent, for a state $\omega \in \mathcal{D}(X)$ and event $E \subseteq X$.

1 $supp(\omega) \subseteq E$;
2 $\omega \models \mathbf{1}_E \;=\; 1$;
3 $\omega \models p \,\&\, \mathbf{1}_E \;=\; \omega \models p$ for each $p \in Obs(X)$.

4.4.6 Let $\omega \in \mathcal{D}(X)$ be an arbitrary distribution with two observables $p, q \in Obs(X)$.

1 Show that there is an inequality:

$$(\omega \models p \,\&\, q)^2 \;\leq\; (\omega \models p^2) \cdot (\omega \models q^2).$$

Hint: Recall $p^2 = p \,\&\, p$. Handle the case $(\omega \models p^2) = 0$ separately. Then define $r \in Obs(X)$ as $r := q - \frac{\omega \models p \& q}{\omega \models p^2} \cdot p$ and exploit that $(\omega \models r^2) \geq 0$ since $r^2 \geq 0$.

2 Deduce the inequality:

$$(\omega \models p)^2 \;\leq\; \omega \models p^2.$$

4.4.7 Let $\omega \in \mathcal{D}(X \times Y)$ have marginals $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$, making it a coupling of σ and τ. Show that for $p \in Obs(X)$ and $q \in Obs(Y)$,

$$\omega \models ((p \otimes \mathbf{1}) + (\mathbf{1} \otimes q)) \;=\; (\sigma \models p) + (\tau \models q).$$

4.4.8 Let p and q be two arbitrary predicates. Prove that the following use of the partial sum operation $\ovee$ for predicates is justified, that is, yields a well-defined new predicate.

$$(p \otimes q) \ovee (p^\perp \otimes q^\perp).$$

4.4.9 Let p_1, p_2, p_3 be predicates on the same set.

1 Write $p_1 \oplus p_2 \coloneqq (p_1^\perp \otimes p_2^\perp)^\perp$ and how that:

$$p_1 \oplus p_2 \;=\; (p_1 \otimes p_2) \ovee (p_1 \otimes p_2^\perp) \ovee (p_1^\perp \otimes p_2) \;=\; (p_1 \otimes \mathbf{1}) \ovee (p_1^\perp \otimes p_2).$$

This $\oplus$ may be called the 'noisy parallel disjunction', just like $(p^\perp \,\&\, q^\perp)^\perp$ is often called 'noisy-or'.

2 Show that the generalisation of the first item to ternary form gives:

$$\begin{aligned} p_1 \oplus p_2 \oplus p_3 \;&\coloneqq\; (p_1^\perp \otimes p_2^\perp \otimes p_3^\perp)^\perp \\ &= (p_1 \otimes p_2 \otimes p_3) \ovee (p_1 \otimes p_2 \otimes p_3^\perp) \ovee (p_1 \otimes p_2^\perp \otimes p_3) \ovee (p_1 \otimes p_2^\perp \otimes p_3^\perp) \\ &\qquad \ovee (p_1^\perp \otimes p_2 \otimes p_3) \ovee (p_1^\perp \otimes p_2 \otimes p_3^\perp) \ovee (p_1^\perp \otimes p_2^\perp \otimes p_3) \\ &= (p_1 \otimes \mathbf{1} \otimes \mathbf{1}) \ovee (p_1^\perp \otimes p_2 \otimes \mathbf{1}) \ovee (p_1^\perp \otimes p_2^\perp \otimes p_3) \end{aligned}$$

3 Generalise to n-ary form.

4 Check that $p \oplus \mathbf{0} = p \otimes \mathbf{1}$, so that $p \oplus \mathbf{0}$ is also weakening of p.

4.4.10 For predicates p, q on the same set, define *Reichenbach* implication $\supset$ as:

$$p \supset q \;\coloneqq\; p^\perp \ovee (p \,\&\, q).$$

This is a fuzzy implication for which $p \supset p \neq \mathbf{1}$, in general, since $p \,\&\, p \neq p$, unless p is sharp.

1 Check that:

$$p \supset q \;=\; (p \,\&\, q^\perp)^\perp$$

from which it easily follows that:

$$p \supset \mathbf{0} \;=\; p^\perp \qquad \mathbf{1} \supset q \;=\; q \qquad p \supset \mathbf{1} \;=\; \mathbf{1} \qquad \mathbf{0} \supset q \;=\; \mathbf{1}.$$

2 Check also that:

$$p^\perp \;\leq\; p \supset q \qquad q \;\leq\; p \supset q.$$

3 Show that:

$$p_1 \supset (p_2 \supset q) \;=\; (p_1 \,\&\, p_2) \supset q.$$

4 For subsets E, D of the same set,

$$\mathbf{1}_E \supset \mathbf{1}_D \;=\; \mathbf{1}_{\neg(E \cap \neg D)} \;=\; \mathbf{1}_{\neg E \cup D}.$$

The subset $\neg(E \cap \neg D) \;=\; \neg E \cup D$ is the standard interpretation of 'E implies D' in Boolean logic (of subsets).

4.4.11 Let p be a predicate on a set X. Prove that the following statements are equivalent.

1 p is sharp;

2 $p \,\&\, p = p$;

3 $p \,\&\, p^\perp = \mathbf{0}$;

4 $q \leq p$ and $q \leq p^\perp$ implies $q = \mathbf{0}$, for each $q \in \mathit{Pred}(X)$.

4.4.12 For two predicates $p, q \colon X \to [0,1]$ write $p \vee q$ for the predicate given by the pointwise join (maximum). In general, such joins of predicates $\vee$ are not preserved by validity. Take $X = \{a, b\}$ and find a distribution $\omega \in \mathcal{D}(X)$ with two predicates p, q on X such that:

$$\omega \models (p \vee q) \;\neq\; (\omega \models p) \;\vee\; (\omega \models q).$$

Similarly, find ω, p, q such that:

$$\omega \models (p \,\&\, q) \;\neq\; (\omega \models p) \cdot (\omega \models q).$$

(There are joint counter examples for both properties.)

4.4.13 For a random variable (ω, p), show that the validity of the observable $p - (\omega \models p) \cdot \mathbf{1}$ is zero, *i.e.*,

$$\omega \models (p - (\omega \models p) \cdot \mathbf{1}) \;=\; 0.$$

Observables of this form are used later on to define (co)variance in Section 5.1.

4.4.14 For a multiset $\varphi \in \mathcal{M}(X)$ on a set X and a factor $p \in \mathit{Fact}(X)$ on the same set, define the multiset $\varphi \bullet p \in \mathcal{M}(X)$ as $(\varphi \bullet p)(x) = \varphi(x) \cdot p(x)$. Show that this gives a monoid action:

$$\mathcal{M}(X) \times \mathit{Fact}(X) \xrightarrow{\bullet} \mathcal{M}(X)$$

with respect to the multiplicative monoid structure $(\mathbf{1}, \&)$ on factors.

4.4.15 Let E be an arbitrary effect algebra. Prove, from the axioms of an effect algebra, that for elements $e, e', d, d', f \in E$ the following properties hold (see Jacobs (2015) for details).

1. Orthosupplement is an involution: $e^{\perp\perp} = e$;
2. Cancellation holds of the form: $e \ovee d = e \ovee d'$ implies $d = d'$;
3. Zero-sum freeness holds: $e \ovee d = 0$ implies $e = d = 0$;
4. Define an order $\leq$ on E via: $e \leq d$ iff $e \ovee f = d$ for some $f \in E$. This is a partial order with 1 as top and 0 as bottom element;
5. $e \leq d$ implies $d^\perp \leq e^\perp$;
6. $e \ovee d$ is defined iff $e \perp d$ iff $e \leq d^\perp$ iff $d \leq e^\perp$;
7. $e \leq d$ and $d \perp f$ implies $e \perp f$ and $e \ovee f \leq d \ovee f$;
8. if $e \leq e'$ and $d \leq d'$ and $e' \ovee d'$ is defined, then also $e \ovee d$ is defined.

4.4.16 In an effect algebra E, as above, define for elements $e, d \in E$,

$$\begin{array}{rcll} e \owedge d & := & (e^\perp \ovee d^\perp)^\perp & \text{if } e^\perp \perp d^\perp \\ e \ominus d & := & (e^\perp \ovee d)^\perp \;=\; e \owedge d^\perp & \text{if } e \geq d. \end{array}$$

Show that:

1. $(E, \owedge, 1)$ is a partial commutative monoid;
2. $e \leq d$ iff $e = d \owedge f$ for some f;
3. $e \ominus 0 = e$ and $1 \ominus e = e^\perp$ and $e \ominus e = 0$;
4. $e \ovee d = f$ iff $e = f \ominus d$; in particular, $(f \ominus d) \ovee d = f$;
5. $e \ovee d \leq f$ iff $e \leq f \ominus d$;
6. $e > d$ iff $e \ominus d > 0$;
7. $f \ominus e = f \ominus d$ implies $e = d$;
8. $e \leq d$ implies $d \ominus e \leq d$ and $d \ominus (d \ominus e) = e$;

9 the function $e \ovee (-)$ preserves all joins $\bigvee_i d_i$ that exist in E: if $e \perp d_i$ for each i, then $e \perp \bigvee_i d_i$ and $e \ovee (\bigvee_i d_i) = \bigvee_i (e \ovee d_i)$;

10 Similarly, $e \owedge (-)$ preserves meets.

11 A homomorphism $f\colon E \to D$ of effect algebras E, D preserves orthosupplement: $f(e^\perp) = f(e)^\perp$, and thus also $f(0) = 0$, $f(e \owedge d) = f(e) \owedge f(d)$ and $f(e \ominus d) = f(e) \ominus f(d)$.

4.4.17 Let E be an effect module. The aim of this exercise is to show that subconvex sums exist in E, see (Pulmannová and Gudder, 1998, Lemma 2.1). This means that for arbitrary elements $e_1, \dots, e_n \in E$ and scalars $r_1, \dots, r_n \in [0, 1]$ with $\sum_i r_i \leq 1$ the sum $r_1 \cdot e_1 \ovee \cdots \ovee r_n \cdot e_n$ exists in E

1 Use induction on $n \geq 1$, and check the base step.

2 For the induction step let $e_1, \dots, e_{n+1}$ and $r_1, \dots, r_{n+1} \in [0, 1]$ with $\sum_{i \leq n+1} r_i \leq 1$ be given. By induction hypothesis the sum $\bigovee_{i \leq n} r_i \cdot e_i$ exists. Use Exercise 4.4.15 to check that the following chain of inequalities holds and suffices to prove that the sum $\bigovee_{i \leq n+1} r_i \cdot e_i$ exists too.

$$r_{n+1} \cdot e_{n+1} \leq r_{n+1} \cdot 1 \leq \Big((\textstyle\sum_{i \leq n} r_i) \cdot 1 \Big)^\perp = \Big(\bigovee_{i \leq n} r_i \cdot 1 \Big)^\perp \leq \Big(\bigovee_{i \leq n} r_i \cdot e_i \Big)^\perp.$$

4.4.18 Consider validity as a function $V\colon Pred(X) \to Pred(\mathcal{D}(X))$, given by $V(p)(\omega) := \omega \models p$. Show that V is a map of effect modules.

4.5 Transformation of observables

One of the basic operations that we have seen so far is state transformation. It is used to transform a state / distribution ω on the domain X of a channel $c\colon X \to\!\!\!\!\circ\; Y$ into a state $c_*(\omega)$ on the codomain Y of the channel. This section introduces transformation of observables c^*. It works in the opposite direction of the channel: an observable q on the codomain Y is transformed into an observable $c^*(q)$ on the domain X. Thus, state transformation works forwardly, in the direction of the channel, whereas observable transformation works backwardly, against the direction of the channel. This operation c^* is often applied only to predicates — and is then called predicate transformation — but here we apply it more generally to observables.

This section introduces observable transformation and lists its key mathematical properties. In the Chapter 6 it will be used for probabilistic reasoning, especially in combination with updating.

Definition 4.5.1 Let $c\colon X \to\!\!\!\!\circ\; Y$ be a channel. An observable $q \in Obs(Y)$ is transformed into $c^*(q) \in Obs(X)$ via the definition:

$$c^*(q)(x) \;:=\; c(x) \models q \;=\; \sum_{y \in Y} c(x)(y) \cdot q(y). \tag{4.14}$$

When q is a point predicate $\mathbf{1}_y$, for an element $y \in Y$, we get:

$$c^*(\mathbf{1}_y)(x) \;=\; c(x)(y) \qquad \text{so that} \qquad c^*(\mathbf{1}_y) \;=\; c(-)(y) : X \to [0, 1].$$

The resulting function $Y \to Pred(X)$, given by $y \mapsto c^*(\mathbf{1}_y)$, is sometimes called the *likelihood* function.

There is a whole series of basic facts about observable transformation.

Lemma 4.5.2 *1 The operation $c^*\colon Obs(Y) \to Obs(X)$ of transforming observables along a channel $c\colon X \rightsquigarrow Y$ restricts, first to factors $c^*\colon Fact(Y) \to Fact(X)$, and then to fuzzy predicates $c^*\colon Pred(Y) \to Pred(X)$, but not to sharp predicates.*

2 Observable transformation c^* is linear: it preserves sums $(\mathbf{0}, +)$ of observables and scalar multiplication of observables.

3 Observable transformation preserves truth $\mathbf{1}$, but not conjunction &.

4 Observable transformation preserves the (pointwise) order on observables:

$$q_1 \leq q_2 \implies c^*(q_1) \leq c^*(q_2).$$

5 Transformation along the unit channel is the identity:

$$unit^*(q) = q.$$

6 Transformation along a composite channel is successive transformation:

$$(d \circ c)^*(q) = c^*(d^*(q)).$$

7 Transformation along a tuple transforms parallel conjunction $\otimes$ into sequential conjunction &, as in:

$$\langle c, d\rangle^*(p \otimes q) = c^*(p) \mathbin{\&} d^*(q).$$

8 Transformation along parallel channels preserves parallel conjunctions:

$$(e \otimes f)^*(p \otimes q) = e^*(p) \otimes f^*(q).$$

9 Transformation along a trivial, deterministic channel f is pre-composition:

$$f^*(q) = q \circ f.$$

Proof. 1 If $q \in Fact(Y)$ then $q(y) \geq 0$ for all y. But then also $c^*(q)(x) = \sum_y c(x)(y) \cdot q(y) \geq 0$, so that $c^*(q) \in Fact(X)$. If in addition $q \in Pred(Y)$, so that $q(y) \leq 1$ for all y, then also $c^*(q)(x) = \sum_y c(x)(y) \cdot q(y) \leq \sum_y c(x)(y) = 1$, since $c(x) \in \mathcal{D}(Y)$, so that $c^*(q) \in Pred(X)$.

The fact that a transformation $c^*(p)$ of a sharp predicate p need not be sharp is demonstrated in Exercise 4.5.2.

2 Easy.

3 $c^*(\mathbf{1})(x) = \sum_y c(x)(y) \cdot \mathbf{1}(y) = \sum_y c(x)(y) \cdot 1 = 1$. The fact that & is not preserved follows from Exercise 4.5.1.

4 Easy.

5 Recall that $unit(x) = 1|x\rangle$, so that $unit^*(q)(x) = \sum_y unit(x)(y) \cdot q(y) = q(x)$.

6 For $c\colon X \rightsquigarrow Y$ and $d\colon Y \rightsquigarrow Z$ and $q \in Obs(Z)$ we have:

$$\begin{aligned}(d \circ c)^*(q)(x) &= \sum_{z \in Z} (d \circ c)(x)(z) \cdot q(z) = \sum_{z \in Z} \left(\sum_{y \in Y} c(x)(y) \cdot d(y)(z) \right) \cdot q(z) \\ &= \sum_{y \in Y} c(x)(y) \cdot \left(\sum_{z \in Z} d(y)(z) \cdot q(z) \right) \\ &= \sum_{y \in Y} c(x)(y) \cdot d^*(q)(y) = c^*(d^*(q))(x).\end{aligned}$$

7 Let $c\colon X \rightsquigarrow Y$, $d\colon X \rightsquigarrow Z$, with $p \in Pred(Y)$ and $q \in Pred(Z)$. Then for $x \in X$,

$$\begin{aligned}
\langle c, d\rangle^*(p \otimes q)(x) &= \sum_{y\in Y,\, z\in Z} \langle c, d\rangle(x)(y,z) \cdot (p \otimes q)(y,z) \\
&= \sum_{y\in Y,\, z\in Z} c(x)(y) \cdot d(x)(z) \cdot p(y) \cdot q(z) \\
&= \Big(\sum_{y\in Y} c(x)(y) \cdot p(y)\Big) \cdot \Big(\sum_{z\in Z} d(x)(z) \cdot q(z)\Big) \\
&= c^*(p)(x) \cdot d^*(q)(x) \;=\; \big(c^*(p) \mathbin{\&} d^*(q)\big)(x).
\end{aligned}$$

8 Similarly,

$$\begin{aligned}
(e \otimes f)^*(p \otimes q)(x,y) &= \sum_{u,v} (e \otimes f)(x,y)(u,v) \cdot (p \otimes q)(u,v) \\
&= \sum_{u,v} e(x)(u) \cdot f(y)(v) \cdot p(u) \cdot q(v) \\
&= \Big(\sum_{u} e(x)(u) \cdot p(u)\Big) \cdot \Big(\sum_{v} f(y)(v) \cdot q(v)\Big) \\
&= e^*(p)(x) \cdot f^*(q)(y) \;=\; \big(e^*(p)(x) \otimes f^*(q)\big)(x,y).
\end{aligned}$$

9 For a function $f\colon X \to Y$, recall that we write $\langle f\rangle := unit \circ f\colon X \rightsquigarrow Y$ if we wish to make explicit that we use the function f as a 'deterministic' channel. Then:

$$\langle f\rangle^*(q)(x) \;=\; \sum_{y\in Y} unit(f(x))(y) \cdot q(y) \;=\; q(f(x)) \;=\; (q \circ f)(x). \qquad \square$$

There is the following fundamental relationship between transformations c_*, c^* and validity $\models$. It generalises Lemma 4.1.5 from deterministic to arbitrary channels.

Proposition 4.5.3 *Let $c\colon X \rightsquigarrow Y$ be a channel with a state $\omega \in \mathcal{D}(X)$ on its domain and an observable $q \in Obs(Y)$ on its codomain. Then:*

$$c_*(\omega) \models q \;=\; \omega \models c^*(q). \tag{4.15}$$

This equation is essentially the law of total expectation, see Remark 4.5.5 below. It is a satisfaction condition, as in the theory of institutions, see Goguen and Burstall (1992).

Proof. The result follows simply by unpacking the relevant definitions:

$$\begin{aligned}
c_*(\omega) \models q \;=\; \sum_{y\in Y} c_*(\omega)(y) \cdot q(y) &= \sum_{y\in Y} \Big(\sum_{x\in X} c(x)(y) \cdot \omega(x)\Big) \cdot q(y) \\
&= \sum_{x\in X} \omega(x) \cdot \Big(\sum_{y\in Y} c(x)(y) \cdot q(y)\Big) \\
&= \sum_{x\in X} \omega(x) \cdot c^*(q)(x) \;=\; \omega \models c^*(q). \qquad \square
\end{aligned}$$

We have already seen several instances of this basic result.

- Earlier we mentioned that marginalisation (of states) and weakening (of observables) are dual

to each other, see Equation (4.11). We can now see this as an instance of (4.15), using a projection $\pi_i \colon X_1 \times \cdots \times X_n \to X_i$ as (trivial) channel, in:

$$(\pi_i)_*(\omega) \models p \;=\; \omega \models (\pi_i)^*(p) \tag{4.16}$$

The left-hand side of (4.16) uses the i-th marginal of state $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n)$, namelY:

$$(\pi_i)_*(\omega) \;=\; \omega[0, \ldots, 0, 1, 0, \ldots, 0] \;\in\; \mathcal{D}(X_i),$$

On the right-hand-side of (4.16) the observable $p \in \mathit{Obs}(X_i)$ is weakenend to the following predicate.

$$(\pi_i)^*(p) \;=\; \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes p \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} \;\in\; \mathit{Pred}(X_1 \times \cdots \times X_n).$$

- Equation (4.3) is also an instance of (4.15), namely for a trivial channel $f \colon X \rightsquigarrow Y$ given by a function $f \colon X \to Y$, as in:

$$\begin{aligned} f_*(\omega) \models p &\overset{(4.15)}{=} \omega \models f^*(p) \\ &= \omega \models q \circ f \qquad \text{by Lemma 4.5.2 (9).} \end{aligned}$$

We continue with a series of observations putting observable transformation in a broader perspective.

Remark 4.5.4 As will be elaborated in Exercise 4.5.7, a (fuzzy) predicate $p \colon X \to [0, 1]$ can be identified with a channel $\widehat{p} := \mathit{flip} \circ p \colon X \rightsquigarrow \mathbf{2}$. Using the point predicate $\mathbf{1}_1$ on $\mathbf{2} = \{0, 1\}$ one can recover the original predicate p via observable / predicate transformation along this channel $\widehat{p}$ as $p = \widehat{p}^*(\mathbf{1}_1)$.

One can thus argue that *fuzzy* predicates, with values in $[0, 1]$, are unnecessary and that *sharp* (or even point) predicates, with values in $\{0, 1\}$, suffice in a probabilistic logic. Indeed, probabilistic reasoning is often based on *events*, which are subsets of the underlying space that can be identified with their characteristic functions (sharp predicates). This argument makes sense, but there is the price, since an additional (binary) channel $X \rightsquigarrow \mathbf{2}$ is needed. In the context of Bayesian networks, this means that an additional (binary) conditional probability table has to be added for each fuzzy predicate.

In the context of this book we like to explicitly use fuzzy predicates $X \to [0, 1]$ as a special subset of observables $X \to \mathbb{R}$. These predicates have their own algebraic properties, formalised in terms of effect modules, see Subsection 4.4.3. But we easily switch back and forth between predicates $X \to [0, 1]$ and channels $X \rightsquigarrow \mathbf{2}$. As we shall see in Section 6.5, in the probabilistic logic programming language ProbLog the difference between the two is hard to see.

Remark 4.5.5 The above equation of validities (4.15) may be seen as an explicit formulation of the law of total expectation. The latter law is formulated, for instance in (Rice, 2006, §4.4.1, Thm. A), as an equation:

$$\mathbb{E}(Y) \;=\; \mathbb{E}\big[\; \mathbb{E}(Y \mid X) \,\big]. \tag{4.17}$$

This formulation involves random variables X, Y but leaves the relevant distribution and channel implicit. Thus some exegesis is needed. The expression $\mathbb{E}(Y \mid X)$ denotes conditional expectation. It involves a conditional probability, forming a channel $c \colon X \rightsquigarrow Y$. Additionally, there is an

observable, which we write as $q\colon Y \to \mathbb{R}$. This gives for each element $x \in X$ a random variable $(c(x), q)$, of which we can form the expected value:

$$c(x) \models q \overset{(4.14)}{=} c^*(q)(x).$$

This conditional expectation $\mathbb{E}(Y \mid X)$ is the function $x \mapsto (c(x) \models q)$ and may thus be identified with the transformed observable $c^*(q)\colon X \to \mathbb{R}$. Together with an implicit distribution $\omega \in \mathcal{D}(X)$ we thus have a random variable $(\omega, c^*(q))$ whose expected value $\omega \models c^*(q)$ is the right-hand-side of (4.17) — and also of (4.15).

On the left-hand-side of (4.17) we simply have an expected value $\mathbb{E}(Y)$. It refers to the expected value for the pushforward distribution $c_*(\omega) \in \mathcal{D}(Y)$. Recall, on Y we had an observable $q\colon Y \to \mathbb{R}$. Together they form a random variable $(c_*(\omega), q)$. Its expected value is $c_*(\omega) \models q$, which is the left-hand-side of (4.15). This shows that the two equations (4.17) and (4.15) express the same property, albeit in quite different ways.

Remark 4.5.6 In a programming context, where a channel $c\colon X \rightsquigarrow Y$ is seen as a program taking inputs from X to outputs in Y, one may call $c^*(q)$ the *weakest precondition* of q, commonly written as $wp(c, q)$, see *e.g.* Dijkstra and Scholten (1990); Kozen (1981); McIver and Morgan (2004). We briefly explain this perspective.

A precondition of q, w.r.t. channel $c\colon X \rightsquigarrow Y$, may be defined as an observable p on the channel's domain X for which:

$$\omega \models p \;\leq\; c_*(\omega) \models q, \qquad \text{for all states } \omega.$$

Proposition 4.5.3 tells that $c^*(q)$ is then a precondition of q. It is also the weakest, since if p is a precondition of q, as described above, then in particular:

$$p(x) \;=\; unit(x) \models p \;\leq\; c_*(unit(x)) \models q \;=\; c(x) \models q \;=\; c^*(q)(x).$$

As a result, $p \leq c^*(q)$.

Lemma 4.5.2 (6) expresses a familiar compositionality property in the theory of weakest preconditions:

$$wp(d \circ c, q) \;=\; (d \circ c)^*(q) \;=\; c^*(d^*(q)) \;=\; wp(c, wp(d, q)).$$

Remark 4.5.7 There is also a purely diagrammatic proof of Proposition 4.5.3, which works in the following way. First we recall from Lemma 4.1.5 that for an observable $p\colon X \to \mathbb{R}$, validity-of-p can be described as a composite:

$$(-) \models p \;=\; \Big(\mathcal{D}(X) \xrightarrow{\mathcal{D}(p)} \mathcal{D}(\mathbb{R}) \xrightarrow{Mean} \mathbb{R}\Big).$$

Moreover, for a channel $c\colon X \rightsquigarrow Y$ and an observable $q\colon Y \to \mathbb{R}$ we can describe the transformed observable $c^*(q)$ as:

$$c^*(q) \;=\; \Big(X \xrightarrow{c} \mathcal{D}(Y) \xrightarrow{\mathcal{D}(q)} \mathcal{D}(\mathbb{R}) \xrightarrow{Mean} \mathbb{R}\Big). \tag{4.18}$$

We can now describe the equation in Proposition 4.5.3 as an equation of maps:

$$\begin{tikzcd} \mathcal{D}(X) \arrow[r, bend left, "c_*(-)\models q"] \arrow[r, bend right, "(-)\models c^*(q)"'] & \mathbb{R} \end{tikzcd}$$

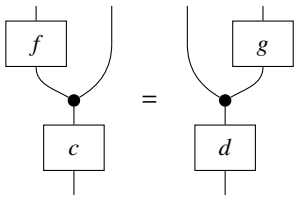
The upper arrow is the east-south route and the lower arrow is the south-east rout in the diagram below.

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "\mathcal{D}(c)"] \arrow[rrr, bend left, "c_*(-)"] \arrow[ddr, bend right, "\mathcal{D}(c^*(q))"'] & \mathcal{D}^2(Y) \arrow[r, "flat"] \arrow[d, "\mathcal{D}^2(q)"'] & \mathcal{D}(Y) \arrow[d, "\mathcal{D}(q)"] \\
& \mathcal{D}^2(\mathbb{R}) \arrow[r, "flat"] \arrow[d, "\mathcal{D}(Mean)"'] & \mathcal{D}(\mathbb{R}) \arrow[d, "Mean"] \\
& \mathcal{D}(\mathbb{R}) \arrow[r, "Mean"'] & \mathbb{R}
\end{tikzcd}$$

This diagram commutes by naturality of flattening and because *Mean* is a $\mathcal{D}$-algebra, see Exercise 4.2.5.

In Remark 4.4.10 we have seen that two joint distributions are equal if and only they give the same validity for (product) predicates. The result below applies this observation in a frequently occurring situation.

Lemma 4.5.8 *Appropriately types channels c, d and f, g satisfy:*



$$\iff \quad \begin{cases} c(x) \models f^*(p) \mathbin{\&} q \\ \quad = d(x) \models p \mathbin{\&} g^*(q) \\ \textit{for all inputs } x \textit{ and predicates } p, q. \end{cases}$$

Algebraically (or categorically) oriented readers may recognise an adjointness in the formulation on the right-hand-side, see also Exercise 4.5.4.

Proof. This follows from the equivalence (4.12) in Remark 4.4.10 using that, by Lemma 4.5.2 (7) and Proposition 4.5.3:

$$\begin{aligned} c(x) \models f^*(p) \mathbin{\&} q &= c(x) \models \langle f, id\rangle^*(p \otimes q) = \langle f, id\rangle_*(c(x)) \models p \otimes q \\ d(x) \models p \mathbin{\&} g^*(q) &= d(x) \models \langle id, g\rangle^*(p \otimes q) = \langle id, g\rangle_*(d(x)) \models p \otimes q. \end{aligned}$$

❑

We close this section with two topics that dig deeper into the nature of transformations. First, we reformulate transformation of states and observables in terms of matrix operations. Then we look closer at the categorical aspects of transformation of observables.

Remark 4.5.9 Let c be a channel with finite sets as domain and codomain. For convenience we write these as $\boldsymbol{n} = \{0, \ldots, n-1\}$ and $\boldsymbol{m} = \{0, \ldots, m-1\}$, so that the channel c has type $\boldsymbol{n} \rightsquigarrow \boldsymbol{m}$. For each $i \in \boldsymbol{n}$ we have that $c(i) \in \mathcal{D}(\boldsymbol{m})$ is given by an m-tuple of numbers in $[0, 1]$ that add up to one. Thus we can associate an $m \times n$ matrix M_c with the channel c, namely:

$$M_c = \begin{pmatrix} c(1)(1) & \cdots & c(n-1)(1) \\ \vdots & & \vdots \\ c(1)(m-1) & \cdots & c(n-1)(m-1) \end{pmatrix}.$$

By construction, the columns of this matrix add up to one. Such matrices are often called *stochastic*.

A state $\omega \in \mathcal{D}(\boldsymbol{n})$ may be identified with a column vector M_ω of length n, as on the left below. It is then easy to see that the matrix $M_{c_*(\omega)}$ of the transformed state, is obtained by *matrix-column multiplication*, as on the right:

$$M_\omega \;=\; \begin{pmatrix} \omega(0) \\ \vdots \\ \omega(n-1) \end{pmatrix} \qquad \text{so that} \qquad M_{c_*(\omega)} \;=\; M_c \cdot M_\omega.$$

Indeed,

$$c_*(\omega)(j) \;=\; \sum\nolimits_i c(i)(j) \cdot \omega(i) \;=\; \sum\nolimits_i (M_c)_{ji} \cdot (M_\omega)_i \;=\; (M_c \cdot M_\omega)_j.$$

An observable $q\colon \boldsymbol{m} \to \mathbb{R}$ on $\boldsymbol{m}$ can be identified with a row vector $M_q = \big(q(0) \;\cdots\; q(m-1)\big)$. Transformation $c^*(q)$ then corresponds to *row-matrix multiplication*:

$$M_{c^*(q)} \;=\; M_q \cdot M_c.$$

Again, this is checked easily:

$$c^*(q)(i) \;=\; \sum\nolimits_j q(j) \cdot c(i)(j) \;=\; \sum\nolimits_j (M_q)_j \cdot (M_c)_{j,i} \;=\; (M_q \cdot M_c)_i.$$

We close this section by making the functoriality of observable transformation explicit, in the style of Proposition 4.4.11. The latter deals with functions, but we now consider functoriality with respect to channels, using the category $\mathbf{Chan} = \mathbf{Chan}(\mathcal{D})$ of probabilistic channels. Notice that the case of sharp predicates is omitted from Proposition 4.4.11, simply because sharp predicates are not closed under predicate transformation (see Exercise 4.5.2). Also, conjunction & is not preserved under transformation, see Exercise 4.5.1 below.

Proposition 4.5.10 *Taking particular observables on a set is functorial, namely, via functors:*

1. *Obs*: $\mathbf{Chan} \to \mathbf{Vect}^{\mathrm{op}}$;
2. *Fact*: $\mathbf{Chan} \to \mathbf{Cone}^{\mathrm{op}}$;
3. *Pred*: $\mathbf{Chan} \to \mathbf{EMod}^{\mathrm{op}}$;

On a channel $c\colon X \rightsquigarrow Y$, all these functors are given by transformation c^*, acting in the opposite direction.

Proof. Essentially, all necessary ingredients are already in Lemma 4.5.2: transformation restricts appropriately (item (1)), transformation preserves identies (item (5)) and composition (item (6)), and the relevant structure (items (2) and (3)). ❑

Exercises

4.5.1 Consider the channel $f\colon \{a,b,c\} \rightsquigarrow \{u,v\}$ from Example 1.11.2, given by:

$$f(a) \;=\; \tfrac{1}{2}|u\rangle + \tfrac{1}{2}|v\rangle \qquad f(b) \;=\; 1|u\rangle \qquad f(c) \;=\; \tfrac{3}{4}|u\rangle + \tfrac{1}{4}|v\rangle.$$

Take as predicates $p, q\colon \{u,v\} \to [0,1]$,

$$p(u) \;=\; \tfrac{1}{2} \qquad p(v) \;=\; \tfrac{2}{3} \qquad q(u) \;=\; \tfrac{1}{4} \qquad q(v) \;=\; \tfrac{1}{6}.$$

Alternatively, in the normal-form notation of Lemma 4.4.3 (2),

$$p \;=\; \tfrac{1}{2} \cdot \mathbf{1}_u + \tfrac{2}{3} \cdot \mathbf{1}_v \qquad\qquad q \;=\; \tfrac{1}{4} \cdot \mathbf{1}_u + \tfrac{1}{6} \cdot \mathbf{1}_v.$$

Compute:

- $f^*(p)$
- $f^*(q)$
- $f^*(p \otimes q)$
- $f^*(p) \otimes f^*(q)$
- $f^*(p \,\&\, q)$
- $f^*(p) \,\&\, f^*(q)$

This will show that predicate transformation along a channel does *not* preserve conjunction &. However, conjunctions are preserved when we restrict to deterministic channels, see Exercise 4.5.5.

4.5.2 1 Still in the context of the previous exercise, consider the sharp (point) predicate $\mathbf{1}_u$ on $\{u, v\}$. Show that the transformed predicate $f^*(\mathbf{1}_u)$ on $\{a, b, c\}$ is not sharp. This proves that sharp predicates are not closed under predicate transformation.

2 Let $h\colon X \to Y$ be a function, considered as deterministic channel, and let $V \subseteq Y$ be an arbitrary subset (event). Check that:

$$h^*(\mathbf{1}_V) \;=\; \mathbf{1}_{h^{-1}(V)}.$$

Conclude that predicate transformation along deterministic channels does preserve sharpness.

4.5.3 Show that an n-test $p_0, \ldots, p_{n-1}$ on a set X, see Definition 4.4.2, can be identified with a channel $c\colon X \twoheadrightarrow \boldsymbol{n}$, with $p_i = c^*(\mathbf{1}_i)$, for $i \in \boldsymbol{n}$.

4.5.4 In the setting of Exercise 2.1.16 let X be a finite set with N elements. We have a function $size := \|-\|\colon \mathcal{N}(X) \to \mathbb{N}$, and in the other direction a channel $size^\dagger\colon \mathbb{N} \to \mathcal{D}(\mathcal{N}(X))$ given by:

$$size^\dagger(k) \;:=\; \sum_{\varphi \in \mathcal{N}[k](X)} \frac{(\varphi)}{N^k} \big|\varphi\rangle \;\in\; \mathcal{D}(\mathcal{N}(X)) \;\hookrightarrow\; \mathcal{D}_\infty(\mathcal{N}(X)).$$

This is well-defined by Exercise 1.7.9.

1 Show that $(size^\dagger)_*(pois[\lambda]) = mpois[\lambda]$, see Examples 2.1.8 (1) and (2).

2 Show that for $p \in Pred(\mathbb{N})$ and $q \in Pred(\mathcal{N}(X))$, there is an adjointness equation:

$$pois[\lambda] \models p \,\&\, (size^\dagger)^*(q) \;=\; mpois[\lambda] \models size^*(p) \,\&\, q.$$

3 Draw the corresponding equation between string diagrams.

4.5.5 Let $h\colon X \to Y$ be an ordinary function. Recall from Lemma 4.5.2 (9) that $h^*(q) = q \circ h$, when h is considered as a deterministic channel. Show that transformation along such deterministic channels does preserve conjunctions:

$$h^*(q_1 \,\&\, q_2) \;=\; h^*(q_1) \,\&\, h^*(q_2),$$

in contrast to the findings in Exercise 4.5.1 for arbitrary channels.

Conclude that weakening preserves conjunction: $\mathbf{1} \otimes (p \,\&\, q) = (\mathbf{1} \otimes p) \,\&\, (\mathbf{1} \otimes q)$.

4.5.6 Let predicate $q_i \in Pred(Y)$ form a test, see Definition 4.4.2, and let $c\colon X \twoheadrightarrow Y$ be channel. Check that the transformed predicates $c^*(q_i)$ form a test on X.

4.5.7 1 Check that a predicate $p\colon X \to [0,1]$ can be identified with a channel $\widehat{p} := \mathit{flip} \circ p\colon X \rightsquigarrow \mathbf{2}$, see also Exercise 4.5.3. Describe this $\widehat{p}$ explicitly in terms of p and show that the predicate p can be recovered from the channel $\widehat{p}$ via transformation, as claimed in Remark 4.5.4.

2 Show that the 'and' channel $ac\colon \mathbf{2} \times \mathbf{2} \rightsquigarrow \mathbf{2}$ from Exercise 2.5.9 satisfies $ac \circ \langle \widehat{p}, \widehat{q} \rangle = \widehat{p \,\&\, q}$.

3 Check also that the 'not' channel $nc\colon \mathbf{2} \rightsquigarrow \mathbf{2}$ from Exercise 2.5.9 satisfies $nc \circ \widehat{p} = \widehat{p^{\perp}}$.

4 Finally, define also a channel $\mathit{scal}(r)\colon \mathbf{2} \rightsquigarrow \mathbf{2}$, for $r \in [0,1]$, so that $\mathit{scal}(r) \circ \widehat{p} = \widehat{r \cdot p}$.

4.5.8 Recall that a state $\omega \in \mathcal{D}(X)$ can be identified with a channel $\mathbf{1} \rightsquigarrow X$ with a trivial domain, and also that a predicate $p\colon X \to [0,1]$ can be identified with a channel $X \rightsquigarrow \mathbf{2}$, see Exercise 4.5.7. Check that under these identifications validity $\omega \models p$ can be identified with:

- state transformation $p_*(\omega)$;
- predicate transformation $\omega^*(p)$;
- channel composition $p \circ \omega$.

4.5.9 This exercises shows how parallel conjunction $\otimes$ and sequential conjunction $\&$ are interdefinable via predicate transformation, using projection channels π_i and copy channels Δ, that is, using weakening and contraction.

1 Let observables p_1 on X_1 and p_2 on X_2 be given. Show that on $X_1 \times X_2$,

$$p_1 \otimes p_2 \;=\; (\pi_1)^*(p_1) \,\&\, (\pi_2)^*(p_2) \;=\; (p_1 \otimes \mathbf{1}) \,\&\, (\mathbf{1} \otimes p_2).$$

The last equation occurred already in Lemma 4.4.9 (2).

2 Let q_1, q_2 be observables on the same set Y. Prove that on Y,

$$q_1 \,\&\, q_2 \;=\; \Delta^*(q_1 \otimes q_2).$$

4.5.10 Let $c_i\colon X \rightsquigarrow Y$ be a finite collection of channels, and let correspondingly, probabilities $r_i \in [0,1]$ satisfy $\sum_i r_i = 1$. We then get a a new channel $\sum_i r_i \cdot c_i\colon X \rightsquigarrow Y$ as (pointwise) convex sum. Show that for a distribution $\omega \in \mathcal{D}(X)$ and an observation q on Y one has:

$$\big(\textstyle\sum_i r_i \cdot c_i\big)_*(\omega) \;=\; \sum_i r_i \cdot c_{i*}(\omega) \qquad\qquad \big(\textstyle\sum_i r_i \cdot c_i\big)^*(q) \;=\; \sum_i r_i \cdot c_i{}^*(q).$$

4.5.11 Let $c\colon X \rightsquigarrow Y$ be a channel, with observables p on Z and q on $Y \times Z$. Check that:

$$(c \otimes \mathit{id})^*\Big((\mathbf{1} \otimes p) \,\&\, q\Big) \;=\; (\mathbf{1} \otimes p) \,\&\, (c \otimes \mathit{id})^*(q).$$

(Recall that $\&$ is not preserved by predicate transformation, see Exercise 4.5.1.)

4.5.12 We take a closer look at the functor

$$\mathbf{Chan} \xrightarrow{\mathit{Pred}} \mathbf{EMod}^{\mathrm{op}}$$

from Proposition 4.4.11. It sends a map $c\colon X \rightsquigarrow Y$ in **Chan**, *i.e.* a channel, to the function $\mathit{Pred}(c)\colon \mathit{Pred}(Y) \to \mathit{Pred}(X)$ via:

$$\mathit{Pred}(c)(q) \;:=\; c^*(q).$$

1 Check in detail that $\mathit{Pred}(c)$ is a morphism of effect modules, see Definition 4.4.1 (3).

2 Show that the functor Pred is *faithful*, in the sense that $\mathit{Pred}(c) = \mathit{Pred}(c')$ implies $c = c'$, for channels $c, c'\colon X \rightsquigarrow Y$.

3 Let Y be a finite set. Show that for each map $h\colon Pred(Y) \to Pred(X)$ in the category **EMod** there is a unique channel $c\colon X \twoheadrightarrow Y$ with $Pred(c) = h$.
Hint: Write a predicate p as finite sum $\bigcirc\!\!\!\!\vee_y\, p(y) \cdot \mathbf{1}_y$, *i.e.* as the normal form described in Lemma 4.4.3 (2), and use the relevant preservation property.
One says that the functor *Pred* is *full and faithful* when restricted to the (sub)category with finite sets as objects. In the context of programming (logics) this property is called *healthiness*, see Dijkstra (1976); Dijkstra and Scholten (1990); McIver and Morgan (2004), or Hino et al. (2016) for an abstract account.

4.6 Validity and drawing

In Chapter 3 we have studied various distributions associated with drawing coloured balls from an urn, such as the multinomial, hypergeometric and Pólya distributions. In this section we look at validity with respect to these draw distributions. This involves means and sampling.

We start with a basic result expressing that the validity of an observable in a distribution can also be described in terms of validity of the same observable in normalised (multinomial) draws, while taking the probabilities of these draws into account.

Proposition 4.6.1 *For a distribution $\omega \in \mathcal{D}(X)$ and an observable $p\colon X \to \mathbb{R}$ one has, for $K \geq 1$,*

$$\omega \models p \;=\; mn[K](\omega) \models \big(Flrn(-) \models p\big), \tag{4.19}$$

where $Flrn(-) \models p\colon \mathcal{N}[K](X) \to \mathbb{R}$ is the observable given by $\varphi \mapsto (Flrn(\varphi) \models p)$.

There are similar results for hypergeometric and Pólya draws, see Exercise 4.6.1.

Proof. We give two proofs, one concrete and one more abstract. The concrete proof uses Lemma 3.3.3 and involves a calculation:

$$\begin{aligned}
mn[K](\omega) \models \big(Flrn(-) \models p\big) &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot (Flrn(\varphi) \models p) \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \Big(\sum_{x \in X} Flrn(\varphi)(x) \cdot p(x)\Big) \\
&= \frac{1}{K} \cdot \sum_{x \in X} \Big(\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x)\Big) \cdot p(x) \\
&= \frac{1}{K} \cdot \sum_{x \in X} (K \cdot \omega(x)) \cdot p(x) \;=\; \sum_{x \in X} \omega(x) \cdot p(x) \;=\; \omega \models p.
\end{aligned}$$

The abstract proof uses a commuting diagram of the form:

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "mn[K]"] \arrow[drrr, equal] & \mathcal{D}(\mathcal{M}[K](X)) \arrow[r, "\mathcal{D}(Flrn)"] & \mathcal{D}(\mathcal{D}(X)) \arrow[r, "\mathcal{D}(-\models p)"] \arrow[d, "flat"'] & \mathcal{D}(\mathbb{R}) \arrow[d, "Mean"] \\
& & \mathcal{D}(X) \arrow[r, "(-)\models p"] & \mathbb{R}
\end{tikzcd}$$

The rectangle on the right commutes by Exercise 4.2.5 (3), and the triangle on the left by Theorem 3.3.4. ❑

Corollary 4.6.2 *For a distribution $\omega \in \mathcal{D}(\mathbb{R})$ on the reals and a number $K > 0$,*

$$mn[K](\omega) \models Mean(Flrn(-)) \;=\; Mean(\omega) \;=\; K \cdot \omega \models \tfrac{(-)}{K}, \tag{4.20}$$

where $K \cdot \omega = \omega + \cdots + \omega \in \mathcal{D}(\mathbb{R})$ is the K-fold convolution.

Proof. The first equation in (4.20) is obtained as special case of (4.19) by taking $p = id$.

For the second equation in (4.20) we use that the K-fold convolution can be expressed via multionials, as: $K \cdot \omega = \mathcal{D}(sum)(mn[K](\omega))$, see Proposition 2.7.10, where $sum(\varphi) = \sum_x \varphi(x) \cdot x \in \mathbb{R}$. We additionally use:

$$\frac{sum(\varphi)}{K} \;=\; \frac{1}{K} \cdot \sum_{x \in \mathbb{R}} \varphi(x) \cdot x \;=\; \sum_{x \in \mathbb{R}} Flrn(\varphi)(x) \cdot x \;=\; Flrn(\varphi) \models id \;=\; Mean(Flrn(\varphi)). \tag{4.21}$$

Now:

$$\begin{array}{rcll} K \cdot \omega \models \frac{(-)}{K} & = & \mathcal{D}(sum)(mn[K](\omega))\omega \models \frac{(-)}{K} & \text{by Proposition 2.7.10} \\ & \overset{(4.3)}{=} & mn[K](\omega) \models \frac{sum(-)}{K} & \\ & \overset{(4.21)}{=} & mn[K](\omega) \models Mean(Flrn(-)) & \\ & = & Mean(\omega) & \text{as just shown.} \end{array}$$

❑

In Example 4.2.2 (2) we have seen the mean of a binomial distribution. A multinomial $mn[K](\omega)$ is a distribution on the set $\mathcal{N}[K](X)$ of natural multisets of size K, and not on (real) numbers. Hence the requirement for a mean, see Definition 4.2.1, does not apply: the space is not a subset of the reals. Still, we have used a generalised mean for multinomials in Proposition 3.3.7. The trick is to include multisets in $\mathcal{N}[K](X)$ in the bigger set $\mathcal{M}(X)$ of multisets, of arbitrary size, with (non-negative) real-valued multiplicities. The latter set $\mathcal{M}(X)$ is a cone, see lemma 1.6.3 (2), and thus has enough structure (addition and scalar multiplication) to compute means, via a function $Mean \colon \mathcal{D}(\mathcal{M}(X)) \to \mathcal{M}(X)$, given as $Mean(\sum_i r_i|\varphi_i\rangle) = \sum_i r_i \cdot \varphi_i$.

Proposition 4.6.3 *We fix a set X and consider means in $\mathcal{M}(X)$.*

1 *For a distribution $\omega \in \mathcal{D}(X)$,*

$$Mean\Big(mn[K](\omega)\Big) := \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi \;=\; K \cdot \omega \;\in\; \mathcal{M}(X).$$

2 *For a non-empty urn $\upsilon \in \mathcal{N}[L](X)$ of size $L \geq K$,*

$$Mean\Big(hg[K](\upsilon)\Big) := \sum_{\varphi \in \leq_K \upsilon} hg[K](\upsilon)(\varphi) \cdot \varphi \;=\; K \cdot Flrn(\upsilon) \;\in\; \mathcal{M}(X).$$

3 *For a non-empty urn $\upsilon \in \mathcal{N}(X)$,*

$$Mean\Big(pl[K](\upsilon)\Big) := \sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} pl[K](\upsilon)(\varphi) \cdot \varphi \;=\; K \cdot Flrn(\upsilon) \;\in\; \mathcal{M}(X).$$

4 *For a distribution $\omega \in \mathcal{D}(X)$ and a rate $\lambda \in \mathbb{R}_{>0}$,*

$$Mean\Big(Pmn[\lambda](\omega)\Big) := \sum_{\varphi \in \mathcal{N}(X)} Pmn[\lambda](\omega)(\varphi) \cdot \varphi \;=\; \lambda \cdot \omega \;\in\; \mathcal{M}(X).$$

The Poisson-iid distribution $Piid[\lambda](\omega) \in \mathcal{D}_\infty(\mathcal{L}(X))$ does not have such a mean since the set $\mathcal{L}(X)$ of lists is not a cone.

Proof. The first three items (1) – (3) follow from Proposition 3.3.7 together with Lemmas 3.5.5 (2) and 3.6.1 (2). Hence we concentrate on the last item, involving the Poission multinomial.

$$\begin{aligned}
Mean\Big(Pmn[\lambda](\omega)\Big)(x) &= \sum_{\varphi \in \mathcal{N}(X)} Pmn[\lambda](\omega)(\varphi) \cdot \varphi(x) \\
&\overset{(3.53)}{=} \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} pois[\lambda](K) \cdot mn[K](\omega)(\varphi) \cdot \varphi(x) \\
&= \sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x) \\
&= \sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot Mean(mn[K](\omega))(x) \\
&\overset{(1)}{=} \sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot K \cdot \omega(x) \\
&= Mean(pois[\lambda]) \cdot \omega(x) \;=\; \lambda \cdot \omega(x), \qquad \text{by Example 4.2.2 (3).} \qquad \square
\end{aligned}$$

One can also describe a 'pointwise mean' for multinomials via point-evaluation observables. For a set X with an element $x \in X$ we can define an observable:

$$\mathcal{M}(X) \xrightarrow{ev_x} \mathbb{R} \qquad \text{by} \qquad ev_x(\varphi) := \varphi(x). \tag{4.22}$$

Then we can compute the validity of this observable as:

$$\begin{aligned}
mn[K](\omega) \models ev_x &\overset{(4.2)}{=} \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot ev_x(\varphi) \;=\; \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x) \\
&= K \cdot \omega(x), \qquad \text{by Lemma 3.3.3.}
\end{aligned}$$

Hence we can alternatively write:

$$Mean\Big(mn[K](\omega)\Big) \;=\; \sum_{x \in X} \big(mn[K](\omega) \models ev_x\big) \,|x\rangle \;=\; K \cdot \omega.$$

The same can be done for the other draw distributions in Proposition 4.6.3, and also in the next situation.

Lemma 4.6.4 *Let X be a non-empty set of size N and let $K \in \mathbb{N}_{>0}$. The three distributions on multisets from (2.57) all have the same uniform mean, as multiset in $\mathcal{M}(X)$:*

$$Mean\Big(ums[X,K]\Big) \;=\; Mean\Big(mc[X,K]\Big) \;=\; Mean\Big(ssms[X,K]\Big) \;=\; \sum_{x \in X} \tfrac{K}{N} \,|x\rangle.$$

Proof. The three cases all follow from Exercise 4.2.2. We do the first one:

$$\begin{aligned}
Mean\Big(ums[K](X)\Big) &= \sum_{x \in X} ums[K](X) \models ev_x \,|x\rangle \\
&= \sum_{x \in X} ums[K](X) \models (id \circ ev_x) \,|x\rangle \\
&\overset{(4.3)}{=} \sum_{x \in X} \mathcal{D}(ev_x)\Big(ums[K](X)\Big) \models id \,|x\rangle \\
&= \sum_{x \in X} Mean\Big(\mathcal{D}(ev_x)\Big(ums[K](X)\Big)\Big) \,|x\rangle \\
&= \sum_{x \in X} \tfrac{K}{N} \,|x\rangle, \qquad \text{by Exercise 4.2.2.} \qquad \square
\end{aligned}$$

Example 4.6.5 We recall a historical example, known as the paradox of the Chevalier De Méré, from the 17th century. He argued informally that the following two outcomes have equal probability.

DM1 Throw a dice 4 times; you get at least one 6.
DM2 Throw a pair of dice 24 times; you get at least one double six, *i.e.* $(6, 6)$.

However, when De Méré betted on (*DM2*) he mostly lost. Puzzled, he wrote to Pascal, who showed that the probabilities differ.

Let us write $\mathit{1six} \colon \mathcal{N}(\mathit{pips}) \to \{0, 1\}$ for the sharp predicate that sends a multiset φ, over the dice space $\mathit{pips} = \{1, 2, 3, 4, 5, 6\}$, to 1 if $\varphi(6) \geq 1$ and to 0 otherwise. It thus tells that the number 6 occurs at least once in the draw φ. We can thus model the first option (*DM1*) as validity:

$$mn[4](\mathit{dice}) \models \mathit{1six}. \tag{DM1}$$

The second option (*DM2*) then becomes:

$$mn[24](\mathit{dice} \otimes \mathit{dice}) \models \mathit{2six}, \tag{DM2}$$

where $\mathit{2six} \colon \mathcal{N}(\mathit{dice} \times \mathit{dice}) \to \{0, 1\}$ is the obvious predicate with $\mathit{2six}(\psi) = 1$ iff $\psi(6, 6) \geq 1$.

Using a computer to calculate the validity (*DM1*) is relatively easy, giving an outcome $\frac{671}{1296}$. It involves summing over $\left(\!\binom{6}{4}\!\right) = 126$ multisets, see Proposition 1.8.7. However, the sum in (*DM1*) is huge, involving $\left(\!\binom{36}{24}\!\right)$ multisets — a number in the order of $2 \cdot 10^{16}$.

The common way to compute (*DM1*) and (*DM2*) is to switch to validity over the distribution *dice*, and using orthosupplement (negations). Getting at least one 6 in 4 throws is the orthosupplement of getting 4 times no 6. This can be represented and computed as:

$$\begin{aligned}
mn[4](\mathit{dice}) \models \mathit{1six} &= acc_*(iid[4](\mathit{dice})) \models \mathit{1six} \qquad \text{by Theorem 2.6.7} \\
&= iid[4](\mathit{dice}) \models acc^*(\mathit{1six}) \\
&= iid[4](\mathit{dice}) \models (acc^*(\mathit{1six}^\perp))^\perp \\
&\overset{(*)}{=} \mathit{dice} \otimes \mathit{dice} \otimes \mathit{dice} \otimes \mathit{dice} \models (\mathbf{1}_6^\perp \otimes \mathbf{1}_6^\perp \otimes \mathbf{1}_6^\perp \otimes \mathbf{1}_6^\perp)^\perp \\
&= 1 - \left(\mathit{dice} \models \mathbf{1}_6^\perp\right)^4 \qquad \text{by Lemmas 4.4.6 and 4.4.8} \\
&= 1 - \left(\tfrac{5}{6}\right)^4 = \tfrac{671}{1296} \approx 0.518.
\end{aligned}$$

The step $\overset{(*)}{=}$ uses for $\vec{x} \in \mathit{pips}^4$,

$$\begin{aligned}
acc^*(\mathit{1six}^\perp)(\vec{x}) = 1 &\iff \mathit{1six}(acc(\vec{x})) = 0 \\
&\iff acc(\vec{x})(6) < 1 \\
&\iff (x_1 \neq 6) \text{ and } (x_2 \neq 6) \text{ and } (x_3 \neq 6) \text{ and } (x_4 \neq 6) \\
&\iff \mathbf{1}_6(x_1) = 0 \text{ and } \mathbf{1}_6(x_2) = 0 \text{ and } \mathbf{1}_6(x_3) = 0 \text{ and } \mathbf{1}_6(x_4) = 0 \\
&\iff \mathbf{1}_6^\perp(x_1) \cdot \mathbf{1}_6^\perp(x_2) \cdot \mathbf{1}_6^\perp(x_3) \cdot \mathbf{1}_6^\perp(x_4) = 1 \\
&\iff (\mathbf{1}_6^\perp \otimes \mathbf{1}_6^\perp \otimes \mathbf{1}_6^\perp \otimes \mathbf{1}_6^\perp)(\vec{x}) = 1.
\end{aligned}$$

Similarly one can compute (*DM2*) as:

$$\begin{aligned}
mn[24](\mathit{dice} \otimes \mathit{dice}) \models \mathit{2six} &= (\mathit{dice} \otimes \mathit{dice})^{24} \models \left(\left((\mathbf{1}_6 \otimes \mathbf{1}_6)^\perp\right)^{24}\right)^\perp \\
&= 1 - \left(\mathit{dice} \otimes \mathit{dice} \models (\mathbf{1}_6 \otimes \mathbf{1}_6)^\perp\right)^{24} \\
&= 1 - \left(\tfrac{35}{36}\right)^{24} \approx 0.491.
\end{aligned}$$

Hence indeed, betting on (*DM2*) is a bad idea.

In this section we look at validities over distributions of draws from an urn, like in (*DM1*) and (*DM2*). In the remainder we establish connections between such validities and validities over the urn, as distribution. This involves free extensions of observables to multisets.

We notice that there are two (obvious) ways to extend an observable $X \to \mathbb{R}$ on a set X to natural multisets over X, since we can choose to use either the additive structure or the multiplicative structure on $\mathbb{R}$. Both these extensions are based on Proposition 1.6.5.

Definition 4.6.6 Let $p\colon X \to \mathbb{R}$ be an observable on a set X.

1 The *additive extension* $\overline{p}^{+}\colon \mathcal{N}(X) \to \mathbb{R}$ of p on multisets over X is defined as:

$$\overline{p}^{+}(\varphi) \;=\; \sum_{x \in supp(\varphi)} \varphi(x) \cdot p(x) \;=\; \sum_{x \in X} \varphi(x) \cdot p(x).$$

2 The *multiplicative extension* $\overline{p}^{\bullet}\colon \mathcal{N}(X) \to \mathbb{R}$ of p is:

$$\overline{p}^{\bullet}(\varphi) \;=\; \prod_{x \in supp(\varphi)} p(x)^{\varphi(x)} \;=\; \prod_{x \in X} p(x)^{\varphi(x)}.$$

By construction, these extensions are homomorphisms of monoids, so that:

$$\begin{cases} \overline{p}^{+}(\mathbf{0}) \;=\; 0 \\ \overline{p}^{\bullet}(\mathbf{0}) \;=\; 1. \end{cases} \qquad \text{and} \qquad \begin{cases} \overline{p}^{+}(\varphi + \psi) \;=\; \overline{p}^{+}(\varphi) + \overline{p}^{+}(\psi) \\ \overline{p}^{\bullet}(\varphi + \psi) \;=\; \overline{p}^{\bullet}(\varphi) \cdot \overline{p}^{\bullet}(\psi) \end{cases} \tag{4.23}$$

Once extended to multisets, we can investigate the validity of these extensions in distributions obtained from drawing. We concentrate on the multinomial and Poisson multinomial cases first where the validity of both the additive and the multiplicative extensions can be formulated in terms of the validity in the underlying state (urn).

Proposition 4.6.7 *Consider a random variable given by a distribution $\omega \in \mathcal{D}(X)$ and an observable $p\colon X \to \mathbb{R}$.*

1 For $K \in \mathbb{N}$, the additive extension $\overline{p}^{+}$ of p forms new random variable, with the multinomial distribution $mn[K](\omega)$ on $\mathcal{N}[K](X)$ as state. The associated validity is:

$$mn[K](\omega) \models \overline{p}^{+} \;=\; K \cdot (\omega \models p).$$

2 The multiplicative extension gives:

$$mn[K](\omega) \models \overline{p}^{\bullet} \;=\; (\omega \models p)^{K}.$$

3 For $\lambda \in \mathbb{R}_{>0}$, the validity of the additive extension in a Poisson multinomial is:

$$Pmn[\lambda](\omega) \models \overline{p}^{+} \;=\; \lambda \cdot (\omega \models p).$$

4 The validity of the multiplicative extension can be expressed as:

$$Pmn[\lambda](\omega) \models \overline{p}^{\bullet} \;=\; e^{-\lambda \cdot (\omega \models p^{\perp})}.$$

Proof. 1 We use Lemma 3.3.3 in:

$$\begin{aligned}
mn[K](\omega) \models \overline{p}^{+} &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \overline{p}^{+}(\varphi) \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \Big(\sum_{x \in X} \varphi(x) \cdot p(x) \Big) \\
&= \sum_{x \in X} p(x) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x) \\
&= \sum_{x \in X} p(x) \cdot K \cdot \omega(x) \;=\; K \cdot (\omega \models p).
\end{aligned}$$

2 By unfolding the multinomial distribution, see (2.54):

$$\begin{aligned}
mn[K](\omega) \models \overline{p}^{\bullet} &= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \overline{p}^{\bullet}(\varphi) \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \binom{}{\varphi} \cdot \Big(\prod\nolimits_x \omega(x)^{\varphi(x)} \Big) \cdot \Big(\prod\nolimits_x p(x)^{\varphi(x)} \Big) \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \binom{}{\varphi} \cdot \prod\nolimits_x \big(\omega(x) \cdot p(x)\big)^{\varphi(x)} \overset{(1.42)}{=} \Big(\textstyle\sum_x \omega(x) \cdot p(x) \Big)^{K} \;=\; (\omega \models p)^{K}.
\end{aligned}$$

3 The validity of the additive extension $\overline{p}^{+}$ in a Poisson point process state is computed as follows, using the first item for multinomials.

$$\begin{aligned}
Pmn[\lambda](\omega) \models \overline{p}^{+} &= \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} pois[\lambda](K) \cdot mn[K](\omega)(\varphi) \cdot \overline{p}^{+}(\varphi) \\
&= \sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot (mn[K](\omega) \models \overline{p}^{+}) \\
&\overset{(1)}{=} \sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot K \cdot (\omega \models p) \;=\; \lambda \cdot (\omega \models p) \qquad \text{by Example 4.2.2 (3).}
\end{aligned}$$

4 In the multiplicative case we get:

$$\begin{aligned}
Pmn[\lambda](\omega) \models \overline{p}^{\bullet} &= \sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot (mn[K](\omega) \models \overline{p}^{\bullet}) \\
&\overset{(2)}{=} \sum_{K \in \mathbb{N}} e^{-\lambda} \cdot \frac{\lambda^{K}}{K!} \cdot (\omega \models p)^{K} \\
&= \frac{e^{-\lambda}}{e^{-\lambda \cdot (\omega \models p)}} \cdot \sum_{K \in \mathbb{N}} e^{-\lambda \cdot (\omega \models p)} \cdot \frac{(\lambda \cdot (\omega \models p))^{K}}{K!} \;=\; e^{-\lambda \cdot (1 - \omega \models p)} \;=\; e^{-\lambda \cdot (\omega \models p^{\perp})}. \qquad \square
\end{aligned}$$

Remark 4.6.8 A validity $\omega \models p$ can be approximated via sampling. This may be useful in situations where the distribution ω has very large support, so that computing the sum $\sum_{x \in supp(\omega)} \omega(x) \cdot p(x) = \omega \models p$ takes too many resources. One can use the approach described in the code fragment below, where $K > 0$ is a parameter for the number of iterations. This is called importance

sampling.

```
v := 0
repeat K times
   x ← ω
   v := v + p(x)
return v/K
```
(4.24)

The justification for this approach is given by Proposition 4.6.7 (1). It takes the probabilities of draws of multisets φ into accounts and computes $\overline{p}^{+}(\varphi) = \sum_x \varphi(x) \cdot p(x)$ as in the above repeat loop. If we do this for all draws φ, divided by K, with their multinomial probabilities, we get:

$$\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \left(\frac{\sum_x \varphi(x) \cdot p(x)}{K} \right) = \frac{1}{K} \cdot \big(mn[K](\omega) \models \overline{p}^{+}(\varphi) \big) = \omega \models p.$$

A justification for (4.24) is given in limit form in Corollary 5.6.3 (1).

For the hypergeometric and Pólya distributions we have results for the additive extension. They resemble the formulation in Proposition 4.6.7 (1) for multinomials.

Proposition 4.6.9 *Consider an observable p and an urn υ.*

$$hg[K](\upsilon) \models \overline{p}^{+} \;=\; K \cdot (Flrn(\upsilon) \models p) \;=\; pl[K](\upsilon) \models \overline{p}^{+}.$$

Proof. Both equations are obtained as for Proposition 4.6.7 (1), this time using Lemmas 3.5.5 (1) and 3.6.1 (1). ❑

For the parallel multinomial law $pml \colon \mathcal{M}[K](\mathcal{D}(X)) \to \mathcal{D}(\mathcal{M}[K](X))$ from Section 3.7 there is a similar result, in the multiplicative case.

Proposition 4.6.10 *For an observable $p \colon X \to \mathbb{R}$ and for a multiset of distributions $\sum_i n_i|\omega_i\rangle \in \mathcal{M}[K](\mathcal{D}(X))$,*

$$pml(\textstyle\sum_i n_i|\omega_i\rangle) \models \overline{p}^{\bullet} \;=\; \prod_i (\omega_i \models p)^{n_i}.$$

Proof. We use the formulation (3.36) of pml in:

$$\begin{aligned}
pml(\textstyle\sum_i n_i|\omega_i\rangle) \models \overline{p}^{\bullet} &= \sum_{i,\varphi_i \in \mathcal{N}[n_i](X)} \big(\textstyle\prod_i mn[n_i](\omega_i)(\varphi_i) \big) \cdot \overline{p}^{\bullet}(\textstyle\sum_i \varphi_i) \\
&= \sum_{i,\varphi_i \in \mathcal{N}[n_i](X)} \big(\textstyle\prod_i mn[n_i](\omega_i)(\varphi_i) \big) \cdot \big(\textstyle\prod_i \overline{p}^{\bullet}(\varphi_i) \big) \\
&\overset{(4.23)}{=} \sum_{i,\varphi_i \in \mathcal{N}[n_i](X)} \textstyle\prod_i mn[n_i](\omega_i)(\varphi_i) \cdot \overline{p}^{\bullet}(\varphi_i) \\
&= \prod_i \sum_{\varphi_i \in \mathcal{N}[n_i](X)} mn[n_i](\omega_i)(\varphi_i) \cdot \overline{p}^{\bullet}(\varphi_i) \\
&= \prod_i mn[n_i](\omega_i) \models \overline{p}^{\bullet} \\
&= \prod_i (\omega_i \models p)^{n_i}, \quad \text{by Proposition 4.6.7 (2).}
\end{aligned}$$

❑

For the following result we use the extension $\mathcal{N}[K] \colon \mathbf{Chan} \to \mathbf{Chan}$ of the (natural) multiset functor to the category of probabilistic channels, see Corollary 3.8.7.

Corollary 4.6.11 *Let $c\colon X \rightsquigarrow Y$ be a channel, with a factor $q\colon Y \to \mathbb{R}_{\geq 0}$ on its codomain. Then we have an equality of predicates on $\mathcal{N}[K](X)$ of the form:*

$$\mathcal{N}[K](c)^*(\overline{q}^\bullet) \;=\; \overline{c^*(q)}^\bullet.$$

Proof. For $\varphi \in \mathcal{N}[K](X)$,

$$\begin{aligned}
\mathcal{N}[K](c)^*(\overline{q}^\bullet)(\varphi) &\overset{(4.14)}{=} \mathcal{N}[K](c)(\varphi) \models \overline{q}^\bullet \overset{(3.42)}{=} pml\big(\mathcal{N}(c)(\varphi)\big) \models \overline{q}^\bullet \\
&= pml\Big(\sum_{x\in X} \varphi(x)\big|c(x)\rangle\Big) \models \overline{q}^\bullet \\
&= \sum_{x\in X} (c(x) \models q)^{\varphi(x)} \qquad \text{by Proposition 4.6.10} \\
&\overset{(4.14)}{=} \sum_{x\in X} c^*(q)(x)^{\varphi(x)} \;=\; \overline{c^*(q)}^\bullet(\varphi).
\end{aligned}$$

❑

Exercises

4.6.1 Let $\upsilon \in \mathcal{N}(X)$ be an urn with an observable $p\colon X \to \mathbb{R}$ and let $K \geq 1$. Prove the following analogues of Proposition 4.6.1.

1 $hg[K](\upsilon) \models \big(Flrn(-) \models p\big) \;=\; Flrn(\upsilon) \models p$, for $K \leq \|\upsilon\|$.

2 $pl[K](\upsilon) \models \big(Flrn(-) \models p\big) \;=\; Flrn(\upsilon) \models p$.

4.6.2 Let $p\colon X \to \mathbb{R}$ be an observable. Show that:

$$\begin{aligned}
\overline{p}^+(acc(x_1, \ldots, x_n)) &= p(x_1) + \ldots + p(x_n) \\
\overline{p}^\bullet(acc(x_1, \ldots, x_n)) &= p(x_1) \cdot \ldots \cdot p(x_n).
\end{aligned}$$

4.6.3 Show that the additive/multiplicative extension preserves the additive/multiplicative structure of observables:

$$\overline{p+q}^+ \;=\; \overline{p}^+ + \overline{q}^+ \qquad\qquad \overline{\mathbf{0}}^+ \;=\; \mathbf{0}$$

and:

$$\overline{p \,\&\, q}^\bullet \;=\; \overline{p}^\bullet \,\&\, \overline{q}^\bullet \qquad\qquad \overline{\mathbf{1}}^\bullet \;=\; \mathbf{1}$$

4.6.4 Let $E \subseteq X$ be a subset (event), corresponding to the sharp indicator predicate $\mathbf{1}_E\colon X \to [0,1]$. Check that for an arbitrary multiset $\varphi \in \mathcal{N}(X)$,

$$\overline{\mathbf{1}_E}^\bullet(\varphi) = 1 \iff supp(\varphi) \subseteq E.$$

Note that this also holds for $\overline{\mathbf{1}_E}^\bullet(\mathbf{0}) = 1$, since a product over an empty set equals one. Conclude that $\overline{\mathbf{1}_E}^\bullet\colon \mathcal{N}(X) \to [0,1]$ is also a sharp predicate.

4.6.5 For a random variable $p\colon X \to \mathbb{R}$ with state $\omega \in \mathcal{D}(X)$ consider the validity $mn[-](\omega) \models \overline{p}^+$ as an observable $\mathbb{N} \to \mathbb{R}$. Show that for a distribution $\sigma \in \mathcal{D}(\mathbb{N})$ one has:

$$\sigma \models \big(mn[-](\omega) \models \overline{p}\big) \;=\; Mean(\sigma) \cdot (\omega \models p).$$

4.6.6 For a random variable (ω, p) write $\sum(\omega, p)\colon \mathbb{N} \to \mathbb{R}$ for the summation observable defined by:

$$\sum(\omega, p)(n) \;:=\; \omega \models p + \cdots + p \quad (n \text{ times}).$$

Show that for a distribution $\sigma \in \mathcal{D}(\mathbb{N})$ one has:

$$\sigma \models \sum(\omega, p) \;=\; Mean(\sigma) \cdot (\omega \models p).$$

4.6.7 The following is often used as illustration of Wald's identity. Roll a dice, and let $n \in pips = \{1, \ldots, 6\}$ be the number that comes up; then roll the dice n more times and record the sum of the resulting pips. What is the expected sum?

1 Use Exercise 4.6.6 to show that the expected number if $\frac{49}{4}$.
2 Obtain this same outcome via Proposition 4.6.7 (1).

4.6.8 Let $p \colon X \to \mathbb{R}$ be an observable. In Definition 4.6.6 we have introduced $\overline{p}^{+}$ and $\overline{p}^{\bullet}$ as extensions of p from X to multisets. We can however also extend p to lists, using freeness, see Proposition 1.4.3. In this exercise we write $\overline{p}^{+} \colon \mathcal{L}(X) \to \mathbb{R}$ and $\overline{p}^{\bullet} \colon \mathcal{L}(X) \to \mathbb{R}$ for these extensions.

1 Describe $\overline{p}^{+}([x_1, \ldots, x_K])$ and $\overline{p}^{\bullet}([x_1, \ldots, x_K])$ concretely.
2 Prove that:

$$iid[K](\omega) \models \overline{p}^{+} \;=\; K \cdot (\omega \models p) \qquad\qquad iid[K](\omega) \models \overline{p}^{\bullet} \;=\; (\omega \models p)^{K}.$$

3 Show next that:

$$Piid[\lambda](\omega) \models \overline{p}^{+} \;=\; \lambda \cdot (\omega \models p) \qquad\qquad Piid[\lambda](\omega) \models \overline{p}^{\bullet} \;=\; e^{-\lambda \cdot (\omega \models p^{+})}.$$

4.6.9 Consider a rate $\lambda \in \mathbb{R}_{\geq 0}$, a finite set X, and a distribution $\omega \in \mathcal{D}(X)$ with full support.

1 Prove that the length function $\| - \| \colon \mathcal{L}(X) \to \mathbb{N}$ satisfies:

$$Piid[\lambda](\omega) \models \| - \| \;=\; \sum_{\ell \in \mathcal{L}(X)} Piid[\lambda](\omega)(\ell) \cdot \|\ell\| \;=\; \lambda.$$

2 Prove that for $\lambda' \in \mathbb{R}_{\geq 0}$ and $\omega \in \mathcal{D}(X)$,

$$D_{KL}\Big(Piid[\lambda](\omega),\, Piid[\lambda'](\omega')\Big) \;=\; D_{KL}\Big(pois[\lambda],\, pois[\lambda']\Big) + \lambda \cdot D_{KL}(\omega, \omega').$$

4.6.10 Similarly, assume $\lambda \in \mathbb{R}_{\geq 0}$ and $\omega \in \mathcal{D}(X)$ with full support.

1 Show that the size function $\| - \| \colon \mathcal{N}(X) \to \mathbb{N}$ satisfies:

$$Pmn[\lambda](\omega) \models \| - \| \;=\; \sum_{\varphi \in \mathcal{N}(X)} Pmn[\lambda](\omega)(\varphi) \cdot \|\varphi\| \;=\; \lambda.$$

2 Prove that for $\lambda' \in \mathbb{R}_{\geq 0}$ and $\omega \in \mathcal{D}(X)$,

$$D_{KL}\Big(Pmn[\lambda](\omega),\, Pmn[\lambda'](\omega')\Big) \;=\; D_{KL}\Big(pois[\lambda],\, pois[\lambda']\Big) + \lambda \cdot D_{KL}(\omega, \omega').$$

4.7 Distributions parameterised by predicates and factors

This chapters introduces observables, including factors and predicates, for probabilistic reasoning. It turns out that they can also be used conveniently as parameters for certain distributions. This the topic of the current section.

The binomial distribution $bn[K](r)$ is parameterised by a bias $r \in [0, 1]$. Similarly, the Poisson distribution $pois[\lambda]$ has a rate $\lambda \geq 0$. The bias can be seen as a predicate $r \colon \mathbf{1} \to [0, 1]$ and the rate as a factor $\lambda \colon \mathbf{1} \to \mathbb{R}_{\geq 0}$, both on the singleton set $\mathbf{1}$. There are various generalisations of the binomial distributions that use multiple biases as parameters. Such multiple biases can

be described jointly as a predicate $p\colon X \to [0,1]$. We shall describe two such distributions with a predicate as parameter, namely the 'Poisson' binomial distribution $pobn[K](p)$ and the 'parallel' binomial distribution $pabn[\psi](p)$. Similarly, the Poisson distribution can be generalised to 'multiple' form, taking a factor $X \to \mathbb{R}_{\geq 0}$ as parameter. In this subsection we briefly discuss these variations.

Aside: another variation is the 'bivariate' binomial $bvbn[K](\gamma)$ for a two-coin distribution $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$, see Definition 4.2.7.

We start with the 'Poisson binomial' distribution. It carries the name 'Poisson' because it was introduced by him. There is no connection to the Poisson distribution $pois[\lambda]$ from Example 2.1.8 (1). For more information about this Poisson binomial, see Chen and Liu (1997); Tang and Tang (2023), for instance as source of the next standard result. It uses the distributions $flip(r) = r|1\rangle + (1-r)|0\rangle$ as distributions on $\mathbb{N}$, so that they can be added via convolution, see Section 2.7. Recall from Proposition 2.7.6 that the binomial distribution $bn[K](r)$ is the K-fold convolution $K \cdot flip(r)$. The next result addresses what happens if we take a convolution sum of flip's with different biases, given by a predicate.

Proposition 4.7.1 *Let $p\colon X \to [0,1]$ be a predicate on a non-empty finite set X, of size $N := |X| > 0$. The convolution sum of the the flip distributions $flip(p(x))$, for $x \in X$, as a distributions on $\{0, 1 \ldots, N\}$ can be described explicitly as:*

$$\sum_{x \in X} flip(p(x)) = \sum_{0 \leq k \leq N} \sum_{U \subseteq X, |U|=k} \prod_{x \in U} p(x) \cdot \prod_{x \in X \setminus U} p^{\perp}(x) \big|k\rangle. \tag{4.25}$$

This distribution is the Poisson binomial distribution, written as $pobn(p) \in \mathcal{D}(\{0, \ldots, K\})$. Clearly it satisfies $pobn(p \circ \pi) = pobn(p)$ for each permutation $\pi\colon X \xrightarrow{\cong} X$.

Notice that if $p(x) = r$ for each $x \in X$, then $pobn(p) = bn[N](r)$.

Proof. We write $pobn(p)$ for the right-hand-side of the equation (4.25). The proof works via induction on $N = |X| \geq 1$. When $N = 1$ the set X is a singleton, say $X = \{*\}$. Then:

$$\begin{aligned} pobn(p) &= \sum_{0 \leq k \leq 1} \sum_{U \subseteq \{*\}, |U|=k} \prod_{x \in U} p(x) \cdot \prod_{x \in \{*\} \setminus U} p^{\perp}(x) \big|k\rangle \\ &= \prod_{x \in \emptyset} p(x) \cdot \prod_{x \in \{*\} \setminus \emptyset} p^{\perp}(x) \big|0\rangle + \prod_{x \in \{*\}} p(x) \cdot \prod_{x \in \{*\} \setminus \{*\}} p^{\perp}(x) \big|1\rangle \\ &= (1 - p(*))|0\rangle + p(*)|1\rangle = flip(p(*)) = \sum_{x \in \{*\}} flip(p(x)). \end{aligned}$$

Next, let $Y = X \cup \{y\}$ where $|X| = N$ and $y \notin X$. We write $p_X\colon X \to [0,1]$ for the restriction of

$p\colon Y \to [0,1]$ to X. Then:

$$\begin{aligned}
pobn(p) &= \sum_{0\le k\le N+1} \sum_{U\subseteq Y, |U|=k} \prod_{x\in U} p(x) \cdot \prod_{x\in Y\setminus U} p^{\perp}(x) \,\big|k\big\rangle \\
&= \sum_{0\le k\le N} \sum_{U\subseteq X, |U|=k} \prod_{x\in U} p(x) \cdot \prod_{x\in Y\setminus U} p^{\perp}(x) \,\big|k\big\rangle \\
&\qquad + \sum_{0\le k\le N} \sum_{U\subseteq X, |U|=k} \prod_{x\in U\cap\{y\}} p(x) \cdot \prod_{x\in Y\setminus(U\cap\{y\})} p^{\perp}(x) \,\big|k+1\big\rangle \\
&= \left(\sum_{0\le k\le N} \sum_{U\subseteq X, |U|=k} \prod_{x\in U} p_X(x) \cdot \prod_{x\in X\setminus U} p_X^{\perp}(x) \,\big|k\big\rangle \right) \cdot p^{\perp}(y) \\
&\qquad + \left(\sum_{0\le k\le N} \sum_{U\subseteq X, |U|=k} \prod_{x\in U} p_X(x) \cdot \prod_{x\in X\setminus U} p_X^{\perp}(x) \,\big|k+1\big\rangle \right) \cdot p(y) \\
&= \mathcal{D}(+)\Big(pobn(p_X) \otimes flip(p(y))\Big) \\
&\overset{\text{(IH)}}{=} \left(\sum_{x\in X} flip(p_X(x)) \right) + flip(p(y)) \;=\; \sum_{x\in Y} flip(p(x)).
\end{aligned}$$

The equation $pobn(p \circ \pi) = pobn(p)$, for a permutation $\pi\colon X \xrightarrow{\cong} X$, follows from the fact that convolution is commutative, see Proposition 2.7.2 (1). ❑

To see an example, for a predicate $p = \frac{3}{4}\cdot \mathbf{1}_a + \frac{2}{3}\cdot \mathbf{1}_b + \frac{1}{2}\cdot \mathbf{1}_c + \frac{1}{3}\cdot \mathbf{1}_d$ one gets:

$$pobn(p) \;=\; \tfrac{1}{36}|0\rangle + \tfrac{13}{72}|1\rangle + \tfrac{7}{18}|2\rangle + \tfrac{23}{72}|3\rangle + \tfrac{1}{12}|4\rangle.$$

We move to a next variation on the binomial, which we call the 'parallel binomial' distribution. We have already seen that the multinomial distribution $mn[K](\omega)$ is a generalisation of the binonmial distribution $bn[K](r)$. Indeed, the single bias probability $r \in [0,1] \cong \mathcal{D}(\mathbf{2})$ can be generalised a distribution $\omega \in \mathcal{D}(X)$ with multiple probabilities. When the set X has two elements, these definitions coincide, up-to-isomorphism, see explicitly in Lemma 2.6.8.

This generalisation is based on the idea that a bias $r \in [0,1]$ corresponds to a (binary) distribution in $\mathcal{D}(\mathbf{2})$. But we can also identify the rate $r \in [0,1]$ with a predicate $\mathbf{1} \to [0,1]$ on a one-element set $\mathbf{1}$. The obvious generalisation of $r \in [0,1]$ to a predicate $p\colon X \to [0,1]$ is used in the above Poisson binomial. This idea also underlies the following definition of what we call a parallel binomial $pabn$.

Definition 4.7.2 Let $\psi \in \mathcal{M}(X)$ be an arbitrary natural multiset. We define the parallel binomial channel $pabn[\psi]\colon Pred(X) \to \mathcal{D}(\mathcal{N}(X))$ on a predicate $p\colon X \to [0,1]$ as:

$$\begin{aligned}
pabn[\psi](p) &:= \sum_{\varphi\le\psi} \binom{\psi}{\varphi} \cdot \prod_{x\in X} p(x)^{\varphi(x)} \cdot (1-p(x))^{\psi(x)-\varphi(x)} \,\big|\varphi\big\rangle \\
&= \sum_{\varphi\le\psi} \prod_{x\in X} \binom{\psi(x)}{\varphi(x)} \cdot p(x)^{\varphi(x)} \cdot (1-p(x))^{\psi(x)-\varphi(x)} \,\big|\varphi\big\rangle \qquad (4.26) \\
&= \sum_{\varphi\le\psi} \prod_{x\in X} bn[\psi(x)]\Big(p(x)\Big)\Big(\varphi(x)\Big) \,\big|\varphi\big\rangle.
\end{aligned}$$

This produces a finite distribution, since there are only finitely many natural sub-multisets $\varphi \le \psi$, for a given natural multiset ψ. In fact, Exercise 1.7.5 gives the number of such sub-multisets.

Example 4.7.3 Consider the multiset $\psi = 3|a\rangle + 1|b\rangle + 1|c\rangle$ with predicate $p = \frac{1}{2}\cdot\mathbf{1}_a + \frac{2}{3}\cdot\mathbf{1}_b + \mathbf{1}_c$. The parallel binomial then takes the form:

$$\begin{aligned} pabn[\psi](p) \;=\; & \tfrac{1}{24}\big|1|c\rangle\big\rangle + \tfrac{1}{8}\big|1|a\rangle + 1|c\rangle\big\rangle + \tfrac{1}{12}\big|1|b\rangle + 1|c\rangle\big\rangle + \tfrac{1}{8}\big|2|a\rangle + 1|c\rangle\big\rangle \\ & + \tfrac{1}{4}\big|1|a\rangle + 1|b\rangle + 1|c\rangle\big\rangle + \tfrac{1}{24}\big|3|a\rangle + 1|c\rangle\big\rangle \\ & + \tfrac{1}{4}\big|2|a\rangle + 1|b\rangle + 1|c\rangle\big\rangle + \tfrac{1}{12}\big|3|a\rangle + 1|b\rangle + 1|c\rangle\big\rangle. \end{aligned}$$

Notice that $1|c\rangle$ occurs in all the multiset in (the support of) this distribution. Indeed, if $\varphi \leq \psi$ satisfies $\varphi(c) = 0$, then the probability associated with φ in (4.26) is zero.

Interestingly, the three ordinary binomial distributions, with sizes and rates given pointwise by ψ and p, are:

$$\begin{aligned} bn[\psi(a)](p(a)) &= \tfrac{1}{8}|0\rangle + \tfrac{3}{8}|1\rangle + \tfrac{3}{8}|2\rangle + \tfrac{1}{8}|3\rangle \\ bn[\psi(b)](p(b)) &= \tfrac{1}{3}|0\rangle + \tfrac{2}{3}|1\rangle \\ bn[\psi(c)](p(c)) &= 1|1\rangle. \end{aligned}$$

Their tensor product is of the form:

$$\begin{aligned} & bn[\psi(a)](p(a)) \otimes bn[\psi(b)](p(b)) \otimes bn[\psi(c)](p(c)) \\ &= \tfrac{1}{24}\big|0,0,1\rangle + \tfrac{1}{8}\big|1,0,1\rangle + \tfrac{1}{12}\big|0,1,1\rangle + \tfrac{1}{8}\big|2,0,1\rangle + \tfrac{1}{4}\big|1,1,1\rangle \\ &\quad + \tfrac{1}{24}\big|3,0,1\rangle + \tfrac{1}{4}\big|2,1,1\rangle + \tfrac{1}{12}\big|3,1,1\rangle. \end{aligned}$$

It contains the multiplicities of the multisets in the above parallel binomial distribution $pabn[\psi](p)$. Thus, the two distributions can be related via the isomorphism $\mathbb{N}^3 \xrightarrow{\cong} \mathcal{N}(\{a,b,c\})$.

In light of this observation it makes sense to define also a parallel version of the Poisson distribution.

Definition 4.7.4 Let X be a finite set, with a factor $\Lambda\colon X \to \mathbb{R}_{\geq 0}$. We define the parallel poisson distribuiton $ppois[\Lambda] \in \mathcal{D}_\infty(\mathcal{N}(X))$ as:

$$ppois[\Lambda] \coloneqq \sum_{\varphi \in \mathcal{N}(X)} \prod_{x \in X} pois[\Lambda(x)](\varphi(x))\,\big|\varphi\rangle. \tag{4.27}$$

This parallel Poisson distribution $ppois[\Lambda]$ may also be identified with a parallel product $\bigotimes_x pois[\Lambda(x)]$, using the identification $\mathcal{N}(X) \cong \mathbb{N}^{|X|}$ that we used in the previous example. With this new definition we can write the Poisson-multinomial $Pmn[\lambda](\omega)$ as parallel Poisson $ppois[\Lambda]$ for $\Lambda(x) = \omega(x)\cdot\lambda$, see Lemma 3.10.3.

Recall from Exercise 3.10.1 the push forward equation $bn[-](r)_*(pois[\lambda]) = pois[r\cdot\lambda]$. There is a parallel version.

Lemma 4.7.5 *Parallel Poisson distributions are closed under pushforward along parallel binomials:*

$$pabn[-](p)_*\big(ppois[\Lambda]\big) = ppois[p \mathbin{\&} \Lambda].$$

Here we write $p \mathbin{\&} \Lambda\colon X \to \mathbb{R}_{\geq 0}$ for the pointwise multiplication: $(p \mathbin{\&} \Lambda)(x) = p(x)\cdot\Lambda(x)$.

Proof. By expanding the definitions (4.27) and (4.26), we get:

$$
\begin{aligned}
& pabn[-](p)_*\Big(ppois[\Lambda]\Big) \\
&= \sum_{\psi\in\mathcal{N}(X)} \sum_{\varphi\leq\psi} ppois[\Lambda](\psi)\cdot pabn[\psi](p)(\varphi)\,\big|\varphi\rangle \\
&= \sum_{\psi\in\mathcal{N}(X)} \sum_{\varphi\leq\psi} \prod_{x\in X} pois[\Lambda(x)](\psi(x))\cdot \prod_{x\in X} bn[\psi(x)](p(x))(\varphi(x))\,\big|\varphi\rangle \\
&= \sum_{\psi\in\mathcal{N}(X)} \sum_{\varphi\leq\psi} \prod_{x\in X} e^{-\Lambda(x)}\cdot \frac{\Lambda(x)^{\psi(x)}}{\psi(x)!}\cdot \binom{\psi(x)}{\varphi(x)}\cdot p(x)^{\varphi(x)}\cdot (1-p(x))^{\psi(x)-\varphi(x)}\,\big|\varphi\rangle \\
&= \sum_{\psi\in\mathcal{N}(X)} \sum_{\varphi\leq\psi} \prod_{x\in X} e^{-p(x)\cdot\Lambda(x)}\cdot \frac{(p(x)\cdot\Lambda(x))^{\varphi(x)}}{\varphi(x)!} \\
&\qquad \cdot e^{-p^{\perp}(x)\cdot\Lambda(x)}\cdot \frac{(p^{\perp}(x)\cdot\Lambda(x)^{\psi(x)-\varphi(x)}}{(\psi(x)-\varphi(x))!}\,\big|\varphi\rangle \\
&= \sum_{\varphi\in\mathcal{N}(X)} \prod_{x\in X} pois[p(x)\cdot\Lambda(x)](\varphi(x)) \\
&\qquad \cdot \sum_{\psi\geq\varphi} pois[p^{\perp}(x)\cdot\Lambda(x)]((\psi-\varphi)(x))\,\big|\varphi\rangle \\
&= \sum_{\varphi\in\mathcal{N}(X)} \prod_{x\in X} pois[(p\ \&\ \Lambda)(x)](\varphi(x))\cdot \sum_{\chi\in\mathcal{N}(X)} pois[(p^{\perp}\ \&\ \Lambda)(x)](\chi(x))\,\big|\varphi\rangle \\
&= \sum_{\varphi\in\mathcal{N}(X)} \prod_{x\in X} pois[(p\ \&\ \Lambda)(x)](\varphi(x))\cdot 1\,\big|\varphi\rangle \\
&= ppois[p\ \&\ \Lambda].
\end{aligned}
$$

❑

Exercises

4.7.1 Show that for a predicate $p\colon X\to[0,1]$ on a finite set X one has:

$$Mean\Big(pobn(p)\Big) = \sum_{x\in X} p(x).$$

Hint: Use Lemma 4.2.4.

4.7.2 Let X be a finite set with a multiset $\psi\in\mathcal{N}(X)$ and a predicate $p\colon X\to[0,1]$.

1 Prove that for each element $y\in X$,

$$\sum_{\varphi\leq\psi} pabn[\psi](p)(\varphi)\cdot\varphi(y) = \psi(y)\cdot p(y).$$

2 Conclude that:

$$Mean\Big(pabn[\psi](p)\Big) = \sum_{y\in X} \psi(y)\cdot p(y)\,\big|y\rangle \in \mathcal{M}(X).$$

4.7.3 Let X be a finite set with a factor $\Lambda\colon X\to\mathbb{R}_{\geq 0}$.

1 Show that for each elements $y\in X$,

$$\sum_{\varphi\in\mathcal{N}(X)} ppois[\Lambda](\varphi)\cdot\varphi(y) =$$

2 Conclude that:

$$Mean\big(ppois[\Lambda]\big) = \sum_{y \in X} \Lambda(y) \,|y\rangle.$$

4.8 Validity-based distances

This section describes two standard distances, between distributions, and between predicates. It shows that these distances can both be formulated in terms of validity $\models$, in a dual form. Basic properties of these distances are included. Earlier, in Section 2.10, we have have seen Kullback-Leibler divergence as a measure of difference between distributions / states. But divergence does not form a metric since it is not symmetric, see Remark 4.8.7 below for an illustration of the difference.

The distances that we focus on in this section are defined as follows. The distance $d(\omega, \omega')$ between two distributions $\omega, \omega' \in \mathcal{D}(X)$, on the same set X, can be defined as the join of the distances in $[0, 1]$, between validities:

$$d(\omega, \omega') := \bigvee_{p \in Pred(X)} \big|\omega \models p \;-\; \omega' \models p\big|. \tag{4.28}$$

Similarly, the distance $d(p, p')$ between two predicates $p, p' \in Pred(X)$ on the same set, is defined as:

$$d(p, p') := \bigvee_{\omega \in \mathcal{D}(X)} \big|\omega \models p \;-\; \omega \models p'\big|. \tag{4.29}$$

Note that the above formulations involve predicates only, not observables in general.

4.8.1 Distance between distributions

The distance defined in (4.28) is commonly called the *total variation* distance, which is a special case of the Wasserstein / Kantorovich distance, see *e.g.* Gibbs and Su (2002); Breugel et al. (2007); McIver et al. (2015); Mardare et al. (2016). Its two alternative characterisations below are standard. We refer to Jacobs and Westerbaan (2020) for more information about the validity-based approach.

Proposition 4.8.1 *Let X be an arbitrary set, with states $\omega, \omega' \in \mathcal{D}(X)$. Then:*

$$d(\omega, \omega') \;=\; \max_{U \subseteq X} \omega \models \mathbf{1}_U \;-\; \omega' \models \mathbf{1}_U \;=\; \tfrac{1}{2} \sum_{x \in X} \big|\omega(x) \;-\; \omega'(x)\big|.$$

We write maximum 'max' instead of join $\bigvee$ to express that the supremum is actually reached by a subset (sharp predicate).

Proof. Let $\omega, \omega' \in \mathcal{D}(X)$ be two discrete probability distributions on the same set X. We will prove the two inequalities labeled (a) and (b) in:

$$\begin{aligned} \tfrac{1}{2} \sum_{x \in X} \big|\omega(x) \;-\; \omega'(x)\big| &\overset{(a)}{\leq} \max_{U \subseteq X} \omega \models \mathbf{1}_U - \omega' \models \mathbf{1}_U \\ &\leq \bigvee_{p \in Pred(X)} \big|\omega \models p - \omega' \models p\big| \overset{(b)}{\leq} \frac{1}{2} \sum_{x \in X} \big|\omega(x) \;-\; \omega'(x)\big|. \end{aligned}$$

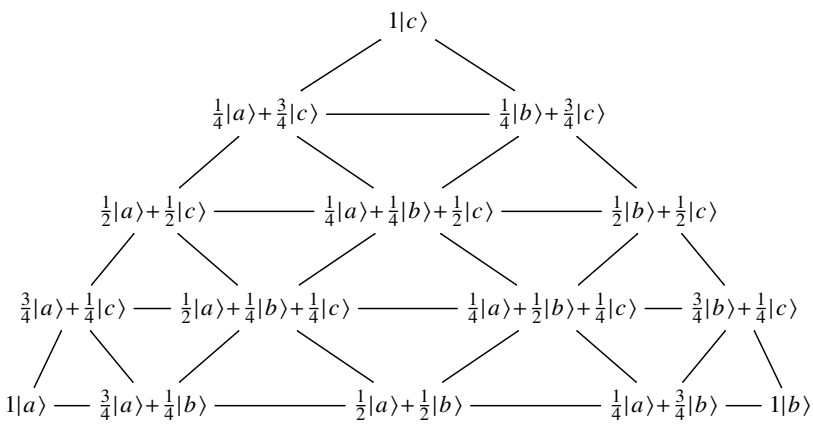


Figure 4.3 Distance graph between distributions, see Example 4.8.3.

This proves Proposition 4.8.1 since the inequality in the middle is trivial.

We start with some preparatory definitions. Let $U \subseteq X$ be an arbitrary subset. We shall write $\omega_i(U) = \sum_{x\in U} \omega_i(x) = (\omega \models \mathbf{1}_U)$. We partition U in three disjoint parts, and take the relevant sums:

$$\begin{cases} U_> = \{x \in U \mid \omega(x) > \omega'(x)\} \\ U_= = \{x \in U \mid \omega(x) = \omega'(x)\} \\ U_< = \{x \in U \mid \omega(x) < \omega'(x)\} \end{cases} \qquad \begin{cases} U\uparrow = \omega(U_>) - \omega'(U_>) \geq 0 \\ U\downarrow = \omega'(U_<) - \omega(U_<) \geq 0. \end{cases}$$

We use this notation in particular for $U = X$. In that case we can use:

$$1 = \omega(X) = \omega(X_>) + \omega(X_=) + \omega(X_<) \qquad 1 = \omega'(X) = \omega'(X_>) + \omega'(X_=) + \omega'(X_<)$$

Hence by subtraction we obtain, since $\omega(X_=) = \omega'(X_=)$,

$$0 = (\omega(X_>) - \omega'(X_>)) + (\omega(X_<) - \omega'(X_<))$$

That is,

$$X\uparrow = \omega(X_>) - \omega'(X_>) = \omega'(X_<) - \omega(X_<) = X\downarrow .$$

As a result:

$$\begin{aligned} \tfrac{1}{2}\sum_{x\in X} \big|\omega(x) - \omega'(x)\big| &= \tfrac{1}{2}\left(\sum_{x\in X_>} (\omega(x) - \omega'(x)) + \sum_{x\in X_<} (\omega'(x) - \omega(x))\right) \\ &= \tfrac{1}{2}\Big((\omega(X_>) - \omega'(X_>)) + (\omega'(X_<) - \omega(X_<))\Big) \\ &= \tfrac{1}{2}(X\uparrow + X\downarrow) \\ &= X\uparrow \end{aligned} \tag{4.30}$$

We have prepared the ground for proving the above inequalities (a) and (b).

(a) We will see that the above maximum is actually reached for the subset $U = X_>$, first of all

because:

$$\frac{1}{2}\sum_{x\in X}\big|\omega(x)-\omega'(x)\big| \overset{(4.30)}{=} X{\uparrow} = \omega(X_>)-\omega'(X_>) = \omega\models \mathbf{1}_{X_>} - \omega'\models \mathbf{1}_{X_>}$$
$$\leq \max_{U\subseteq X}\ \omega\models \mathbf{1}_U - \omega'\models \mathbf{1}_U.$$

(*b*) Let $p\in Pred(X)$ be an arbitrary predicate. We have: $(\mathbf{1}_U \ \&\ p)(x) = \mathbf{1}_U(x)\cdot p(x)$, which is $p(x)$ if $x\in U$ and 0 otherwise. Then:

$$\begin{aligned}
\big|\omega\models p - \omega'\models p\big| &= \Big|(\omega\models \mathbf{1}_{X_>}\&\, p + \omega\models \mathbf{1}_{X_=}\&\, p + \omega\models \mathbf{1}_{X_<}\&\, p) \\
&\qquad - (\omega'\models \mathbf{1}_{X_>}\&\, p + \omega'\models \mathbf{1}_{X_=}\&\, p + \omega'\models \mathbf{1}_{X_<}\&\, p)\Big| \\
&= \Big|(\omega\models \mathbf{1}_{X_>}\&\, p - \omega'\models \mathbf{1}_{X_>}\&\, p) - (\omega'\models \mathbf{1}_{X_<}\&\, p - \omega\models \mathbf{1}_{X_<}\&\, p)\Big| \\
&= \begin{cases} (\omega\models \mathbf{1}_{X_>}\&\, p - \omega'\models \mathbf{1}_{X_>}\&\, p) - (\omega'\models \mathbf{1}_{X_<}\&\, p - \omega\models \mathbf{1}_{X_<}\&\, p) \\ \quad \text{if } \omega\models \mathbf{1}_{X_>}\&\, p - \omega'\models \mathbf{1}_{X_>}\&\, p \overset{(*)}{\geq} \omega'\models \mathbf{1}_{X_<}\&\, p - \omega\models \mathbf{1}_{X_<}\&\, p \\ (\omega'\models \mathbf{1}_{X_<}\&\, p - \omega\models \mathbf{1}_{X_<}\&\, p) - (\omega\models \mathbf{1}_{X_>}\&\, p - \omega'\models \mathbf{1}_{X_>}\&\, p) \\ \quad \text{otherwise} \end{cases} \\
&\leq \begin{cases} \omega\models \mathbf{1}_{X_>}\&\, p - \omega'\models \mathbf{1}_{X_>}\&\, p & \text{if } (*) \\ \omega'\models \mathbf{1}_{X_<}\&\, p - \omega\models \mathbf{1}_{X_<}\&\, p & \text{otherwise} \end{cases} \\
&= \begin{cases} \sum_{x\in X_>}(\omega(x)-\omega'(x))\cdot p(x) & \text{if } (*) \\ \sum_{x\in X_<}(\omega'(x)-\omega(x))\cdot p(x) & \text{otherwise} \end{cases} \\
&\leq \begin{cases} \sum_{x\in X_>}\omega(x)-\omega'(x) & \text{if } (*) \\ \sum_{x\in X_<}\omega'(x)-\omega(x) & \text{otherwise} \end{cases} \\
&= \begin{cases} X{\uparrow} & \text{if } (*) \\ X{\downarrow} = X{\uparrow} & \text{otherwise} \end{cases} \\
&= X{\uparrow} \overset{(4.30)}{=} \tfrac{1}{2}\textstyle\sum_{x\in X}\big|\omega(x)-\omega'(x)\big|.
\end{aligned}$$

❑

With this formulation of the distance we can make an earlier result precise, namely that for large urns hypergeometric and Pólya distributions look like multinomial distributions, for the frequentist learning of the urn, see Proposition 3.6.5.

Proposition 4.8.2 *Let $X = \{x_1,\ldots,x_N\}$ be a finite set with $N\geq 2$ elements.*

1 *When the urn size increases, the (total variation) distance between hypergeometric and multinomial distributions goes to zero, as in:*

$$\lim_{n_1,\ldots n_N\to\infty} d\Big(hg[K]\big(\textstyle\sum_i n_i|x_i\rangle\big),\ mn[K]\big(Flrn(\textstyle\sum_i n_i|x_i\rangle)\big)\Big) = 0.$$

2 *Similarly, for Pólya distributions:*

$$\lim_{n_1,\ldots n_N\to\infty} d\Big(pl[K]\big(\textstyle\sum_i n_i|x_i\rangle\big),\ mn[K]\big(Flrn(\textstyle\sum_i n_i|x_i\rangle)\big)\Big) = 0.$$

Proof. We only prove the first point, since the proof of the second one works similarly. We use

Proposition 3.6.5 (1) in:

$$\begin{aligned}
&\lim_{n_1,\ldots n_N\to\infty} d\Big(hg[K](\textstyle\sum_i n_i|x_i\rangle),\ mn[K](Flrn(\textstyle\sum_i n_i|x_i\rangle))\Big) \\
&= \lim_{n_1,\ldots n_N\to\infty} \tfrac{1}{2} \sum_{\varphi\in\mathcal{N}[K](X)} \Big| hg[K]\big(\textstyle\sum_i n_i|x_i\rangle\big)(\varphi) \;-\; mn[K]\Big(Flrn(\textstyle\sum_i n_i|x_i\rangle)\Big)(\varphi) \Big| \\
&= \tfrac{1}{2} \sum_{\varphi\in\mathcal{N}[K](X)} \lim_{n_1,\ldots n_N\to\infty} hg[K](\textstyle\sum_i n_i|x_i\rangle)(\varphi) \\
&\qquad\qquad \cdot \left| \frac{mn[K](Flrn(\sum_i n_i|x_i\rangle))(\varphi)}{hg[K](\sum_i n_i|x_i\rangle)(\varphi)} \;-\; 1 \right| \\
&\leq \tfrac{1}{2} \sum_{\varphi\in\mathcal{N}[K](X)} \left| \lim_{n_1,\ldots n_N\to\infty} \frac{mn[K](Flrn(\sum_i n_i|x_i\rangle))(\varphi)}{hg[K](\sum_i n_i|x_i\rangle)(\varphi)} \;-\; 1 \right| \\
&= 0.
\end{aligned}$$

❑

We may reformulate these results as 'urn limits' of the form:

$$\lim_{\upsilon\to\infty} d\Big(hg[K](\upsilon),\ mn[K](Flrn(\upsilon))\Big) \;=\; 0 \;=\; \lim_{\upsilon\to\infty} d\Big(pl[K](\upsilon),\ mn[K](Flrn(\upsilon))\Big).$$

The meaning of $\upsilon \to \infty$ for $\upsilon \in \mathcal{N}(X)$ is that all multiplicities $\upsilon(x)$ of the urn go to infinity, for each element x of the finite set X. The equations are called the laws of large urns in Jacobs (2025c). Later on we shall also see a law of large draws, see Theorem 5.6.1.

Example 4.8.3 Consider the set of 'fractional' distributions:

$$\Big\{ Flrn(\varphi) \;\Big|\; \varphi \in \mathcal{N}[4](\{a,b,c\}) \Big\}.$$

Its $\binom{\binom{3}{4}}{} = 15$ elements form a triangle, as described in Figure 4.3, with each edge describing a (total variation) distance of $\frac{1}{4}$.

The sum-formulation in Proposition 4.8.1 is useful in many situations, for instance in order to prove that the above distance function d between states is a metric.

Lemma 4.8.4 *The distance $d(\omega,\omega')$ between states $\omega,\omega' \in \mathcal{D}(X)$ in (4.28) turns the set of distributions $\mathcal{D}(X)$ into a metric space, with $[0,1]$-valued metric.*

Proof. If $d(\omega,\omega') = \frac{1}{2}\sum_{x\in X} |\,\omega(x) - \omega'(x)\,| = 0$, then $|\,\omega(x) - \omega'(x)\,| = 0$ for each $x \in X$, so that $\omega(x) = \omega'(x)$, and thus $\omega = \omega'$. Obviously, $d(\omega,\omega') = d(\omega',\omega)$. The triangle inequality holds for d since it holds for the standard distance on $[0,1]$.

$$\begin{aligned}
d(\omega_1,\omega_3) &= \tfrac{1}{2}\sum_{x\in X} \big|\,\omega_1(x) - \omega_3(x)\,\big| \\
&\leq \tfrac{1}{2}\sum_{x\in X} \big|\,\omega_1(x) - \omega_2(x)\,\big| + \big|\,\omega_2(x) - \omega_3(x)\big| \\
&= \tfrac{1}{2}\sum_{x\in X} \big|\,\omega_1(x) - \omega_2(x)\,\big| + \tfrac{1}{2}\sum_{x\in X} \big|\,\omega_2(x) - \omega_3(x)\,\big| \;=\; d(\omega_1,\omega_2) + d(\omega_2,\omega_3).
\end{aligned}$$

❑

We use this same sum-formulation for the following result. It uses the notion of short (non-expansive) function, which is frequently used as choice of mapping between metric spaces. Explicitly $f\colon (X,d) \to (Y,d)$ is short if $d(f(x),f(x')) \leq d(x,x')$, for all $x,x' \in X$. Such short functions are automatically continuous, since if $d(x_n,x) \to 0$ as n goes to infinity, then also $d(f(x_n),f(x)) \to 0$.

Lemma 4.8.5 *State transformation is short: for a channel $c\colon X \rightsquigarrow Y$ one has:*

$$d(c_*(\omega), c_*(\omega')) \;\leq\; d(\omega, \omega').$$

Proof. Since:

$$\begin{aligned}
d(c_*(\omega), c_*(\omega')) &= \tfrac{1}{2} \sum_{y \in Y} \big| c_*(\omega)(y) - c_*(\omega')(y) \big| \\
&= \tfrac{1}{2} \sum_{y \in Y} \Big| \sum_{x \in X} \omega(x) \cdot c(x)(y) - \sum_{x \in X} \omega'(x) \cdot c(x)(y) \Big| \\
&= \tfrac{1}{2} \sum_{y \in Y} \Big| \sum_{x \in X} c(x)(y) \cdot (\omega(x) - \omega'(x)) \Big| \\
&\leq \tfrac{1}{2} \sum_{y \in Y} \sum_{x \in X} c(x)(y) \cdot \big| \omega(x) - \omega'(x) \big| \\
&= \tfrac{1}{2} \sum_{x \in X} \Big(\sum_{y \in Y} c(x)(y) \Big) \cdot \big| \omega(x) - \omega'(x) \big| \;=\; d(\omega, \omega').
\end{aligned}$$

❑

We recall from Definition 3.1.5 that a *coupling* of two states $\omega, \omega' \in \mathcal{D}(X)$ is a joint state $\tau \in \mathcal{D}(X \times X)$ that marginalises to ω and ω', *i.e.* that satisfies $\tau[1, 0] = \omega$ and $\tau[0, 1] = \omega'$. Such couplings give an alternative formulation of the distance between states, which is commonly called the Wasserstein distance, and can be described in terms of 'optimal transport', from one distribution to another, see Lindvall (1987); Villani (2009) for details. The proof of the next result is standard and is included in order to be complete. It is often called KR-duality, for Kantorovich-Rubenstein Kantorovich and Rubinshtein (1958). For an overview of applications in computer science, see Deng and Du (2009).

Proposition 4.8.6 *For distributions $\omega, \omega' \in \mathcal{D}(X)$,*

$$d(\omega, \omega') \;=\; \bigwedge \{\tau \models Eq^{\perp} \mid \tau \in \mathcal{D}(X \times X) \textit{ is a coupling of } \omega, \omega'\}, \tag{4.31}$$

where $Eq\colon X \times X \to [0, 1]$ is the equality predicate from Definition 4.1.1 (4).

The binary predicate $Eq^{\perp}$ is the discrete distance on the set X, since for $x, x' \in X$ one has:

$$\begin{aligned}
Eq^{\perp}(x, x') \;=\; 1 - Eq(x, x') &= \begin{cases} 1 - 1 & \text{if } x = x' \\ 1 - 0 & \text{if } x \neq x' \end{cases} \\
&= \begin{cases} 0 & \text{if } x = x' \\ 1 & \text{if } x \neq x'. \end{cases}
\end{aligned}$$

The formula in the proposition can be generalised to a a 'lifting' of the distribution functor $\mathcal{D}$ from sets to metric spaces. As such it is used above for a set with the discrete metric $Eq^{\perp}$.

As we have seen in Exercise 2.3.8 there can be infinitely many couplings of two given distributions. This makes computing the infimum in (4.31) a challenge. However, the Wasserstein distance can be computed via (linear) optimisation[2].

Proof. We use the notation and results from the proof of Proposition 4.8.1. We first prove the

[2] This is implemented for instance in Python's ScyPi library.

inequality ($\leq$). So let $\tau \in \mathcal{D}(X \times X)$ be a coupling of $\omega, \omega' \in \mathcal{D}(X)$. Write $X_> = \{x \in X \mid \omega(x) > \omega'(x)\}$. Then:

$$\omega(x) \;=\; \sum_{y \in X} \tau(x,y) \;=\; \tau(x,x) + \sum_{y \neq x} \tau(x,y) \;\leq\; \omega'(x) + (\tau \models (\mathbf{1}_x \otimes \mathbf{1}) \;\&\; Eq^\perp).$$

This means that $\omega(x) - \omega'(x) \leq \tau \models (\mathbf{1}_x \otimes \mathbf{1}) \;\&\; Eq^\perp$ for $x \in X_>$. We similarly have:

$$\omega'(x) \;=\; \sum_{y \in X} \tau(y,x) \;=\; \tau(x,x) + \sum_{y \neq x} \tau(y,x) \;\leq\; \omega(x) + (\tau \models (\mathbf{1} \otimes \mathbf{1}_x) \;\&\; Eq^\perp).$$

Hence $\omega'(x) - \omega(x) \leq \tau \models (\mathbf{1} \otimes \mathbf{1}_x) \;\&\; Eq^\perp$ for $x \notin X_>$. Putting this together gives:

$$\begin{aligned}
d(\omega,\omega') \;&=\; \tfrac{1}{2} \sum_{x \in X} \big|\omega(x) - \omega'(x)\big| \\
&=\; \tfrac{1}{2} \sum_{x \in X_>} \omega(x) - \omega'(x) \;+\; \tfrac{1}{2} \sum_{x \in \neg X_>} \omega'(x) - \omega(x) \\
&\leq\; \tfrac{1}{2} \cdot \sum_{x \in X_>} \tau \models (\mathbf{1}_x \otimes \mathbf{1}) \;\&\; Eq^\perp \;+\; \tfrac{1}{2} \cdot \sum_{x \in \neg X_>} \tau \models (\mathbf{1} \otimes \mathbf{1}_x) \;\&\; Eq^\perp \\
&=\; \tfrac{1}{2} \cdot \Big(\tau \models (\mathbf{1}_{X_>} \otimes \mathbf{1}) \;\&\; Eq^\perp\Big) \;+\; \tfrac{1}{2} \cdot \Big(\tau \models (\mathbf{1} \otimes \mathbf{1}_{\neg X_>}) \;\&\; Eq^\perp\Big) \\
&\leq\; \tfrac{1}{2} \cdot \Big(\tau \models Eq^\perp\Big) \;+\; \tfrac{1}{2} \cdot \Big(\tau \models Eq^\perp\Big) \;=\; \tau \models Eq^\perp.
\end{aligned}$$

For the inequality ($\geq$) one uses what is called an *optimal coupling* $\rho \in \mathcal{D}(X \times X)$ of ω, ω'. It can be defined as:

$$\rho(x,y) \;:=\; \begin{cases} \min\big(\omega(x), \omega'(x)\big) & \text{if } x = y \\ \dfrac{(\omega(x) \dot{-} \omega'(x)) \cdot (\omega'(y) \dot{-} \omega(y))}{d(\omega,\omega')} & \text{otherwise,} \end{cases} \tag{4.32}$$

where $a \dot{-} b := \max(a - b, 0)$, for $a, b \in \mathbb{R}$. This formula occurs for instance in (Villani, 2009, proof of Thm. 6.15) (for the continuous case).

We first check that this ρ is a coupling. Let $x \in X_>$ so that $\omega(x) > \omega'(x)$; then:

$$\begin{aligned}
\sum_{y \in X} \rho(x,y) \;&=\; \omega'(x) + (\omega(x) - \omega'(x)) \cdot \sum_{y \neq x} \frac{\omega'(y) \dot{-} \omega(y)}{d(\omega,\omega')} \\
&=\; \omega'(x) + (\omega(x) - \omega'(x)) \cdot \frac{\sum_{y \in X_<} \omega'(y) - \omega(y)}{d(\omega,\omega')} \\
&=\; \omega'(x) + (\omega(x) - \omega'(x)) \cdot \frac{X{\downarrow}}{d(\omega,\omega')} \\
&=\; \omega'(x) + (\omega(x) - \omega'(x)) \cdot 1 \quad \text{see the proof of Proposition 4.8.1} \\
&=\; \omega(x).
\end{aligned}$$

If $x \notin X_>$, so that $\omega(x) \leq \omega'(x)$, then it is obvious that $\sum_y \rho(x,y) = \omega(x) + 0 = \omega(x)$. This shows $\rho[1,0] = \omega$. In a similar way one obtains $\rho[0,1] = \omega'$. Finally,

$$\begin{aligned}
\rho \models Eq \;=\; \sum_{x \in X} \rho(x,x) \;=\; \sum_{x \in X} \min\big(\omega(x), \omega'(x)\big) \;&=\; \sum_{x \in X_>} \omega'(x) \;+\; \sum_{x \notin X_>} \omega(x) \\
&=\; \omega'(X_>) + 1 - \omega(X_>) \\
&=\; 1 - \big(\omega(X_>) - \omega'(X_>)\big) \;=\; 1 - d(\omega,\omega').
\end{aligned}$$

Hence $d(\omega,\omega') = 1 - (\rho \models Eq) = \rho \models Eq^\perp$. ❑

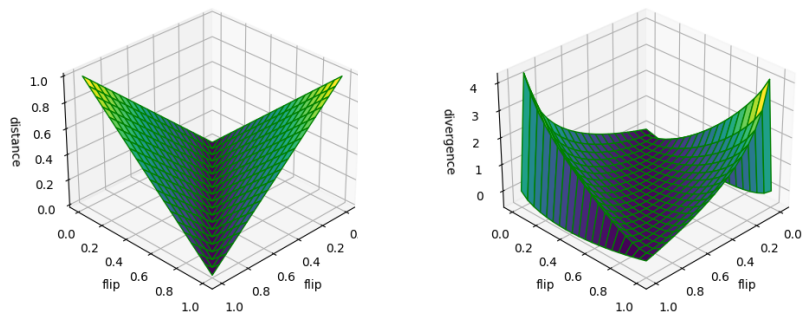


Figure 4.4 Visual comparance of total variation distance and divergence between flip distributions, see Remark 4.8.7 for details.

Now that we have a good understanding of the total variation distance on distributions, there are a couple of comparisons to make.

Remark 4.8.7 In Definition 2.10.1 we have seen Kullback-Leibler divergence D_{KL}, as a measure of difference between distributions. However, this D_{KL} is not a proper metric, since it is not symmetric, see Exercise 2.10.1. Nevertheless, it is often used as distance between distributions, especially in minimisation problems (see Theorem 3.4.2).

Figure 4.4 compares the total variation distance and the Kullback-Leibler divergence between two flip states:

$$d\Big(\mathit{flip}(r), \mathit{flip}(s)\Big) \qquad\qquad D_{KL}\Big(\mathit{flip}(r), \mathit{flip}(s)\Big).$$

for $r, s \in [0,1]$, where, recall, $\mathit{flip}(r) = r|1\rangle + (1-r)|0\rangle$. The distance and divergence are zero when $r = s$ and increases on both sides of the diagonal. The distance ascends via straight planes, but the divergence has a more baroque shape.

There is in fact a clear relationship between total variation distance and Kullback-Leibler divergence, known as Pinsker's inequality, see item (2) below. The proof below follows Canonne (2023).

Theorem 4.8.8 *1 For numbers $r, s \in [0,1]$,*

$$2(r-s)^2 \;=\; 2 \cdot d\Big(\mathit{flip}(r),\ \mathit{flip}(s)\Big)^2 \;\leq\; D_{KL}\Big(\mathit{flip}(r),\ \mathit{flip}(s)\Big).$$

2 For distributions $\omega, \rho \in \mathcal{D}(X)$,

$$2 \cdot d(\omega, \rho)^2 \;\leq\; D_{KL}(\omega, \rho).$$

Proof. 1 Without loss of generality, we assume $r \geq s$. Clearly,

$$d\Big(\mathit{flip}(r),\ \mathit{flip}(s)\Big) \;=\; \tfrac{1}{2} \cdot \Big(\big|\, r - s \,\big| + \big|\, (1-r) - (1-s) \,\big|\Big) \;=\; \big|\, r - s \,\big|.$$

We have to prove the inequality:

$$2(r-s)^2 \;\leq\; D_{KL}\Big(\mathit{flip}(r),\ \mathit{flip}(s)\Big) \;=\; r \cdot \ln\big(\tfrac{r}{s}\big) + (1-r) \cdot \ln\big(\tfrac{1-r}{1-s}\big) \;=\; g(r) - g(s),$$

for $g(x) = r \cdot \ln(x) + (1-r) \cdot \ln(1-x)$. The latter difference can be computed via an integral, as:

$$\begin{aligned} g(r) - f(s) &= \int_s^r f'(x)\,\mathrm{d}x = \int_s^r \frac{r}{x} - \frac{1-r}{1-x}\,\mathrm{d}x \\ &= \int_s^r \frac{r(1-x)-(1-r)x}{x(1-x)}\,\mathrm{d}x \\ &\geq 4 \cdot \int_s^r r - x\,\mathrm{d}x \qquad \text{since } x(1-x) \leq \tfrac{1}{4} \text{ for } x \in [0,1] \\ &= 4 \cdot \left[\, rx - \tfrac{1}{2}x^2 \,\right]_s^r \\ &= 4(r^2 - \tfrac{1}{2}r^2 - rs + \tfrac{1}{2}s^2) \;=\; 2(r^2 - 2rs + s^2) \;=\; 2(r-s)^2. \end{aligned}$$

2 We use the formulation for total variation distance in the middle of Proposition 4.8.1. So let $U \subseteq X$ be an arbitrary subset. We use it do define a flip channel $f_U \colon X \to \mathcal{D}(\{H,T\})$, namely:

$$f_U(x) \coloneqq \begin{cases} 1|H\rangle & \text{if } x \in U \\ 1|T\rangle & \text{if } x \notin U. \end{cases}$$

Then, for a distribution $\omega \in \mathcal{D}(X)$,

$$\begin{aligned} f_{U*}(\omega)(H) &= \sum_{x \in X} \omega(x) \cdot f_U(x)(H) \;=\; \sum_{x \in U} \omega(x) \;=\; \omega \models \mathbf{1}_U \\ f_{U*}(\omega)(T) &= \sum_{x \in X} \omega(x) \cdot f_U(x)(T) \;=\; \sum_{x \in \neg U} \omega(x) \;=\; 1 - (\omega \models \mathbf{1}_U). \end{aligned}$$

Thus: $f_{U*}(\omega) = \mathit{flip}(\omega \models \mathbf{1}_U)$. Now we get, by the previous item:

$$\begin{aligned} \left|\, \omega \models \mathbf{1}_U \;-\; \rho \models \mathbf{1}_U \,\right| &= \sqrt{(\,\omega \models \mathbf{1}_U \;-\; \rho \models \mathbf{1}_U\,)^2} \\ &= \sqrt{d\big(\mathit{flip}(\omega \models \mathbf{1}_U), \mathit{flip}(\rho \models \mathbf{1}_U)\big)^2} \\ &\leq \sqrt{\tfrac{1}{2} \cdot D_{KL}\big(\mathit{flip}(\omega \models \mathbf{1}_U), \mathit{flip}(\rho \models \mathbf{1}_U)\big)} \\ &= \sqrt{\tfrac{1}{2} \cdot D_{KL}\big(f_{U*}(\omega), f_{U*}(\rho)\big)} \\ &\leq \sqrt{\tfrac{1}{2} \cdot D_{KL}(\omega,\rho)} \qquad \text{by Proposition 2.10.6 (3).} \end{aligned}$$

Since this holds for all subsets $U \subseteq X$ we get, as in Proposition 4.8.1,

$$d(\omega,\rho) = \max_{U \subseteq X} \omega \models \mathbf{1}_U \;-\; \rho \models \mathbf{1}_U \;\leq\; \sqrt{\tfrac{1}{2} \cdot D_{KL}(\omega,\rho)}.$$

This yields: $2 \cdot d(\omega,\rho)^2 \leq D_{KL}(\omega,\rho)$, as required. ❑

Remark 4.8.9 Let X be a finite set, say with N elements. We can view the set $\mathcal{D}(X)$ of distributions on X as a subset of the N-dimensional cube $[0,1]^N$. The total variation distance can be seen as coming from the norm on $[0,1]^N$ given by:

$$\|\, x \,\|_{tv} \;\coloneqq\; \tfrac{1}{2} \cdot \textstyle\sum_i x_i, \qquad \text{for } x \in [0,1]^N.$$

However, on n-tuples of real numbers the Euclidean norm is common, defined as:

$$\|\, x \,\|_{eu} \;\coloneqq\; \sqrt{\textstyle\sum_i x_i^2}, \qquad \text{here for } x \in [0,1]^N.$$

Is this difference relevant?

It is not, from a topological point of view. Both norms induce the same topology on $[0,1]^N$. This follows from a general result about norms, see *e.g.* (Conway, 1990, III, Prop. 1.5). For this it suffices to show that the norms are related via constants. In our case we have:

$$\tfrac{2}{N} \cdot \| x \|_{tv} \leq \| x \|_{eu} \leq 2\sqrt{N} \cdot \| x \|_{tv}$$

For the first inequality we use that $x_i = \sqrt{x_i^2} \leq \| x \|_{eu}$, so $\| x \|_{tv} = \frac{1}{2} \cdot \sum_i x_i \leq \frac{N}{2} \cdot \| x \|_{eu}$. For the second inequality we abbreviate $s = \| x \|_{tv}$ and observe that $x_i \leq 2s$. Hence $x_i^2 \leq 4s^2$ and thus $\sum_i x_i^2 \leq 4N \cdot s^2$. But then: $\| x \|_{eu} = \sqrt{\sum_i x_i^2} \leq 2\sqrt{N} \cdot s$.

We conclude that from a topological perspective it does not matter if we use distributions (on a finite set) with the total variation distance or with the Euclidean distance, induced by the underlying cube of unit intervals. This also means that the induced Borel measures are the same. This will be relevant later on, in Section **??**, when we consider the Dirichlet probability measure on $\mathcal{D}(X)$.

The relation between distributions and multisets is a recurring theme. Now that we have a distance function on distributions we can speak about approximation of a distribution via a chain of multisets. This is what the next remark is about. This topic returns as a law of large draws in Section 5.6, see especially Theorem 5.6.1. Here we take a an algorithmic perspective.

Remark 4.8.10 Let a distribution $\omega \in \mathcal{D}(X)$ be given. One can ask: is there a sequence of natural multisets $\varphi_K \in \mathcal{N}[K](X)$ with $Flrn(\varphi_K)$ getting closer and closer to ω, in the total variation distance d, as K goes to infinity?

The answer is yes. Here is one way to do it. Assume the distribution ω has support $\{x_1, \ldots, x_N\}$, of course with $N > 0$. Below we use $k \cdot \omega \in \mathcal{M}(\{x_1, \ldots, x_n\})$, for $k \geq 0$.

- Pick $\varphi_1 = 1|x_j\rangle$ where ω takes a maximum at x_j, *i.e.*, $\omega(x_j) \geq \omega(x_i)$ for all i. Set variable $cp := j$, for 'current position'.
- Look for the first position i after cp where $\varphi_K(x_i) < (n+1) \cdot \omega(x_i)$. Then you set $\varphi_{K+1} := \varphi_K + 1|x_i\rangle$, and $cp := i$. This search wraps around, if needed. When no i is found we are done and have $Flrn(\varphi_K) = \omega$.

Concretely, if $\omega = \frac{1}{6}|x_1\rangle + \frac{1}{2}|x_2\rangle + \frac{1}{3}|x_3\rangle$. Then, consecutively,

- $0 \cdot \omega = \mathbf{0}$ and $\varphi_1 = 1|x_2\rangle$
- $1 \cdot \omega = \frac{1}{6}|x_1\rangle + \frac{1}{2}|x_2\rangle + \frac{1}{3}|x_3\rangle$ and $\varphi_2 = 1|x_2\rangle + 1|x_3\rangle$
- $2 \cdot \omega = \frac{1}{3}|x_1\rangle + 1|x_2\rangle + \frac{2}{3}|x_3\rangle$ and $\varphi_3 = 1|x_1\rangle + 1|x_2\rangle + 1|x_3\rangle$
- $3 \cdot \omega = \frac{1}{2}|x_1\rangle + \frac{3}{2}|x_2\rangle + 1|x_3\rangle$ and $\varphi_4 = 1|x_1\rangle + 2|x_2\rangle + 1|x_3\rangle$
- $4 \cdot \omega = \frac{2}{3}|x_1\rangle + 2|x_2\rangle + \frac{4}{3}|x_3\rangle$ and $\varphi_5 = 1|x_1\rangle + 2|x_2\rangle + 2|x_3\rangle$
- $5 \cdot \omega = \frac{5}{6}|x_1\rangle + \frac{5}{2}|x_2\rangle + \frac{5}{3}|x_3\rangle$ and $\varphi_6 = 1|x_1\rangle + 3|x_2\rangle + 2|x_3\rangle$
- $6 \cdot \omega = 1|x_1\rangle + 3|x_2\rangle + 2|x_3\rangle$, giving rise to a halt, since $\varphi_6 = 6 \cdot \omega$ and thus $Flrn(\varphi_6) = \omega$.

In this case we get a finite sequence of multisets approaching ω. The sequence is infinite for, *e.g.*,

$$\omega = \tfrac{1}{7}|x_1\rangle + \tfrac{4}{7}|x_2\rangle + \tfrac{2}{7}|x_3\rangle \approx 0.1429|x_1\rangle + 0.5714|x_2\rangle + 0.2857|x_3\rangle.$$

Running the above algorithm gives, for instance:

- $\varphi_{10} = 2|x_1\rangle + 5|x_2\rangle + 3|x_3\rangle$
- $\varphi_{100} = 15|x_1\rangle + 56|x_2\rangle + 29|x_3\rangle$
- $\varphi_{1000} = 143|x_1\rangle + 571|x_2\rangle + 286|x_3\rangle$
- $\varphi_{10000} = 1429|x_1\rangle + 5713|x_2\rangle + 2858|x_3\rangle$.

The next result summarises some basic topological properties of metric spaces of distributions. In short, when X is finite, then $\mathcal{D}(X)$ is a compact Polish space: it is complete and has a countable dense subset, given by fractional distributions, see Exercise 2.4.4. Like any metric space, $\mathcal{D}(X)$ is Hausdorff: two distinct distributions are contained in two disjoint open balls. Additionally, $\mathcal{D}(X)$ is convex, so $\mathcal{D}(X)$ is a convex compact space, as studied for instance in Jacobs et al. (2016).

Theorem 4.8.11 *1 For an arbitrary set X, the set $\mathcal{D}(X)$ has a dense subset $\mathcal{FD}(X) \subseteq \mathcal{D}(X)$ of fractional distributions:*

$$\begin{aligned} \mathcal{FD}(X) &:= \bigcup_{K\in\mathbb{N}} \mathcal{D}[K](X) \;\subseteq\; \mathcal{D}(X) \qquad \textit{where} \\ \mathcal{D}[K](X) &:= \{\mathit{Flrn}(\varphi) \mid \varphi \in \mathcal{N}[K](X)\}. \end{aligned}$$

2 If X is a finite set, then $\mathcal{D}(X)$, with the total variation distance d, is a complete metric space.

3 When X is finite, then $\mathcal{D}(X)$ is a Polish space*: a complete metric space with a countable dense subset, namely $\mathcal{FD}(X) \subseteq \mathcal{D}(X)$.*

4 When X is finite, the space $\mathcal{D}(X)$ is totally bounded, and thus compact.

The finiteness requirement in item (2) is necessary, see Example 4.8.13 below.

Proof. 1 Let $\omega \in \mathcal{D}(X)$ and $\varepsilon > 0$. We need to find a multiset $\varphi \in \mathcal{N}(X)$ with $d(\omega, \mathit{Flrn}(\varphi)) < \varepsilon$. There is a systematic way to find such multisets via the decimal representation of the probabilities in ω. This works as follows. Assume we have:

$$\omega = 0.383914217\ldots|a\rangle + 0.406475610\ldots|b\rangle + 0.209610173\ldots|c\rangle.$$

For each n we chop off after n decimals and multiply with 10^n, giving:

$$\begin{aligned} \varphi_1 &:= 3|a\rangle + 4|b\rangle + 2|c\rangle && \text{with} \quad d(\omega, \mathit{Flrn}(\varphi_1)) \leq \tfrac{1}{2}\cdot 3\cdot 10^{-1} \\ \varphi_2 &:= 38|a\rangle + 40|b\rangle + 20|c\rangle && \text{with} \quad d(\omega, \mathit{Flrn}(\varphi_2)) \leq \tfrac{1}{2}\cdot 3\cdot 10^{-2} \\ \varphi_3 &:= 383|a\rangle + 406|b\rangle + 209|c\rangle && \text{with} \quad d(\omega, \mathit{Flrn}(\varphi_3)) \leq \tfrac{1}{2}\cdot 3\cdot 10^{-3} \qquad \textit{etc.} \end{aligned}$$

In general, for a distribution ω with $\mathit{supp}(\omega) = \{x_1, \ldots, x_M\}$ we can thus construct a sequence of multisets $\varphi_n \in \mathcal{N}(\{x_1, \ldots, x_M\})$ with $d(\omega, \mathit{Flrn}(\varphi_n)) \leq \frac{1}{2}\cdot M\cdot 10^{-n}$. This distance becomes less than any given $\varepsilon > 0$, by choosing n sufficiently large. This shows that the fractional distributions are dense in $\mathcal{D}(X)$.

2 Let X have M elements, say $X = \{x_1, \ldots, x_M\}$ and $\omega_i \in \mathcal{D}(X)$ be a Cauchy sequence. Fix n. Then for all i, j,

$$\big|\omega_i(x_n) - \omega_j(x_n)\big| \leq 2\cdot d(\omega_i, \omega_j).$$

Hence, the sequence $\omega_i(x_n) \in [0,1]$ is a Cauchy sequence, say with limit $r_n \in [0,1]$. Take $\omega = \sum_n r_n|x_n\rangle \in \mathcal{D}(X)$. This is the limit of the distributions ω_i.

3 When X is finite, say with M elements, we know from Proposition 1.8.7 that $\mathcal{N}[K](X)$ contains $\left(\!\binom{M}{K}\!\right)$ multisets, so that $\mathcal{D}[K](X)$ contains $\left(\!\binom{M}{K}\!\right)$ distributions. Hence a countable union $\bigcup_K \mathcal{D}[K](X)$ of such finite sets is countable.

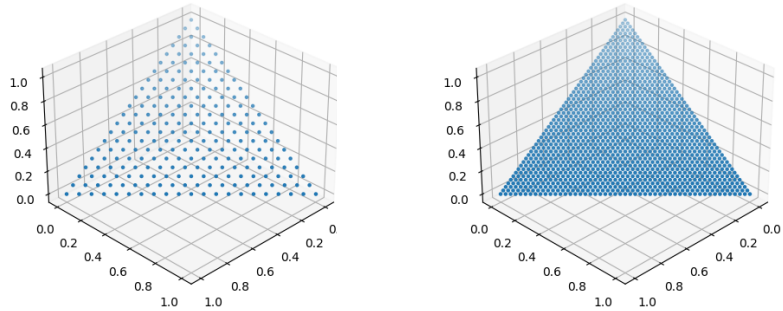


Figure 4.5 Fractional distributions from $\mathcal{D}[K](\{a, b, c\})$ as points in the cube $[0, 1]^3$, for $K = 20$ on the left and $K = 50$ on the right. The plot on the left contains $\left(\!\binom{3}{20}\!\right) = 231$ dots (distributions) and the one on the right $\left(\!\binom{3}{50}\!\right) = 1326$, see Proposition 1.8.7.

4 We use the standard result that a metric space is compact if and only if it is complete and totally bounded. Hence, by item (2), it suffices to show that $\mathcal{D}(X)$ is totally bounded. Suppose X has M elements. Let $\varepsilon > 0$ be given. We need to find a finite number of ε-balls $B_\varepsilon(\sigma)$ whose union contains $\mathcal{D}(X)$. Take K so that $\frac{1}{2} \cdot M \cdot 10^{-K} < \varepsilon$. The above argument for item (1) shows that for each $\omega \in \mathcal{D}(X)$ there is then an $\varphi \in \mathcal{N}[K](X)$ with $d(\omega, Flrn(\varphi)) < \varepsilon$. This shows:

$$\mathcal{D}(X) \subseteq \bigcup_{\varphi \in \mathcal{N}[K](X)} \{\omega \in \mathcal{D}(X) \mid d(\omega, Flrn(\varphi)) < \varepsilon\} = \bigcup_{\varphi \in \mathcal{N}[K](X)} B_\varepsilon(Flrn(\varphi)). \qquad \square$$

Notice that the sequence of multisets φ_K approaching ω described in Remark 4.8.10 has the special property that $\|\varphi_K\| = K$. This does not hold for the sequence φ_n in the above proof. Such a size property is not needed there.

The denseness of the fractional distributions is illustrated in Figure 4.5. A different way to approximate a distribution ω is described in Section 5.6, via what is called the law of large numbers.

The denseness in Theorem 4.8.11 means that every distribution can be approximated by normalisations of multisets — as illustrated in Remark 4.8.10. The entropy of the distribution can then also be approximated via these multisets, in a somewhat unexpected manner, namely via their (multiset) coefficients. This occurs for instance in (Jaynes, 2003, Eqn. (11.29)). This uses the *Stirling approximation* of the factorial: for a large number n one has:

$$\ln(n!) \approx n \cdot ln(n) - n \qquad \text{or, equivalently,} \qquad n! \approx \left(\frac{n}{e}\right)^n. \tag{4.33}$$

Proposition 4.8.12 *Let $\varphi_n \in \mathcal{N}(X)$ be an infinite sequence of natural multisets on a finite set X, with $\lim_{n\to\infty} Flrn(\varphi_n) = \omega \in \mathcal{D}(X)$. Then:*

$$H(\omega) = \lim_{n\to\infty} \frac{\ln(\!(\varphi_n)\!)}{\|\varphi_n\|}.$$

Proof. Write $X = \{x_1, \ldots, x_L\}$. Via Stirling's approximation (4.33) we get:

$$\begin{aligned}
\frac{\ln(\mathbf{(}\varphi_n\mathbf{)})}{\|\varphi_n\|} &= \frac{1}{\|\varphi_n\|} \cdot \ln\left(\frac{\|\varphi_n\|!}{\prod_i \varphi_n(x_i)!}\right) \\
&= \frac{1}{\|\varphi_n\|} \cdot \left(\ln(\|\varphi_n\|!) - \sum_i \ln(\varphi_n(x_i)!)\right) \\
&\approx \frac{1}{\|\varphi_n\|} \cdot \left(\|\varphi_n\| \cdot \ln(\|\varphi_n\|) - \|\varphi_n\| - \sum_i \Big(\varphi_n(x_i) \cdot \ln(\varphi_n(x_i)) - \varphi_n(x_i)\Big)\right) \\
&= \ln(\|\varphi_n\|) - 1 - \sum_i \frac{\varphi_n(x_i)}{\|\varphi_n\|} \cdot \ln(\varphi_n(x_i)) + \frac{1}{\|\varphi_n\|} \cdot \sum_i \varphi_n(x_i) \\
&= -\sum_i \frac{\varphi_n(x_i)}{\|\varphi_n\|} \cdot \ln\left(\frac{\varphi_n(x_i)}{\|\varphi_n\|}\right) \\
&= H(Flrn(\varphi_n)) \longrightarrow H(\omega).
\end{aligned}$$

❑

Example 4.8.13 The finiteness requirement in Theorem 4.8.11 (2) is necessary for completeness. Consider the distributions $\omega_n \in \mathcal{D}(\mathbb{N})$, for $n > 0$, given by:

$$\begin{aligned}
\omega_1 &= \tfrac{1}{2}|0\rangle + \tfrac{1}{2}|1\rangle \\
\omega_2 &= \tfrac{1}{2}|0\rangle + \tfrac{1}{4}|1\rangle + \tfrac{1}{4}|2\rangle \\
\omega_3 &= \tfrac{1}{2}|0\rangle + \tfrac{1}{4}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|3\rangle \\
\omega_4 &= \tfrac{1}{2}|0\rangle + \tfrac{1}{4}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{16}|3\rangle + \tfrac{1}{16}|4\rangle, \qquad \textit{etc.}
\end{aligned}$$

The general formula is:

$$\omega_n := \sum_{0 \le i < n} \frac{1}{2^{i+1}} \big|i\rangle + \frac{1}{2^n} \big|n\rangle.$$

Recall, *e.g.* from the proof of Theorem 1.7.4 (2)

$$(1 - r) \cdot \left(1 + r + r^2 + \cdots + r^{n-1}\right) = 1 - r^n.$$

When we apply this to $r = \frac{1}{2}$ we get:

$$\tfrac{1}{2} + \tfrac{1}{2^2} + \tfrac{1}{2^3} + \cdots + \tfrac{1}{2^n} + \tfrac{1}{2^n} = 1.$$

Thus:

$$\begin{aligned}
d(\omega_n, \omega_{n+1}) &= \left|\tfrac{1}{2^{n-1}} - \tfrac{1}{2^n}\right| + \left|0 - \tfrac{1}{2^n}\right| = \tfrac{1}{2^n} + \tfrac{1}{2^n} = \tfrac{1}{2^{n-1}} \\
d(\omega_n, \omega_{n+2}) &= \left|\tfrac{1}{2^{n-1}} - \tfrac{1}{2^n}\right| + \left|0 - \tfrac{1}{2^{n+1}}\right| + \left|0 - \tfrac{1}{2^{n+1}}\right| = \tfrac{1}{2^n} + 2 \cdot \tfrac{1}{2^{n+1}} = \tfrac{1}{2^{n-1}}.
\end{aligned}$$

Clearly, $d(\omega_n, \omega_{n+m}) = \frac{1}{2^{n-1}}$ for all $m > 0$. This makes these (ω_n) a Cauchy sequence. This sequence does not have a limit in $\mathcal{D}(\mathbb{N})$, because the support of distributions in $\mathcal{D}(\mathbb{N})$ is finite.

As an aside, the Cauchy sequence (ω_n) does have a limit in $\mathcal{D}_\infty(\mathbb{N})$. This limit is $\sum_{n \ge 0} \frac{1}{2^{n+1}}|n\rangle$.

4.8.2 Distance between predicates

What we have to say about the validity-based distance (4.29) between predicates is rather brief. First, there is also a pointwise formulation.

Lemma 4.8.14 *For two predicates $p, p' \in \mathrm{Pred}(X)$,*

$$d(p, p') = \bigvee_{x \in X} \big| p(x) - p'(x) \big|.$$

This distance function d makes the set $\mathrm{Pred}(X)$ into a metric space.

Proof. First, we have for each $x \in X$,

$$\begin{aligned} d(p, p') &\overset{(4.29)}{=} \bigvee_{\omega \in \mathcal{D}(X)} \big| \omega \models p - \omega \models p' \big| \\ &\geq \big| \mathit{unit}(x) \models p - \mathit{unit}(x) \models p' \big| = \big| p(x) - p'(x) \big|. \end{aligned}$$

Hence $d(p, p') \geq \bigvee_x \big| p(x) - p'(x) \big|$.

The other direction follows from:

$$\begin{aligned} \big| \omega \models p - \omega \models p' \big| &= \left| \sum_{z \in X} \omega(z) \cdot p(z) - \sum_{z \in X} \omega(z) \cdot p'(z) \right| \\ &\leq \sum_{z \in X} \omega(z) \cdot \big| p(z) - p'(z) \big| \\ &\leq \sum_{z \in X} \omega(z) \cdot \bigvee_{x \in X} \big| p(x) - p'(x) \big|. \\ &= \left(\sum_{z \in X} \omega(z) \right) \cdot \bigvee_{x \in X} \big| p(x) - p'(x) \big| = \bigvee_{x \in X} \big| p(x) - p'(x) \big|. \end{aligned}$$

The fact that we get a metric space is now straightforward. ❑

There is an analogue of Lemma 4.8.5.

Lemma 4.8.15 *Predicate transformation is also short: for a channel $c \colon X \rightsquigarrow Y$ one has, for predicates $p, p' \in \mathrm{Pred}(Y)$,*

$$d(c^*(p), c^*(p')) \leq d(p, p').$$

Proof. Via the formulation of Lemma 4.8.14 we get:

$$\begin{aligned} d(c^*(p), c^*(p')) &= \bigvee_{x \in X} \big| c^*(p)(x) - c^*(p')(x) \big| \\ &= \bigvee_{x \in X} \left| \sum_{y \in Y} c(x)(y) \cdot p(y) - \sum_{y \in Y} c(x)(y) \cdot p'(y) \right| \\ &\leq \bigvee_{x \in X} \sum_{y \in Y} c(x)(y) \cdot \big| p(y) - p'(y) \big| \\ &\leq \bigvee_{x \in X} \left(\sum_{y \in Y} c(x)(y) \right) \cdot d(p, p') = d(p, p'). \end{aligned}$$

❑

Exercises

4.8.1 Recall from Fact 2.3.10 that for an arbitrary distribution $\omega \in \mathcal{D}(X)$ the joint distributions $\mathcal{D}(\Delta)(\omega)$ and $\omega \otimes \omega$ in $\mathcal{D}(X \times X)$ are different, in general. Show that the (total variation)

distance between them is:

$$d\Big(\mathcal{D}(\Delta)(\omega),\ \omega \otimes \omega\Big) \;=\; \tfrac{1}{2} - \tfrac{1}{2}\sum_{x \in X} \omega(x)^2.$$

The latter expression is called the Gini-Simpson index, see Leinster (2021); Perrone (2024a), and is used in ecology to quantify diversity. Note the analogy with Definition 2.10.7 (2).

4.8.2 Prove that the total variation distance satisfies, for distributions $\omega_1, \omega_2 \in \mathcal{D}(X)$ and a drawsize number $K \in \mathbb{N}$,

$$d(\omega_1,\ \omega_2) \;\leq\; d(mn[K](\omega_1),\ mn[K](\omega_2)).$$

4.8.3 1 Prove that for states $\omega, \omega' \in \mathcal{D}(X)$ and $\rho, \rho' \in \mathcal{D}(Y)$ there is an inequality:

$$d(\omega \otimes \rho, \omega' \otimes \rho') \;\leq\; d(\omega, \omega') + d(\rho, \rho').$$

(In this situation there is an actual equality for Kullback-Leibler divergence, see Proposition 2.10.2 (1).)

2 Prove similarly that for predicates $p, p' \in \mathit{Pred}(X)$ and $q, q' \in \mathit{Pred}(Y)$ one gets:

$$d(p \otimes q, p' \otimes q') \;\leq\; d(p, p') + d(q, q').$$

4.8.4 Let $\varphi, \psi \in \mathcal{N}[K](X)$ be two different natural multisets of the same size K. Check that the distance between the corresponding fractional distrbutions is at least $\frac{1}{K}$ in:

$$d\Big(\mathit{Flrn}(\varphi),\ \mathit{Flrn}(\psi)\Big) \;\geq\; \tfrac{1}{K}.$$

(See also Figure 4.3.)

4.8.5 1 Show that for a state $\omega \in \mathcal{D}(X)$, the 'validity in ω' function (predicate):

$$\mathit{Pred}(X) \xrightarrow{\omega \models (-)} [0, 1]$$

is short.

2 Similarly, show that for an observation $p \in \mathit{Obs}(X)$ the "validity of p" function is short:

$$\mathcal{D}(X) \xrightarrow{(-) \models p} \mathbb{R}$$

4.8.6 In the context of Remark 4.8.7, check that:

$$\begin{aligned} d\Big(\mathit{flip}(0), \mathit{flip}(1)\Big) \;&=\; 1 \;=\; d\Big(\mathit{flip}(1), \mathit{flip}(0)\Big) \\ D_{KL}\Big(\mathit{flip}(0), \mathit{flip}(1)\Big) \;&=\; 0 \;=\; D_{KL}\Big(\mathit{flip}(1), \mathit{flip}(0)\Big). \end{aligned}$$

(Using that the logarithm of zero is defined to be zero.)

4.8.7 This exercise uses the distance between a joint state and the product of its marginals as measure of entwinedness, like in Jacobs (2017b).

1 Take $\sigma_2 := \frac{1}{2}|00\rangle + \frac{1}{2}|11\rangle \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$, for $\mathbf{2} = \{0, 1\}$. Show that:

$$d\Big(\sigma_2,\ \sigma_2[1, 0] \otimes \sigma_2[0, 1]\Big) \;=\; \tfrac{1}{2}.$$

2 Take $\sigma_3 := \frac{1}{2}|000\rangle + \frac{1}{2}|111\rangle \in \mathcal{D}(\mathbf{2} \times \mathbf{2} \times \mathbf{2})$. Show that:

$$d\Big(\sigma_3,\ \sigma_3[1, 0, 0] \otimes \sigma_3[0, 1, 0] \otimes \sigma_3[0, 0, 1]\Big) \;=\; \tfrac{3}{4}.$$

3 Now define $\sigma_n \in \mathcal{D}(\mathbf{2}^n)$ for $n \geq 2$ as:

$$\sigma_n \coloneqq \tfrac{1}{2}|\underbrace{0\cdots0}_{n \text{ times}}\rangle + \tfrac{1}{2}|\underbrace{1\cdots1}_{n \text{ times}}\rangle.$$

Show that:

- each marginal $(\pi_i)_*(\sigma_n)$ equals $\frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle$;
- the product $\bigotimes_i (\pi_i)_*(\sigma_n)$ of the marginals is the uniform state on 2^n;
- $d\big(\sigma_n, \bigotimes_i (\pi_i)_*(\sigma_n)\big) = \dfrac{2^{n-1}-1}{2^{n-1}}$.

Note that the latter distance goes to 1 as n goes to infinity.

4.8.8 The next 'splitting lemma' is attributed to Jones Jones (1989), see *e.g.* Jung and Tix (1998); Mardare et al. (2016). For $\omega_1, \omega_2 \in \mathcal{D}(X)$ with distance $d \coloneqq d(\omega_1, \omega_2)$ one can find distributions $\omega_1', \omega_2', \sigma \in \mathcal{D}(X)$ so that both ω_1 and ω_2 can be written as convex sum:

$$\omega_i = d \cdot \omega_i' + (1-d) \cdot \sigma.$$

Prove this result.
Hint: Use the optimal coupling ρ from (4.32) to define $\sigma(x) = \frac{\rho(x,x)}{1-d}$.

5

Variance and covariance

The previous chapter introduced validity $\omega \models p$, of an observable p in a state / distribution ω. The current chapter uses validity to define the standard statistical concepts of variance and covariance, and the associated notions of standard deviation and correlation. Informally, for a random variable (ω, p), the variance $\mathit{Var}(\omega, p)$ describes the extent to which the observable p differs from the expected value $\omega \models p$, that is, how much much p varies or is spread out. Together, $\omega \models p$ and $\mathit{Var}(\omega, p)$ are representational values that capture the statistical essence of a random variable. The standard deviation of a random variable is the square root of its variance.

The notion of covariance is used to compare *two* random variables. What do we mean by *two*? We can have:

1 two random variables (ω, p_1) and (ω, p_2), with (possibly) different observables $p_1, p_2 \colon X \to \mathbb{R}$, but with the same shared state $\omega \in \mathcal{D}(X)$;
2 a joint state $\tau \in \mathcal{D}(X_1 \times X_2)$ together with two observables $q_1 \colon X_1 \to \mathbb{R}$ and $q_2 \colon X_2 \to \mathbb{R}$ on the two components X_1, X_2. Via weakening of the observables we get two random variables:

$$(\tau, q_1 \otimes \mathbf{1}) \qquad\qquad (\tau, \mathbf{1} \otimes q_2)$$

like in the first point, involving two observables $q_1 \otimes \mathbf{1}$ and $\mathbf{1} \otimes q_2$, now on the same set $X_1 \times X_2$.

These differences are significant, but the two cases are not always clearly distinguished in the literature. One of the principles in this book is to make states explicit. Hence we shall clearly distinguish between the first *shared-state* form of covariance and the second *joint-state* form.

Apart from these subtleties, covariance captures to what extent two random variables change together. Covariance may be positive, when the variables change together in the same direction, or negative, meaning that they change in opposite directions.

This chapter first introduces the basic definitions and results for variance, and for covariance in shared-state form. They are applied to draw distributions, from Chapter 3, in Section 5.2. The joint-state version of covariance is introduced in Section 5.3 and illustrated in several examples. Section 5.4 then establishes the equivalence between:

- non-entwinedness of a joint state, meaning that it is the product of its (two) marginals;
- joint-state independence of random variables on this state
- joint-state covariance is zero.

See Theorem 5.4.6 for details. Such equivalences do not hold for shared-state formulations. This is one important reason for being careful about the distinction between a shared state and a joint state.

Covariance and correlation of (observables on) joint states is relevant in the setting of updating,

in Chapter 6. In presence of such correlation, updating in one (product) component has crossover influence in the other component.

This chapter also introduces the notion of moment of a random variable. It is an infinite sequence of representational numbers, where epected value (validity) and variance correspond to the first few moments. More generally, for each number $n \in \mathbb{N}$, the so-called n-th moment exists for a random variable. We look in particular at moments for distributions on the reals — with the identity as observable. Then, the sequence of moments entirely determines the distribution. Moments are especially useful for K-fold convolutions, where the number K goes to infinity. As already briefly shown at the end of Section 2.7, when K goes to infinity, the K-fold convolution takes the form of a Bell curve, for any distribution. This can now be made more precise.

Finally, Section 5.6, introduces the law of large draws, as a metric variation of the law of large numbers. This law of large draws says that the (total variation) distance $d(\omega, \mathit{Flrn}(\varphi))$ between a distribution ω and the frequentist learning $\mathit{Flrn}(\varphi)$ of a large multinomial draw φ from ω, goes to zero. This requires that the probabilities of the draws are taken into account. It is thus a limit result, 'in probability'. This is an alternative way of expressing the denseness of fractional distributions among all distributions, as formulated in Theorem 4.8.11. It is shown how the results that are commonly called the (weak) laws of large numbers can be derived from the law of large draws. Recall, there are also laws of large urns, expressing that for a very large urn, hypergeometric and Pólya draws are essentially multinomial draws, see Proposition 4.8.2.

5.1 Variance and shared-state covariance

This section describes the standard notions of variance, covariance and correlation within the setting of this book. It uses the validity relation $\models$ and the operations on observables from Section 4.4. Recall from Definition 4.3.1 that we understand a random variable here as a pair (ω, p) consisting of a state $\omega \in \mathcal{D}(X)$ and an observable $p\colon X \to \mathbb{R}$. The validity $\omega \models p$ is a real number, and can thus be used as a scalar, in the sense of Section 4.4. The truth predicate forms an observable $\mathbf{1} \in \mathit{Obs}(X)$; scalar multiplication yields a new observable $(\omega \models p) \cdot \mathbf{1} \in \mathit{Obs}(X)$. It can be subtracted[1] from p, and then squared, giving an observable:

$$(p - (\omega \models p) \cdot \mathbf{1})^2 \;=\; (p - (\omega \models p) \cdot \mathbf{1}) \,\&\, (p - (\omega \models p) \cdot \mathbf{1}) \;\in\; \mathit{Obs}(X).$$

This observable denotes the function that sends $x \in X$ to $(p(x) - (\omega \models p))^2 \in \mathbb{R}_{\geq 0}$. It is thus a factor. Its validity in the original state ω is called variance. It captures how far the values of p are spread out from their expected value.

Definition 5.1.1 For a random variable (ω, p), the *variance* $\mathit{Var}(\omega, p)$ is the non-negative number defined by:

$$\mathit{Var}(\omega, p) \;\coloneqq\; \omega \models (p - (\omega \models p) \cdot \mathbf{1})^2.$$

When the support of ω is a subset of $\mathbb{R}$, that is, when $\omega \in \mathcal{D}(\mathbb{R})$, and when p is the identity observable, we simply write $\mathit{Var}(\omega)$ for $\mathit{Var}(\omega, \mathit{id})$.

[1] Subtraction expressions like these occur more frequently in mathematics. For instance, an eigenvalue λ of a matrix M may be defined as the scalar that forms a solution to the equation $M - \lambda \cdot \mathbf{1} = 0$, where $\mathbf{1}$ is the identity matrix. A similar expression is used to define the elements in the spectrum of a C^*-algebra. See also Excercise 4.4.13.

The name *standard deviation* is used for the square root of the variance; thus:

$$\mathit{StDev}(\omega, p) \;:=\; \sqrt{\mathit{Var}(\omega, p)}.$$

Example 5.1.2 1 We recall Example 4.1.3 (1), with distribution $\mathit{flip}(\frac{3}{10}) = \frac{3}{10}|1\rangle + \frac{7}{10}|0\rangle$ and observable $v(0) = -50$ and $v(1) = 100$. We had $\omega \models v = -5$, and so we get:

$$\begin{aligned}\mathit{Var}\Big(\mathit{flip}(\tfrac{3}{10}), v\Big) &= \sum_{x\in\{0,1\}} \mathit{flip}(\tfrac{3}{10})(x) \cdot (v(x) + 5)^2 \\ &= \tfrac{3}{10} \cdot (100 + 5)^2 + \tfrac{7}{10} \cdot (-50 + 5)^2 \;=\; 4725.\end{aligned}$$

The standard deviation is around 68.7.

2 For a (fair) dice we have $\mathit{pips} = \{1, 2, 3, 4, 5, 6\} \hookrightarrow \mathbb{R}$ and $\mathit{Mean}(\mathit{dice}) = \frac{7}{2}$ so that:

$$\begin{aligned}\mathit{Var}(\mathit{dice}) &= \sum_{x\in\mathit{pips}} \mathit{dice}(x) \cdot (x - \tfrac{7}{2})^2 \\ &= \tfrac{1}{6} \cdot \Big((\tfrac{5}{2})^2 + (\tfrac{3}{2})^2 + (\tfrac{1}{2})^2 + (\tfrac{1}{2})^2 + (\tfrac{3}{2})^2 + (\tfrac{5}{2})^2\Big) \;=\; \tfrac{35}{12}.\end{aligned}$$

Via a suitable shift-and-rescale one can standardise an observable so that its validity becomes 0 and its variance becomes 1, see Exercise 5.1.19.

The following result is known as the partition of variance into expected values; it is often useful in calculations. But it also has some important consequences, formulated in a subsequent corollary.

Lemma 5.1.3 *Variance satisfies:*

$$\mathit{Var}(\omega, p) \;=\; (\omega \models p^2) - (\omega \models p)^2.$$

Proof. We have:

$$\begin{aligned}\mathit{Var}(\omega, p) &= \omega \models (p - (\omega \models p) \cdot \mathbf{1})^2 \\ &= \sum_{x\in X} \omega(x) \cdot (p(x) - (\omega \models p))^2 \\ &= \sum_{x\in X} \omega(x) \cdot \Big(p(x)^2 - 2(\omega \models p) \cdot p(x) + (\omega \models p)^2\Big) \\ &= \left(\sum_{x\in X} \omega(x)p^2(x)\right) - 2(\omega \models p) \cdot \left(\sum_{x\in X} \omega(x) \cdot p(x)\right) + \left(\sum_{x\in X} \omega(x) \cdot (\omega \models p)^2\right) \\ &= (\omega \models p^2) - 2(\omega \models p) \cdot (\omega \models p) + (\omega \models p)^2 \\ &= (\omega \models p^2) - (\omega \models p)^2.\end{aligned}$$

❑

Corollary 5.1.4 *Let $\omega \in \mathcal{D}(X)$ be a distribution with an observable p on the same set X.*

1 *There is an inequality:*

$$\omega \models p^2 \;\geq\; (\omega \models p)^2.$$

Later, in Proposition 6.1.8 (1) we shall see that this inequality generalises to arbitrary powers $n \in \mathbb{N}$.

2 *In the setting of the previous item there is an inequality:*

$$\omega \models \big|p - (\omega \models p) \cdot \mathbf{1}\big| \;\leq\; \sqrt{\mathit{Var}(\omega, p)}.$$

3 *For a channel $c\colon X \rightsquigarrow Y$ and an observable q on Y there is a pointwise inequality of observables (see Subsection 4.4.1):*

$$c^*(q^2) \;\geq\; (c^*(q))^2 \qquad \textit{i.e.} \qquad c^*(q \,\&\, q) \;\geq\; c^*(q) \,\&\, c^*(q).$$

Proof. 1 The inequality follows directly from Lemma 5.1.3, where we use that variance is non-negative — since it is defined as validity of a square. The inequality also occurs in Exercise 4.4.6 (2).

2 We abbreviate $q := \big|\, p - (\omega \models p) \cdot \mathbf{1} \,\big|$, so that $q(x) = \big|\, p(x) - (\omega \models p) \,\big|$. Then:

$$Var(\omega, p) \;=\; \omega \models q^2 \;\geq\; (\omega \models q)^2, \qquad \text{by the previous point.}$$

As a result,

$$\omega \models \big|\, p - (\omega \models p) \cdot \mathbf{1} \,\big| \;=\; \omega \models q \;\leq\; \sqrt{Var(\omega, p)}.$$

3 Let $x \in X$ be arbitrary. Then:

$$\begin{aligned} c^*(q \,\&\, q)(x) &= c(x) \models q^2 && \text{see Definition 4.5.1} \\ &\geq (c(x) \models q)^2 && \text{by the previous item} \\ &= c^*(q)(x) \cdot c^*(q)(x) \\ &= \big(c^*(q) \,\&\, c^*(q)\big)(x). \end{aligned}$$

❑

The formulation in Lemma 5.1.3 is convenient to compute variances, as illustrated in the following example. It introduces specific discrete distributions that approximate continuous 'normal' distributions, see also Section 5.5 for a more general description.

Example 5.1.5 We fix a number $N \in \mathbb{N}_{>0}$ and form the set:

$$S_N := \{0, 1, \ldots, 2N\} = \{0, \ldots, N-1, N, N+1, \ldots, 2N\}.$$

It has $2N + 1$ elements, with N sitting in the middle.

For $K \in \mathbb{N}_{>0}$ we consider the *average* function:

$$(S_N)^K \xrightarrow{avg[K]} \mathbb{Q} \qquad \text{where} \qquad avg[K](i_1, \ldots, i_K) := \frac{i_1 + \cdots + i_K}{K}.$$

We define a sequence of distributions $\omega_K \in \mathcal{D}(\mathbb{Q})$, as:

$$\begin{aligned} \omega_K &:= \mathcal{D}(avg[K])\big(unif_{S_N} \otimes \cdots \otimes unif_{S_N}\big) \\ &= \frac{unif_{S_N} + \cdots + unif_{S_N}}{K} && \text{in the style of Proposition 2.7.2} \\ &= \frac{1}{K} * K \cdot unif_{S_N} && \text{using scalar multiplication (2.64).} \end{aligned} \tag{5.1}$$

The support of this distribution ω_K is given by the set of numbers from 0 to $2N$, with steps of $\frac{1}{K}$. Thus:

$$\Big\{0, \tfrac{1}{K}, \ldots, N - \tfrac{1}{K}, N, N + \tfrac{1}{K}, \ldots, 2N - \tfrac{1}{K}, 2N\Big\}.$$

It has $2K \cdot N + 1$ elements, with N as midpoint. Figure 5.1 contains plots of the resulting distributions, for $N = 5$ and $K = 2, 3, 4, 5$. We see that they approximate a Bell curve, which is typical for normal distribuitions. Our aim is to compute the mean and variance of these distributions. From the pictures it is clear that the mean is N, but we like to establish this formally.

We shall derive the mean and variance of the distributions $\omega_K \in \mathcal{D}(S_N)$, namely:

$$Mean(\omega_K) = N \qquad \text{and} \qquad Var(\omega_K) = \frac{N \cdot (N+1)}{3K}. \tag{5.2}$$

The first equation is as expected. The proof involves the inclusion $S_N \subseteq \mathbb{R}$ and uses Proposition 1.2.6 (4):

$$\begin{aligned} Mean(\omega_K) = \omega_K \models id &= \sum_{0 \leq i_1, \ldots, i_K \leq 2N} (unif_{S_N})^K(i_1, \ldots, i_K) \cdot avg(i_1, \ldots, i_K) \\ &= \sum_{0 \leq i_1, \ldots, i_K \leq 2N} \frac{1}{(2N+1)^K} \cdot \frac{i_1 + \cdots + i_K}{K} \\ &= \frac{1}{(2N+1)^K \cdot K} \cdot \frac{K \cdot 2N \cdot (2N+1)^K}{2} = N. \end{aligned}$$

To prove the variance equation in (5.2) we use Lemma 5.1.3 and Proposition 1.2.6 (5):

$$\begin{aligned} Var(\omega_K) &= (\omega_K \models id^2) - (\omega_K \models id)^2 \\ &= \sum_{0 \leq i_1, \ldots, i_K \leq 2N} \frac{1}{(2N+1)^K} \cdot \Big(\frac{i_1 + \cdots + i_K}{K}\Big)^2 - N^2 \\ &= \frac{1}{(2N+1)^K \cdot K^2} \cdot \frac{K \cdot 2N \cdot (2N+1)^K \cdot ((3K+1) \cdot 2N + 2)}{12} - N^2 \\ &= \frac{N \cdot ((3K+1) \cdot 2N + 2) - 6K \cdot N^2}{6K} = \frac{2N^2 + 2N}{6K} = \frac{N \cdot (N+1)}{3K}. \end{aligned}$$

The mean is constant, for fixed N, independently of the value of K. The variance does depend on K and goes to zero as K goes to infinity. In the pictures in Figure 5.1 the bells become narrower, as K increases. In this way one can choose discrete distributions with specific (fractional) variances. Of course, the mean can be moved to an arbitrary position via the shift function of Exercise 2.7.8.

The formulation of Lemma 5.1.3 is useful for calculating the variance in specific cases.

Proposition 5.1.6 *The mean and variance of the nomial distribution* (2.3) *are:*

$$Mean\big(nom[N,K]\big) = \frac{(N-1) \cdot K}{2} \qquad \textit{and} \qquad Var\big(nom[N,K]\big) = \frac{(N+1) \cdot (N-1) \cdot K}{12}.$$

Proof. We use Lemma 1.10.4. First:

$$Mean\big(nom[N,K]\big) = \sum_{0 \leq i \leq (N-1) \cdot K} \frac{C_N(K,i) \cdot i}{N^K} = \frac{\frac{(N-1) \cdot K}{2} \cdot N^K}{N^K} = \frac{(N-1) \cdot K}{2}.$$

Next:

$$\begin{aligned} Var\big(nom[N,K]\big) &= \sum_{0 \leq i \leq (N-1) \cdot K} \frac{C_N(K,i) \cdot i^2}{N^K} - Mean\big(nom[N,K]\big)^2 \\ &= \Big(\frac{(N-1) \cdot K}{2} + \frac{N+1}{6}\Big) \cdot \frac{(N-1) \cdot K}{2} - \frac{(N-1)^2 \cdot K^2}{4} \\ &= \frac{(N+1) \cdot (N-1) \cdot K}{12}. \end{aligned}$$

❑

We continue with covariance and correlation, which involve *two* random variables, instead one, as for variance. We can distinguish situations, namely:

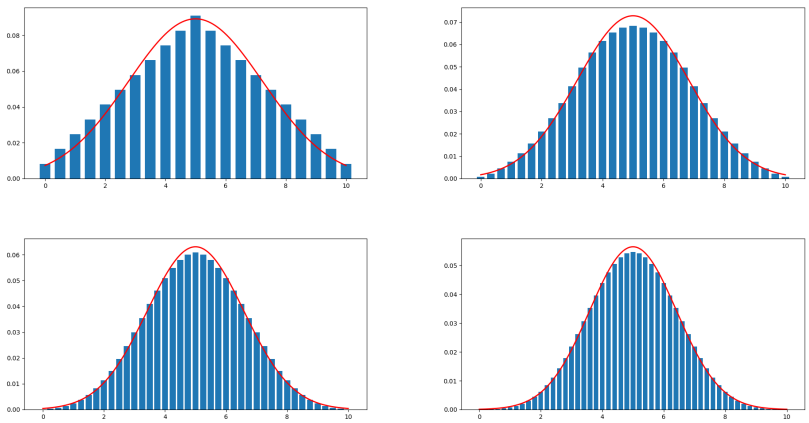


Figure 5.1 Averages ω_K of uniform distributions on $\{0, 1, \ldots, 10\}$ from (5.1), for $K = 2, 3$ at the top and $K = 4, 5$ at the bottom. The red line is (a scaled version of) the probability density function of the continuous normal distribution $\frac{1}{K} \cdot Norm(N, \sqrt{N \cdot (N+1)/3K})$, see Section **??**.

- The two random variables are of the form (ω, p_1) and (ω, p_2), where they share their state ω.
- There is a joint state $\tau \in \mathcal{D}(X_1 \times X_2)$ together with two observables $q_1 \colon X_1 \to \mathbb{R}$ and $q_2 \colon X_2 \to \mathbb{R}$ on the two components X_1, X_2. This situation can be seen as a special case of the previous point by first weakening the two observables to the product space, via: $(\pi_1)^*(q_1) = q_1 \otimes \mathbf{1}$ and $(\pi_2)^*(q_2) = \mathbf{1} \otimes q_2$. In this way we obtain two random variable with a shared state:

$$(\tau,\ (\pi_1)^*(q_1)) \qquad \text{and} \qquad (\tau,\ (\pi_2)^*(q_2)).$$

  These observable transformations $(\pi_i)^*(q_i)$ along a deterministic channel $\pi_i \colon X_1 \times X_2 \to X_i$ can also be described simply as function composition $q_i \circ \pi_i \colon X_1 \times X_2 \to \mathbb{R}$, see Lemma 4.5.2 (9).

We start with the situation in the first bullet above, and deal with the second bullet in Definition 5.3.1 in Section 5.3.

Definition 5.1.7 Let (ω, p_1) and (ω, p_2) be two random variable with a shared state $\omega \in \mathcal{D}(X)$.

1 The *covariance* of these random variables is defined as the validity:

$$Cov\big(\omega, p_1, p_2\big) \ \coloneqq\ \omega \models (p_1 - (\omega \models p_1) \cdot \mathbf{1})\ \&\ (p_2 - (\omega \models p_2) \cdot \mathbf{1}).$$

2 The *correlation* between (ω, p_1) and (ω, p_2) is the covariance divided by their standard deviations:

$$Cor\big(\omega, p_1, p_2\big) \ \coloneqq\ \frac{Cov(\omega, p_1, p_2)}{StDev(\omega, p_1) \cdot StDev(\omega, p_2)}.$$

Notice that variance $Var(\omega, p)$ is a special case of covariance $Cov(\omega, p, p)$, namely with equal observables. Correlation is normalised covariance, so that the outcome is in the interval $[-1, 1]$, see Exercise 5.1.22 below. In ordinary language two phenomena are called correlated when there is relation between them. More technically, two random variables are called correlated if their

correlation, as defined above, is non-zero — or equivalently, when their covariance is non-zero. Positive correlation means that the observables move together in the same direction, whereas negative correlation means that they move in opposite directions. This will be made more precise in terms of updating, see Proposition 6.11.1. When the correlation equals 1 (resp. -1), one speaks of perfect correlation (resp. anticorrelation).

Before we go on, we state the following analogue of Lemma 5.1.3, leaving the proof to the reader.

Lemma 5.1.8 *Covariance can be reformulated as:*

$$Cov\big(\omega, p_1, p_2\big) = (\omega \models p_1 \;\&\; p_2) - (\omega \models p_1) \cdot (\omega \models p_2). \qquad \square$$

Example 5.1.9 1 We have seen in Definition 4.1.2 (2) how the average of an observable can be computed as its validity in a uniform state. The same approach is used to compute the covariance (and correlation) in a uniform joint state. Consider the following to lists a and b of numerical data, of the same length.

$$a = [5, 10, 15, 20, 25] \qquad\qquad b = [10, 8, 10, 15, 12]$$

We will identify a and b with random variables, namely with $a, b \colon \mathbf{5} \to \mathbb{R}$, where $\mathbf{5} = \{0, 1, 2, 3, 4\}$. Hence there are obvious definitions: $a(0) = 5$, $a(1) = 10$, $a(2) = 15$, $a(3) = 20$, $a(4) = 25$, and simililarly for b. Then we can compute their averages as validities in the uniform state $unif_{\mathbf{5}}$ on the set $\mathbf{5}$, namely as:

$$avg(a) = unif_{\mathbf{5}} \models a = 15 \qquad\qquad avg(a) = unif_{\mathbf{5}} \models a = 11.$$

We will calculate the covariance between a and b with respect to the uniform state $unif_{\mathbf{5}}$, as:

$$\begin{aligned} Cov(unif_{\mathbf{5}}, a, b) &= unif_{\mathbf{5}} \models (a - (unif_{\mathbf{5}} \models a) \cdot \mathbf{1}) \;\&\; (b - (unif_{\mathbf{5}} \models b) \cdot \mathbf{1}) \\ &= \textstyle\sum_i \frac{1}{5} \cdot (a(i) - 15) \cdot (b(i) - 11) = 11. \end{aligned}$$

2 In order to obtain the correlation between a, b, we first need to compute their variances:

$$\begin{aligned} Var(unif_{\mathbf{5}}, a) &= \textstyle\sum_i \frac{1}{5}(a(i) - 15)^2 = 50 \\ Var(unif_{\mathbf{5}}, b) &= \textstyle\sum_i \frac{1}{5}(b(i) - 11)^2 = 5.6 \end{aligned}$$

Then:

$$Cor(unif_{\mathbf{5}}, a, b) = \frac{Cov(unif_{\mathbf{5}}, a, b)}{\sqrt{Var(unif_{\mathbf{5}}, a)} \cdot \sqrt{Var(unif_{\mathbf{5}}, b)}} = \frac{11}{\sqrt{50} \cdot \sqrt{5.6}} \approx 0.66.$$

The next result collects several linearity properties for (co)variance and correlation. This shows that one can do quite a bit of re-scaling and stretching of observables without changing the outcome.

Theorem 5.1.10 *Consider a state $\omega \in \mathcal{D}(X)$, with observables $p, p_1, p_2 \in Obs(X)$ and numbers $r, s \in \mathbb{R}$.*

1 Covariance satisfies:

$$\begin{aligned}
Cov(\omega, p_1, p_2) &= Cov(\omega, p_2, p_1) \\
Cov(\omega, p_1, \mathbf{1}) &= 0 \\
Cov(\omega, r \cdot p_1, p_2) &= r \cdot Cov(\omega, p_1, p_2) \\
Cov(\omega, p_1 + p_2, p) &= Cov(\omega, p_1, p) + Cov(\omega, p_2, p) \\
Cov(\omega, p_1 + r \cdot \mathbf{1}, p_2 + s \cdot \mathbf{1}) &= Cov(\omega, p_1, p_2).
\end{aligned}$$

2 Variance satisfies:

$$\begin{aligned}
Var(\omega, r \cdot p) &= r^2 \cdot Var(\omega, p) \\
Var(\omega, p + r \cdot \mathbf{1}) &= Var(\omega, p) \\
Var(\omega, p_1 + p_2) &= Var(\omega, p_1) + 2 \cdot Cov(\omega, p_1, p_2) + Var(\omega, p_2).
\end{aligned}$$

3 For correlation we have:

$$\begin{aligned}
Cor(\omega, p_1, p_2) &= Cor(\omega, p_2, p_1) \\
Cor(\omega, p_1, \mathbf{1}) &= 0 \\
Cor(\omega, r \cdot p_1, s \cdot p_2) &= \begin{cases} Cor(\omega, p_1, p_2) & \textit{if } r, s \textit{ have the same sign} \\ -Cor(\omega, p_1, p_2) & \textit{otherwise.} \end{cases} \\
Cor(\omega, p_1 + r \cdot \mathbf{1}, p_2 + s \cdot \mathbf{1}) &= Cor(\omega, p_1, p_2).
\end{aligned}$$

Proof. 1 Obviously, covariance is symmetric and covariance with truth is 0. Covariance preserves scalar multiplication in each (observable) argument since by Lemma 5.1.8:

$$\begin{aligned}
Cov(\omega, r \cdot p_1, p_2) &= (\omega \models (r \cdot p_1) \,\&\, p_2) - (\omega \models r \cdot p_1) \cdot (\omega \models p_2) \\
&= r \cdot (\omega \models p_1 \,\&\, p_2) - r \cdot (\omega \models p_1) \cdot (\omega \models p_2) \\
&= r \cdot Cov(\omega, p_1, p_2).
\end{aligned}$$

For preservation of sums we reason from the definition:

$$\begin{aligned}
Cov(\omega, p_1 + p_2, p) &= (\omega \models (p_1 + p_2) \,\&\, p) - (\omega \models p_1 + p_2) \cdot (\omega \models p) \\
&= (\omega \models (p_1 \,\&\, p) + (p_2 \,\&\, p)) - ((\omega \models p_1) + (\omega \models p_2)) \cdot (\omega \models p) \\
&= (\omega \models p_1 \,\&\, p) + (\omega \models p_2 \,\&\, p) \\
&\qquad - (\omega \models p_1) \cdot (\omega \models p) - (\omega \models p_2) \cdot (\omega \models p) \\
&= Cov(\omega, p_1, p) + Cov(\omega, p_2, p)
\end{aligned}$$

The equation $Cov(\omega, p_1 + r \cdot \mathbf{1}, p_2 + s \cdot \mathbf{1}) = Cov(\omega, p_1, p_2)$ follows from the previous equations.

2 The first property holds by what we have just seen:

$$Var(\omega, r \cdot p) = Cov(\omega, r \cdot p, r \cdot p) = r^2 \cdot Cov(\omega, p, p) = r^2 \cdot Var(\omega, p).$$

Similarly, $Var(\omega, p + r \cdot \mathbf{1}) = Var(\omega, p)$. Next:

$$\begin{aligned}
Var(\omega, p_1 + p_2) &= Cov(\omega, p_1 + p_2, p_1 + p_2) \\
&= Cov(\omega, p_1 + p_2, p_1) + Cov(\omega, p_1 + p_2, p_2) \\
&= Cov(\omega, p_1, p_1) + Cov(\omega, p_2, p_1) + Cov(\omega, p_1, p_2) + Cov(\omega, p_2, p_2) \\
&= Var(\omega, p_1) + 2 \cdot Cov(\omega, p_1, p_2) + Var(\omega, p_2).
\end{aligned}$$

3 The first two equations are obvious, given their counterparts in item (1). By unpacking the definition of correlation and using the previous two items we get:

$$\begin{aligned}
Cor(\omega, r \cdot p_1, s \cdot p_2) &= \frac{Cov(\omega, r \cdot p_1, s \cdot p_2)}{\sqrt{Var(\omega, r \cdot p_1)} \cdot \sqrt{Var(\omega, s \cdot p_2)}} \\
&= \frac{r \cdot s \cdot Cov(\omega, p_1, p_2)}{\sqrt{r^2 \cdot Var(\omega, p_1)} \cdot \sqrt{s^2 \cdot Var(\omega, p_2)}} \\
&= \frac{r \cdot s \cdot Cov(\omega, p_1, p_2)}{|r| \cdot \sqrt{Var(\omega, p_1)} \cdot |s| \cdot \sqrt{Var(\omega, p_2)}} \\
&= \begin{cases} Cor(\omega, p_1, p_2) & \text{if } r, s \text{ have the same sign} \\ -Cor(\omega, p_1, p_2) & \text{otherwise.} \end{cases}
\end{aligned}$$

(The same sign means: either both $r \geq 0$ and $s \geq 0$ or both $r \leq 0$ and $s \leq 0$.)

The final equation $Cor(\omega, p_1 + r \cdot \mathbf{1}, p_2 + s \cdot \mathbf{1}) = Cor(\omega, p_1, p_2)$ holds since both variance and covariance are closed under addition of constants. ❑

We conclude this section with bivariate binomial distributions, introduced in Definition 4.2.7. The next result is the analogue of Proposition 4.2.10. It involves variances and covariances of distributions on two-dimensional spaces. For the case of the ordinary binomial distribution, see Exercise 5.1.15 below.

Proposition 5.1.11 *Let $bvbn[K](\gamma) \in \mathcal{D}\big(\{0, \ldots, K\} \times \{0, \ldots, K\}\big)$ be a bivariate binomial distribution, as introduced in Definition 4.2.7, for a draw size $K \in \mathbb{N}$ and a two-coin distribution $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$.*

For clarity we use different notation for the various projections:

$$\mathbf{2} \times \mathbf{2} \underset{\pi_2}{\overset{\pi_1}{\rightrightarrows}} \mathbf{2} \qquad\qquad \{0, \ldots, K\} \times \{0, \ldots, K\} \underset{\Pi_2}{\overset{\Pi_1}{\rightrightarrows}} \{0, \ldots K\}$$

We use these projections as observables in:

1 $Var\big(bvbn[K](\gamma), \Pi_i\big) = K \cdot Var\big(\gamma, \pi_i\big)$;

2 $Cov\big(bvbn[K](\gamma), \Pi_i, \Pi_j\big) = K \cdot Cov\big(\gamma, \pi_i, \pi_j\big)$, *for* $i, j \in \{1, 2\}$.

Proof. 1 We do the case for $i = 1$. First,

$$\begin{aligned}
Var\big(\gamma, \pi_1\big) &= \big(\gamma \models \pi_1 \;\&\; \pi_1\big) - \big(\gamma \models \pi_1\big)^2 \\
&= \left(\sum_{b_1, b_2 \in \mathbf{2}} \gamma(b_1, b_2) \cdot b_1 \cdot b_1\right) - \left(\sum_{b_1, b_2 \in \mathbf{2}} \gamma(b_1, b_2) \cdot b_1\right)^2 \\
&= \gamma(1, 0) + \gamma(1, 1) - \big(\gamma(1, 0) + \gamma(1, 1)\big)^2.
\end{aligned}$$

We turn to the bivariate binomial and make use of Lemma 3.3.3 and Exercise 3.3.8.

$$
\begin{aligned}
&Var\Big(bvbn[K](\gamma), \Pi_i\Big)\\
&= \Big(bvbn[K](\gamma) \models \Pi_1 \;\&\; \Pi_1\Big) - \Big(bvbn[K](\gamma) \models \Pi_1\Big)^2\\
&= \left(\sum_{n_1,n_2\in\mathbb{N}} bvbn[K](\gamma)(n_1,n_2)\cdot n_1\cdot n_1\right) - \left(\sum_{n_1,n_2\in\mathbb{N}} bvbn[K](\gamma)(n_1,n_2)\cdot n_1\right)^2\\
&= \left(\sum_{\varphi\in\mathcal{N}[K](\mathbf{2}\times\mathbf{2})} mn[K](\gamma)(\varphi)\cdot(\varphi(1,0)+\varphi(1,1))^2\right)\\
&\qquad - \left(\sum_{\varphi\in\mathcal{N}[K](\mathbf{2}\times\mathbf{2})} mn[K](\gamma)(\varphi)\cdot(\varphi(0,1)+\varphi(1,1))\right)^2\\
&= K\cdot(K-1)\cdot\gamma(1,0)^2 + K\cdot\gamma(1,0) \;+\; 2\cdot K\cdot(K-1)\cdot\gamma(1,0)\cdot\gamma(1,1)\\
&\qquad + K\cdot(K-1)\cdot\gamma(1,1)^2 + K\cdot\gamma(1,1) - \Big(K\cdot\gamma(1,0)+K\cdot\gamma(1,1)\Big)^2\\
&= K\cdot\Big(-\gamma(1,0)^2+\gamma(1,0)-2\cdot\gamma(1,0)\cdot\gamma(1,1)-\gamma(1,1)^2+\gamma(1,1)\Big)\\
&= K\cdot\Big(\gamma(1,0)+\gamma(1,1)-(\gamma(1,0)+\gamma(1,1))^2\Big)\\
&= K\cdot Var\Big(\gamma,\pi_1\Big) \quad \text{as shown above.}
\end{aligned}
$$

2 The cases where $i = j$ are covered by the previous point, since covariance with equal observables is variance. We do the case $i = 1, j = 2$, which gives the same outcome as for $i = 2, j = 1$. We first look at the covariance of the two-coin distribution γ.

$$
\begin{aligned}
Cov\Big(\gamma,\pi_1,\pi_2\Big) &= \Big(\gamma\models\pi_1 \;\&\; \pi_2\Big) - \Big(\gamma\models\pi_1\Big)\cdot\Big(\gamma\models\pi_2\Big)\\
&= \left(\sum_{b_1,b_2\in\mathbf{2}} \gamma(b_1,b_2)\cdot b_1\cdot b_2\right) - \left(\sum_{b_1,b_2\in\mathbf{2}} \gamma(b_1,b_2)\cdot b_1\right)\cdot\left(\sum_{b_1,b_2\in\mathbf{2}} \gamma(b_1,b_2)\cdot b_2\right)\\
&= \gamma(1,1) \;-\; \Big(\gamma(1,0)+\gamma(1,1)\Big)\cdot\Big(\gamma(0,1)+\gamma(1,1)\Big).
\end{aligned}
$$

Next, again by Lemma 3.3.3 and Exercise 3.3.8,

$$
\begin{aligned}
&Cov\Big(bvbn[K](\gamma), \Pi_1, \Pi_2\Big)\\
&= \Big(bvbn[K](\gamma) \models \Pi_1 \;\&\; \Pi_2\Big) - \Big(bvbn[K](\gamma) \models \Pi_1\Big)\cdot\Big(bvbn[K](\gamma) \models \Pi_2\Big)\\
&= \left(\sum_{n_1,n_2\in\mathbb{N}} bvbn[K](\gamma)(n_1,n_2)\cdot n_1\cdot n_2\right)\\
&\qquad - \left(\sum_{n_1,n_2\in\mathbb{N}} bvbn[K](\gamma)(n_1,n_2)\cdot n_1\right)\cdot\left(\sum_{n_1,n_2\in\mathbb{N}} bvbn[K](\gamma)(n_1,n_2)\cdot n_2\right)\\
&= \left(\sum_{\varphi\in\mathcal{N}[K](\mathbf{2}\times\mathbf{2})} mn[K](\gamma)(\varphi)\cdot(\varphi(1,0)+\varphi(1,1))\cdot(\varphi(0,1)+\varphi(1,1))\right)\\
&\qquad - \Big(K\cdot\gamma(1,0)+K\cdot\gamma(1,1)\Big)\cdot\Big(K\cdot\gamma(0,1)+K\cdot\gamma(1,1)\Big)\\
&= K\cdot(K-1)\cdot\Big(\gamma(1,0)\cdot\gamma(0,1)+\gamma(1,0)\cdot\gamma(1,1)+\gamma(1,1)\cdot\gamma(0,1)\\
&\qquad + \gamma(1,1)^2\Big) + K\cdot\gamma(1,1) \;-\; K^2\cdot\Big(\gamma(1,0)+\gamma(1,1)\Big)\cdot\Big(\gamma(0,1)+\gamma(1,1)\Big)\\
&= K\cdot\Big(\gamma(1,1) \;-\; (\gamma(1,0)+\gamma(1,1))\cdot(\gamma(0,1)+\gamma(1,1))\Big)\\
&= Cov\Big(\gamma,\pi_1,\pi_2\Big), \quad \text{see before.}
\end{aligned}
$$

❑

Exercises

5.1.1 Show that the pyramid distributions $\Lambda_n \in \mathcal{D}(\mathbb{Z})$ from Exercise 2.1.1 satisfy:

$$Mean(\Lambda_n) \;=\; 0 \qquad \text{and} \qquad Var(\Lambda_n) \;=\; \frac{(n-1)\cdot(n+1)}{6} \;=\; \frac{\binom{n+1}{3}}{n}.$$

Hint: Use Proposition 1.2.6.

(The outcomes $\binom{n+1}{3}$, for $n \geq 2$, are known as Tetrahedral numbers.)

5.1.2 Let (ω, p) be a random variable, where $\omega \in \mathcal{D}(X)$ and $p\colon X \to \mathbb{R}$.

1 Show that that the pushforward distribution $\mathcal{D}(p)(\omega) \in \mathcal{D}(\mathbb{R})$ satisfies:

$$Mean(\mathcal{D}(p)(\omega)) \;=\; \omega \models p \qquad\qquad Var(\mathcal{D}(p)(\omega)) \;=\; Var(\omega, p).$$

(The first equation already occurred in Definition 4.3.1 (2) but is repeated to illustrate the uniformity.)

2 Consider the function $f := (p - (\omega \models p)\cdot \mathbf{1})^2\colon X \to \mathbb{R}$ used in the definition of variance. Show that the pushforward $\mathcal{D}(f)(\omega) \in \mathcal{D}(\mathbb{R})$ satisfies:

$$Mean(\mathcal{D}(f)(\omega)) \;=\; Var(\omega, p).$$

5.1.3 Let ω be a distribution and p be a factor on the same set. Define for $v \in \mathbb{R}_{\geq 0}$,

$$f(v) \;:=\; \omega \models (p - v\cdot \mathbf{1})^2.$$

Show that the function $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ takes its minimum value at $\omega \models p$.

5.1.4 Let $\omega \in \mathcal{D}(X)$ and $p \in Obs(X)$. Prove that:

$$Var(\omega, p) = 0 \iff p \text{ is constant on the subset } supp(\omega) \subseteq X.$$

5.1.5 Let $\tau \in \mathcal{D}(X \times Y)$ be a joint state with an observable $p\colon X \to \mathbb{R}$. We can turn them into a random variable in two ways, by marginalisation and weakening:

$$(\tau[1,0], p) \qquad \text{and} \qquad (\tau, (\pi_1)^*(p)),$$

where $(\pi_1)^*(p) = p \otimes \mathbf{1}$ is an observable on $X \times Y$.

These two random variables have the same expected value, by (4.11). Show that they have the same variance too:

$$Var(\tau[1,0], p) \;=\; Var(\tau, (\pi_1)^*(p)).$$

As a result, the standard deviations are also the same.

5.1.6 Define $g(r) := flip(r) \otimes flip(1-r) \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$ and show that for each $r \in [0,1]$,

$$Cov(g(r)) \;=\; Cov(g(r), \pi_1, \pi_2) \;=\; 0.$$

5.1.7 Use Jensen's inequality, see Lemma 2.10.3, to prove the inequality $\omega \models p^2 \;\geq\; (\omega \models p)^2$ in Corollary 5.1.4 (1).

5.1.8 Let $\lambda \in \mathbb{R}_{\geq 0}$. Recall that $Mean(pois[\lambda]) = \lambda$, see Example 4.2.2 (3).

1 Prove that $\displaystyle\sum_{k \geq 1} pois[\lambda](k)\cdot k\cdot(k-1) = \lambda^2$;

2 Use this equation to show that $\displaystyle\sum_{k \geq 1} pois[\lambda](k)\cdot k^2 = \lambda^2 + \lambda$;

3 Deduce that $Var(pois[\lambda]) = \lambda$. Thus, for the Poisson distribution the mean and variance are equal — see Example 4.2.2 (3) for the mean.

5.1.9 Prove Lemma 5.1.8 along the lines of the proof of Lemma 5.1.3.

5.1.10 Show for a predicate p,

1 $Cov(\omega, p, p^{\perp}) = -Var(\omega, p) \leq 0$;

2 $Var(\omega, p^{\perp}) = Var(\omega, p)$.

5.1.11 Generalise the third equation of Theorem 5.1.10 (2) to:

$$Var\big(\omega, \textstyle\sum_i p_i\big) = \textstyle\sum_{i,j} Cov(\omega, p_i, p_j),$$

where $\omega \in \mathcal{D}(X)$ and $p_1, \ldots, p_n \colon X \to \mathbb{R}$.

5.1.12 Recall from Lemma 4.2.4 the equation $Mean(\omega + \rho) = Mean(\omega) + Mean(\rho)$, for distributions $\omega, \rho \in \mathcal{D}(\mathbb{R})$ with convolution $\omega + \rho \in \mathcal{D}(\mathbb{R})$. Show that also:

$$\begin{aligned} Var(\omega + \rho) &= Var(\omega) + Var(\rho) \qquad \text{and thus also:} \\ Cov(\omega + \rho) &= Cov(\omega) + Cov(\rho). \end{aligned}$$

5.1.13 Recall the translation operation $shift(a, \omega) = 1|a\rangle + \omega = \mathcal{D}(a + (-))(\omega)$ from Exercise 2.7.8, which we use here for $a \in \mathbb{R}$ and $\omega \in \mathcal{D}(\mathbb{R})$.

1 Show that shifting a distribution $\omega \in \mathcal{D}(\mathbb{R})$ also shifts its mean, that is:

$$Mean(shift(a, \omega)) = a + Mean(\omega).$$

2 Show that the variance is stable under shifting:

$$Var(shift(a, \omega)) = Var(\omega).$$

5.1.14 For $\omega \in \mathcal{D}(\mathbb{R})$ and $s \in \mathbb{R}$ show that:

$$Mean(s * \omega) = s \cdot Mean(\omega) \qquad Var(s * \omega) = s^2 \cdot Var(\omega).$$

(The first equation occurred already in Lemma 4.2.4 but is repeated for clarity.)

5.1.15 Let $r \in [0, 1]$ be given.

1 Show that the distribution $flip(r) = r|1\rangle + (1 - r)|0\rangle$ satisfies:

$$Mean(flip(r)) = r \qquad Var(flip(r)) = r \cdot (1 - r).$$

2 Use the previous exercise and Proposition 2.7.6 (1) to conclude that the binomial distribution $bn[K](r) \in \mathcal{D}(\{0, 1, \ldots, K\})$, for $K \in \mathbb{N}$, satisfies:

$$Mean(bn[K](r)) = K \cdot r \qquad Var(bn[K](r)) = K \cdot r \cdot (1 - r).$$

Conclude that a binomial distribution is entirely determined by its mean and variance: if we know that $\omega \in \mathcal{D}(\mathbb{N})$ is a binomial distribution of which we know the mean m and the variance v, then we can find $K \in \mathbb{N}$ and $r \in [0, 1]$ such that $\omega = bn[K](r)$. This fact may be applied in a situation like in Example 3.4.3 (2): when the draw size K is not known, we may obtain it (after rounding) from the variance of the data ψ.

5.1.16 Suppose we have a distribution $\sigma \in \mathcal{D}\big(\{0,\dots,K\} \times \{0,\dots,K\}\big)$, for some given $K \in \mathbb{N}$. We know that σ is a bivariate binomial distribution, of the form $\sigma = bvbn[K](\gamma)$ but we do not know $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$. Use Proposition 5.1.11 to show that one can obtain the probabilities of γ, and thus γ itself, in the following way — using projections $\Pi_i \colon \{0,\dots,K\}^2 \to \{0,\dots,K\}$ as in Proposition 5.1.11.

- $\gamma(1,1) = \frac{1}{K} \cdot \big(Cov(\sigma, \Pi_1, \Pi_2) - (\sigma \models \Pi_1) \cdot (\sigma \models \Pi_2)\big)$;
- $\gamma(1,0) = \frac{1}{K} \cdot (\sigma \models \Pi_1) - \gamma(1,1)$;
- $\gamma(0,1) = \frac{1}{K} \cdot (\sigma \models \Pi_2) - \gamma(1,1)$;
- $\gamma(0,0) = 1 - \gamma(1,0) - \gamma(0,1) - \gamma(1,1)$.

Thus, when the draw size K is given, a bivariate binomial distribution is entirely determined by its mean and covariance, since its mean is given by the pair of validities ($bvbn[K](\gamma) \models \Pi_1$, $bvbn[K](\gamma) \models \Pi_2$), see (4.6).

5.1.17 Let $h \colon X \to Y$ be a function, with a state $\omega \in \mathcal{D}(X)$ on its domain an an observable $q \colon Y \to \mathbb{R}$ on its codomain. Show that:

$$Var\big(\omega,\, q \circ h\big) \;=\; Var\big(\mathcal{D}(h)(\omega),\, q\big).$$

Hint: Recall Exercise 4.5.5.

5.1.18 Consider a distribution $\omega \in \mathcal{D}(X)$, a channel $c \colon X \rightsquigarrow Y$ and an observable $q \colon Y \to \mathbb{R}$. Prove the 'law of total variance', see *e.g.* (Rice, 2006, §4.4.1, Thm. B), in analogy with the law of total expectation in Remark 4.5.5:

$$Var(c_*(\omega), q) \;=\; Var(\omega, c^*(q)) \;+\; (\omega \models Var(c(-), q)).$$

Conclude that $Var(c_*(\omega), q) \geq Var(\omega, c^*(q))$.

5.1.19 Let (ω, p) be a random variable on a space X, with non-zero variance (and thus non-zero standard deviation). Define a new *standard score* observable $p' = StSc(\omega, p) \colon X \to \mathbb{R}$ by:

$$p' \;\coloneqq\; \frac{1}{StDev(\omega, p)} \cdot (p - (\omega \models p) \cdot \mathbf{1}).$$

That is, pointwise, for $x \in X$,

$$p'(x) \;=\; \frac{p(x) - (\omega \models p)}{StDev(\omega, p)}.$$

This pair of ω with p' is also called the Z-random variable. It is normalised in the sense that:

1 $\omega \models p' \;=\; 0$;
2 $Var(\omega, p') = StDev(\omega, p') = 1$.

Prove these two items.

5.1.20 In the previous exercise we have standardised a random variable by adapting the observable. A distribution on the reals can be also be standardised, by adapting the distribution itself. This works as follows.

Let $\omega \in \mathcal{D}(\mathbb{R})$ be given with mean $\mu \coloneqq Mean(\omega) = \omega \models id$ and standard deviation $\sigma \coloneqq StDev(\omega) = \sqrt{Var(\omega, id)}$. We assume $\sigma \neq 0$. Define, in the style of the previous exercise (and Exercise 5.1.13), a new distribution $\omega' = StSc(\omega) \in \mathcal{D}(\mathbb{R})$ via scalar multiplication and shifting:

$$\omega' \;\coloneqq\; \tfrac{1}{\sigma} * shift(-\mu, \omega) \;=\; \mathcal{D}\Big(\tfrac{(-)-\mu}{\sigma}\Big)(\omega),$$

where $*$ is scalar multiplication from (2.64).

1 Prove that: $Mean(\omega') = 0$ and $Var(\omega') = StDev(\omega') = 1$.

2 Show that applying this standardisation producedure to the distribution $flip(r) = r|1\rangle + (1-r)|0\rangle$ yields:

$$flip'(r) = r\left|\sqrt{\tfrac{1-r}{r}}\right\rangle + (1-r)\left|-\sqrt{\tfrac{r}{1-r}}\right\rangle.$$

5.1.21 We now apply the standardisation $StSc$ from the previous exercise to a K-fold convolution $K \cdot \omega = \omega + \cdots + \omega$, for $K > 0$ and $\omega \in \mathcal{D}(\mathbb{R})$. Prove that:

$$StSc(K \cdot \omega) = \tfrac{1}{\sqrt{K}} * \big(K \cdot StSc(\omega)\big).$$

5.1.22 Recall the Cauchy-Schwarz inequality, for real numbers $a_i, b_i \in \mathbb{R}$,

$$\big(\textstyle\sum_i a_i b_i\big)^2 \leq \big(\textstyle\sum_i a_i^2\big) \cdot \big(\textstyle\sum_i b_i^2\big).$$

Use this inequality to prove that correlation is in the interval $[-1, 1]$.

5.1.23 Let $\omega \in \mathcal{D}(X)$ be a state with an n-test $\vec{p} = p_1, \ldots, p_n$, see Definition 4.4.2.

1 Define the (symmetric) covariance matrix as:

$$CovMat(\omega, \vec{p}) := \begin{pmatrix} Cov(\omega, p_1, p_1) & \cdots & Cov(\omega, p_1, p_n) \\ \vdots & & \vdots \\ Cov(\omega, p_n, p_1) & \cdots & Cov(\omega, p_n, p_n) \end{pmatrix}$$

Prove that all rows and all colums add up to 0 and that the entries on the diagonal are non-negative and have a sum below 1.

2 Next consider the vector v of validities and the symmetric matrix A of conjunctions:

$$v := \begin{pmatrix} \omega \models p_1 \\ \vdots \\ \omega \models p_n \end{pmatrix} \qquad A := \begin{pmatrix} \omega \models p_1 \,\&\, p_1 & \cdots & \omega \models p_1 \,\&\, p_n \\ \vdots & & \vdots \\ \omega \models p_n \,\&\, p_1 & \cdots & \omega \models p_n \,\&\, p_n \end{pmatrix}$$

Check that $CovMat(\omega, \vec{p}) = A - v \cdot v^T$, where $(-)^T$ is transpose.

5.1.24 This exercise elaborates an elementary proof of Corollary 5.1.4 (1). It relies on the square of sums formula:

$$(v_1 + \cdots + v_n)^2 = \sum_i v_i^2 + \sum_{j<i} 2 v_i v_j.$$

Now let $\omega \in \mathcal{D}(X)$ and $p \in Obs(X)$ be given, say with $supp(\omega) = \{x_1, \ldots, x_n\}$; write $\omega_i = \omega(x_i)$ and $p_i = p(x_i)$. Use the above the square of sums formula to prove the inequality in the middle:

$$(\omega \models p)^2 = \big(\textstyle\sum_i \omega_i \cdot p_i\big)^2 \leq \textstyle\sum_i \omega_i \cdot p_i^2 = \omega \models p \,\&\, p.$$

5.1.25 In linear regression a finite collection $(a_i, b_i)_{1 \leq i \leq n}$ of real numbers $a_i, b_i \in \mathbb{R}$ is given. The aim is to find coefficients $v, w \in \mathbb{R}$ of a line $y = vx + w$ that best approximates these points. The error that is minimised is the 'sum of squared residuals' given as:

$$f(v, w) := \sum\nolimits_i \big(b_i - (v a_i + w)\big)^2.$$

We redescribe the a_i, b_i as observables $a, b \colon \{1, 2, \ldots, n\} \to \mathbb{R}$, with $a(i) = a_i, b(i) = b_i$,

together with the uniform distribution *unif* on the space $\{1, 2, \ldots, n\}$. Thus we have two random variables $(unif, a)$ and $(unif, b)$ with a shared state. Write:

$$\overline{a} \coloneqq unif \models a = \tfrac{1}{n} \textstyle\sum_i a_i \qquad \overline{b} \coloneqq unif \models b = \tfrac{1}{n} \textstyle\sum_i b_i.$$

By taking partial derivatives $\frac{\partial f}{\partial v}$ and $\frac{\partial f}{\partial w}$, setting them to zero, and using some elementary calculus, one obtains the best linear approximation of the (a_i, b_i) via coefficients given by the familiar formulas:

$$\hat{v} = \frac{\sum_i a_i(b_i - \overline{b})}{\sum_i a_i(a_i - \overline{a})} \qquad \hat{w} = \overline{b} - \hat{v}\,\overline{a}.$$

1 Derive the above formulas for $\hat{v}$ and $\hat{w}$.

2 Show that one can also write the slope $\hat{v}$ of the best line as:

$$\hat{v} = \frac{\sum_i (a_i - \overline{a})(b_i - \overline{b})}{\sum_i (a_i - \overline{a})^2}.$$

3 Check that this yields:

$$\hat{v} = \frac{Cov(unif, a, b)}{Var(unif, a)}.$$

4 Thus, with the above values $\hat{v}$, $\hat{w}$ the sum of squares of $b - (\hat{v} \cdot a + \hat{w} \cdot \mathbf{1})$ is minimal. Show that the latter expression can also be written in terms of standard scores, see Exercise 5.1.19, namely as:

$$b - (\hat{v} \cdot a + \hat{w} \cdot \mathbf{1}) = StDev(b) \cdot \Big(StSc(b) - Cor(a, b) \cdot StSc(a)\Big),$$

where we have omitted the uniform distribution for convenience. The right-hand-side shows that by using correlation as scalar one can bring the standard score of a closest to the standard score of b.

Linear regression is described here as a technique for obtaining the 'best' line, from data points (a_i, b_i). Once this line is found, one can use it for *prediction*: if we have an arbitrary first coordinate a we can predict the corresponding second coordinate as $\hat{v} \cdot a + \hat{w}$. For instance, if a_i is number of hours spent learning by student i, and b_i is the resulting mark of student i, then linear regression may give a reasonable prediction of the mark given a (new) number a of time spent on learning. Chapter 8 is devoted to learning techniques, of which linear regression is simple instance.

5.2 Draw distributions and their (co)variances

This section establishes standard (co)variance results, not only for the multinomial, hypergeometric and Pólya draw distributions, but also for Poisson multinomials. We continue the approach of Section 4.6 and use inclusion functions $incl\colon \mathcal{N}[K](X) \hookrightarrow \mathcal{M}(X)$, from the set of K-sized natural multisets into the set of arbitrary multisets, so that we can exploit $\mathcal{M}(X)$'s cone structure (addition and scalar multiplication), as in Proposition 4.6.3.

In the literature one sometimes finds descriptions of such variances as vectors, but they presuppose an (unnecessary) ordering of the points of the underlying space as sequence. The multiset description given below does not require such an ordering. It forms an analogue of Proposition 4.6.3.

Proposition 5.2.1 *1 For a distribution $\omega \in \mathcal{D}(X)$,*

$$Var\Big(mn[K](\omega)\Big) = K \cdot \omega \cdot (\mathbf{1} - \omega) \in \mathcal{M}(X).$$

2 For a non-empty urn $\upsilon \in \mathcal{N}[L](X)$ of size $L \geq K$,

$$Var\Big(hg[K](\upsilon)\Big) = K \cdot \frac{L-K}{L-1} \cdot Flrn(\upsilon) \cdot (\mathbf{1} - Flrn(\upsilon)) \in \mathcal{M}(X).$$

3 For a non-empty urn $\upsilon \in \mathcal{N}(X)$ of size L,

$$Var\Big(pl[K](\upsilon)\Big) = K \cdot \frac{L+K}{L+1} \cdot Flrn(\upsilon) \cdot (\mathbf{1} - Flrn(\upsilon)).$$

4 For a distribution $\omega \in \mathcal{D}(X)$ and a rate $\lambda \in \mathbb{R}_{>0}$,

$$Var\Big(Pmn[\lambda](\omega)\Big) = \lambda \cdot \omega \in \mathcal{M}(X).$$

Proof. 1 We use the formulation of Lemma 5.1.3 and compute in $\mathcal{M}(X)$,

$$\begin{aligned}
Var\Big(mn[K](\omega)\Big) &= \Big(mn[K](\omega) \models incl^2\Big) - \Big(mn[K](\omega) \models incl\Big)^2 \\
&= \left(\sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi^2\right) - Mean\Big(mn[K](\omega)\Big)^2 \\
&= K \cdot (K-1) \cdot \omega^2 + K \cdot \omega - (K \cdot \omega)^2 \\
&\qquad \text{by Exercise 3.3.8 and Proposition 4.6.3 (1)} \\
&= K \cdot \omega \cdot (\mathbf{1} - \omega).
\end{aligned}$$

2 For a non-empty urn $\upsilon \in \mathcal{N}[L](X)$ of size $L \geq K$, we use Exercise 3.5.5 and Proposition 4.6.3 (2).

$$\begin{aligned}
Var\Big(hg[K](\upsilon)\Big) &= \left(\sum_{\varphi \leq_K \upsilon} hg[K](\upsilon)(\varphi) \cdot \varphi^2\right) - Mean\Big(hg[K](\upsilon)\Big)^2 \\
&= K \cdot Flrn(\upsilon) \cdot \frac{(K-1) \cdot \upsilon + (L-K)}{L-1} - \Big(K \cdot Flrn(\upsilon)\Big)^2 \\
&= K \cdot Flrn(\upsilon) \cdot \left(\frac{L-K}{L-1} + \frac{L \cdot (K-1) \cdot \upsilon - (L-1) \cdot K \cdot \upsilon}{L \cdot (L-1)}\right) \\
&= K \cdot Flrn(\upsilon) \cdot \left(\frac{L-K}{L-1} - \frac{(L-K) \cdot \upsilon}{L \cdot (L-1)}\right) = K \cdot \frac{L-K}{L-1} \cdot Flrn(\upsilon) \cdot (\mathbf{1} - Flrn(\upsilon)).
\end{aligned}$$

3 Similarly, for a non-empty urn $\upsilon \in \mathcal{N}[L](X)$, we now use Exercise 3.6.2 and Proposition 4.6.3 (3).

$$\begin{aligned}
Var\Big(pl[K](\upsilon)\Big) &= \left(\sum_{\varphi \in \mathcal{N}[K](supp(\upsilon))} pl[K](\upsilon)(\varphi) \cdot \varphi^2\right) - Mean\Big(pl[K](\upsilon)\Big)^2 \\
&= K \cdot Flrn(\upsilon) \cdot \frac{(K-1) \cdot \upsilon + (L+K)}{L+1} - \Big(K \cdot Flrn(\upsilon)\Big)^2 \\
&= K \cdot Flrn(\upsilon) \cdot \left(\frac{L+K}{L+1} + \frac{L \cdot (K-1) \cdot \upsilon - (L+1) \cdot K \cdot \upsilon}{L \cdot (L+1)}\right) \\
&= K \cdot Flrn(\upsilon) \cdot \left(\frac{L+K}{L+1} - \frac{(L+K) \cdot \upsilon}{L \cdot (L+1)}\right) = K \cdot \frac{L+K}{L+1} \cdot Flrn(\upsilon) \cdot (\mathbf{1} - Flrn(\upsilon)).
\end{aligned}$$

4 Finally, using Proposition 4.6.3 (4) together with Exercises 3.3.8, and 5.1.8,

$$\begin{aligned}
Var\Big(Pmn[\lambda](\omega)\Big) &= \Bigg(\sum_{\varphi\in\mathcal{N}(X)} Pmn[\lambda](\omega)(\varphi)\cdot\varphi^2\Bigg) - Mean\Big(Pmn[\lambda](\omega)\Big)^2 \\
&= \Bigg(\sum_{K\in\mathbb{N}} pois[\lambda](K)\cdot \sum_{\varphi\in\mathcal{N}[K](X)} mn[K](\omega)(\varphi)\cdot\varphi^2\Bigg) - \Big(\lambda\cdot\omega\Big)^2 \\
&= \Bigg(\sum_{K\in\mathbb{N}} pois[\lambda](K)\cdot\Big(K\cdot(K-1)\cdot\omega^2 + K\cdot\omega\Big)\Bigg) - \lambda^2\cdot\omega^2 \\
&= \Bigg(\sum_{K\in\mathbb{N}} pois[\lambda](K)\cdot K\cdot(K-1)\Bigg)\cdot\omega^2 + \Bigg(\sum_{K\in\mathbb{N}} pois[\lambda](K)\cdot K\Bigg)\cdot\omega - \lambda^2\cdot\omega^2 \\
&= \lambda^2\cdot\omega^2 + \lambda\cdot\omega - \lambda^2\cdot\omega^2 \;=\; \lambda\cdot\omega.
\end{aligned}$$

❑

Recall from (4.22) the multiset-evaluation observable $ev_x\colon \mathcal{M}(X) \to \mathbb{R}_{\geq 0}$ given by $ev_x(\varphi) = \varphi(x)$. The above formulations of variance as multisets can also be written as:

$$Var\Big(mn[K](\omega)\Big) = \sum_{x\in X} Var(mn[K](\omega), ev_x)\,\big|x\rangle \;\in\; \mathcal{M}(X). \tag{5.3}$$

(And similarly for $hg[K](\upsilon)$, $pl[K](\upsilon)$ and $Pmn[\lambda](\omega)$.)

There are anologous results and definitions for covariance.

Proposition 5.2.2 *Let X be a set with a two different elements $y \neq z$ from X and with a number $K \in \mathbb{N}$.*

1 For an urn-distribution $\omega \in \mathcal{D}(X)$,

$$Cov\Big(mn[K](\omega)\Big) := \sum_{y,z\in X} Cov(mn[K](\omega, ev_y, ev_z)\,\big|y,z\rangle \;=\; -K\cdot(\omega\otimes\omega) \;\in\; \mathcal{M}(X\times X).$$

2 For an urn $\upsilon \in \mathcal{N}[L](X)$ of size $L \geq K$,

$$\begin{aligned}
Cov\Big(hg[K](\upsilon)\Big) &:= \sum_{y,z\in X} Cov(hg[K](\upsilon), ev_y, ev_z)\,\big|y,z\rangle \\
&= -K\cdot\frac{L-K}{L-1}\cdot(Flrn(\upsilon)\otimes Flrn(\upsilon)) \;=\; -K\cdot\frac{L-K}{L-1}\cdot Flrn(\upsilon\otimes\upsilon) \;\in\; \mathcal{M}(X\times X).
\end{aligned}$$

This last equation comes from Lemma 2.3.2 (5).

3 In the Pólya case:

$$\begin{aligned}
Cov\Big(pl[K](\upsilon)\Big) &:= \sum_{y,z\in X} Cov(pl[K](\upsilon), ev_y, ev_z)\,\big|y,z\rangle \\
&= -K\cdot\frac{L+K}{L+1}\cdot(Flrn(\upsilon)\otimes Flrn(\upsilon)) \;=\; -K\cdot\frac{L+K}{L+1}\cdot Flrn(\upsilon\otimes\upsilon) \;\in\; \mathcal{M}(X\times X).
\end{aligned}$$

4 The Poisson point processes always has zero covariance:

$$Cov\Big(Pmn[\lambda](\omega)\Big) := \sum_{y,z\in X} Cov(Pmn[\lambda](\omega), ev_y, ev_z)\,\big|y,z\rangle \;=\; \mathbf{0} \;\in\; \mathcal{M}(X\times X).$$

Proof. 1 Using the covariance formula of Lemma 5.1.8 and Exercise 3.3.8 we compute:

$$\begin{aligned} Cov(mn[K](\omega),\, ev_y,\, ev_z) &= (mn[K](\omega) \models ev_y \,\&\, ev_z) - (mn[K](\omega) \models ev_y) \cdot (mn[K](\omega) \models ev_z) \\ &= K \cdot (K-1) \cdot \omega(y) \cdot \omega(z) - K \cdot \omega(y) \cdot K \cdot \omega(z) \\ &= -K \cdot \omega(y) \cdot \omega(z). \\ &= -K \cdot (\omega \otimes \omega)(y, z). \end{aligned}$$

2 In the same way, using Exercise 3.5.5 and Proposition 4.6.3 (2):

$$\begin{aligned} Cov(hg[K](\upsilon),\, ev_y,\, ev_z) &= (hg[K](\upsilon) \models ev_y \,\&\, ev_z) - (hg[K](\upsilon) \models ev_y) \cdot (hg[K](\upsilon) \models ev_z) \\ &= K \cdot (K-1) \cdot Flrn(\upsilon)(y) \cdot \frac{\upsilon(z)}{L-1} - K \cdot Flrn(\upsilon)(y) \cdot K \cdot Flrn(\upsilon)(z) \\ &= K \cdot Flrn(\upsilon)(y) \cdot \frac{L \cdot (K-1) \cdot \upsilon(z) - K \cdot (L-1) \cdot \upsilon(z)}{L \cdot (L-1)} \\ &= -K \cdot \frac{L-K}{L-1} \cdot Flrn(\upsilon)(y) \cdot Flrn(\upsilon)(z). \end{aligned}$$

3 Similarly, via Exercise 3.6.2 and Proposition 4.6.3 (3):

$$\begin{aligned} Cov(pl[K](\upsilon),\, ev_y,\, ev_z) &= (pl[K](\upsilon) \models ev_y \,\&\, ev_z) - (pl[K](\upsilon) \models ev_y) \cdot (pl[K](\upsilon) \models ev_z) \\ &= K \cdot (K-1) \cdot Flrn(\upsilon)(y) \cdot \frac{\upsilon(z)}{L+1} - K \cdot Flrn(\upsilon)(y) \cdot K \cdot Flrn(\upsilon)(z) \\ &= K \cdot Flrn(\upsilon)(y) \cdot \frac{L \cdot (K-1) \cdot \upsilon(z) - K \cdot (L+1) \cdot \upsilon(z)}{L \cdot (L+1)} \\ &= -K \cdot \frac{L+K}{L+1} \cdot Flrn(\upsilon)(y) \cdot Flrn(\upsilon)(z). \end{aligned}$$

4 Via Proposition 4.6.3 (4) together with Exercises 3.3.8, and 5.1.8 we get:

$$\begin{aligned} &Cov(Pmn[\lambda](\omega), ev_y, ev_z) \\ &= (Pmn[\lambda](\omega) \models ev_y \,\&\, ev_z) - (Pmn[\lambda](\omega) \models ev_y) \cdot (Pmn[\lambda](\omega) \models ev_z) \\ &= \left(\sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot (mn[K](\omega) \models ev_y \,\&\, ev_z) \right) \\ &\qquad - \left(\sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot (mn[K](\omega) \models ev_y) \right) \cdot \left(\sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot (mn[K](\omega) \models ev_z) \right) \\ &= \left(\sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot K \cdot (K-1) \cdot \omega(y) \cdot \omega(z) \right) \\ &\qquad - \left(\sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot K \cdot \omega(y) \right) \cdot \left(\sum_{K \in \mathbb{N}} pois[\lambda](K) \cdot K \cdot \omega(z) \right) \\ &= \lambda^2 \cdot \omega(y) \cdot \omega(z) - \lambda \cdot \omega(y) \cdot \lambda \cdot \omega(z) \;=\; 0. \end{aligned}$$

❑

On the diagonal in these joint multisets, where $y = z$, one recovers the variances. When the elements of the space X are ordered as lists, these covariance multisets can be seen as matrices. We elaborate an illustration.

Example 5.2.3 Consider a group of 50 people of which 25 vote for the green party (G), 10 vote liberal (L) and the remaining ones vote for the christian-democractic party (C). We thus have a set of vote options $V = \{G, L, C\}$ with a (natural) voter multiset $\nu = 25|G\rangle + 10|L\rangle + 15|C\rangle$.

We select five people from the group and look at their votes. These five people are obtained

in hypergeometric mode, where selected individuals step out of the group and are no longer available for subsequent selection.

The hypergeometric mean is introduced in Proposition 4.6.3 (2). It gives the multiset:

$$Mean\Big(hg[5](\nu)\Big) \;=\; 5 \cdot Flrn(\nu) \;=\; \tfrac{5}{2}|G\rangle + \tfrac{3}{2}|L\rangle + 1|C\rangle.$$

The deviations of the mean are given by the variance, see Proposition 5.2.1 (2), as a multiset on V.

$$\begin{aligned} Var\Big(hg[5](\nu)\Big) \;&=\; 5 \cdot \frac{50-5}{50-1} \cdot \sum_{x \in V} Flrn(\nu)(x) \cdot (1 - Flrn(\nu)(x)) \,\big|x\rangle \\ &=\; \tfrac{225}{196}|G\rangle + \tfrac{27}{28}|L\rangle + \tfrac{36}{49}|C\rangle \;\approx\; 1.148|G\rangle + 0.9643|L\rangle + 0.7347|C\rangle. \end{aligned}$$

The covariances give a 2-dimensional multiset on the product $V \times V$, see Proposition 5.2.2 (2).

$$\begin{aligned} Cov\Big(hg[5](\nu)\Big) \;=\; & \tfrac{225}{196}|G,G\rangle - \tfrac{135}{196}|G,L\rangle - \tfrac{45}{98}|G,C\rangle \\ & - \tfrac{135}{196}|L,G\rangle + \tfrac{27}{28}|L,L\rangle - \tfrac{27}{98}|L,C\rangle \\ & - \tfrac{45}{98}|C,G\rangle - \tfrac{27}{98}|C,L\rangle + \tfrac{36}{49}|C,C\rangle \\ \approx\; & 1.148|G,G\rangle - 0.6888|G,L\rangle - 0.4592|G,C\rangle \\ & - 0.6888|L,G\rangle + 0.9643|L,L\rangle - 0.2755|L,C\rangle \\ & - 0.4592|C,G\rangle - 0.2755|C,L\rangle + 0.7347|C,C\rangle. \end{aligned}$$

When these covariances are seen as a matrix, we recognise that the matrix is symmetric and has variances on its diagonal.

We recall from Definition 4.6.6 the extension of an observable p on a set X to an observable on natural multisets $\mathcal{N}(X)$ over X. This can be done additively and multiplicatively. Interestingly, the additive extension $\overline{p}^+$ interacts well with (co)variance in the multinomial case, like with validity in Proposition 4.6.7 (1). For convenience, we recall that $\overline{p}^+(\varphi) = \sum_x \varphi(x) \cdot p(x)$.

Proposition 5.2.4 *Let $p, q \colon X \to \mathbb{R}$ be observables on a set X, with their addtive extensions $\overline{p}^+, \overline{q}^+ \colon \mathcal{N}(X) \to \mathbb{R}$. For a distribution $\omega \in \mathcal{D}(X)$ and a number $K \in \mathbb{N}$,*

1 Variance of $\overline{p}^+$ over multinomial draws is related to variance of p over X, via:

$$Var\Big(mn[K](\omega),\, \overline{p}^+\Big) \;=\; K \cdot Var(\omega, p).$$

2 Similarly for covariance:

$$Cov\Big(mn[K](\omega),\, \overline{p}^+,\, \overline{q}^+\Big) \;=\; K \cdot Cov(\omega, p, q).$$

Proof. 1 By Proposition 4.6.7 (1) and Exercise 3.3.8:

$$\begin{aligned}
&Var\big(mn[K](\omega), \overline{p}^{+}\big) \\
&= mn[K](\omega) \models \overline{p}^{+} \;\&\; \overline{p}^{+} - (mn[K](\omega) \models \overline{p}^{+})^2 \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \overline{p}^{+}(\varphi) \cdot \overline{p}^{+}(\varphi) - K^2 \cdot (\omega \models p)^2 \\
&= \sum_{x,y \in X} p(x) \cdot p(y) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x) \cdot \varphi(y) - K^2 \cdot (\omega \models p)^2 \\
&= \sum_{x,y \in X} p(x) \cdot p(y) \cdot \left\{ \begin{array}{ll} K \cdot (K-1) \cdot \omega(x) \cdot \omega(y) & \text{if } x \neq y \\ K \cdot (K-1) \cdot \omega(x)^2 + K \cdot \omega(x) & \text{if } x = y \end{array} \right\} - K^2 \cdot (\omega \models p)^2 \\
&= K \cdot (K-1) \cdot (\omega \models p)^2 + K \cdot (\omega \models p \;\&\; p) - K^2 \cdot (\omega \models p)^2 \\
&= K \cdot (\omega \models p \;\&\; p) - K \cdot (\omega \models p)^2 \\
&= K \cdot Var(\omega, p).
\end{aligned}$$

2 Similarly. ❑

5.2.1 Distributions of validities and of variances

Fix a random variable (ω, p) on a set X and a number K. We can form the multinomial distribution $mn[K](\omega)$ on the set $\mathcal{N}[K](X)$ of natural multisets φ of size K, over X. Applying frequentist learning to such multisets φ gives new distributions $Flrn(\varphi) \in \mathcal{D}(X)$, and thus new random variables $(Flrn(\varphi), p)$. We can look a the validity and variance of the latter. This gives what we call *distributions of validity* and *distributions of variance*. Such distributions can be defined for the three modes of drawing: multinomial, hypergeometric and Pólya. Below we shall elaborate the multinomial case. The hypergeometric and Pólya cases are handled similarly and are left as exercises.

For the multinomial case we define two new distributions:

$$\begin{aligned}
mn\text{-}val[K](\omega, p) &:= \mathcal{D}(Flrn(-) \models p)\big(mn[K](\omega)\big) \\
&= \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \,\big|\, Flrn(\varphi) \models p \,\rangle \\
mn\text{-}var[K](\omega, p) &:= \mathcal{D}(Var(Flrn(-), p))\big(mn[K](\omega)\big) \\
&= \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \,\big|\, Var(Flrn(\varphi), p) \,\rangle.
\end{aligned} \tag{5.4}$$

Such distributions are useful in hypothesis testing in statistics, for instance when ω is a huge distribution for which we wish we check the validity of the observable (or predicate) p. We can then take small samples φ from ω via the multinomial distribution and check the validity of p in the normalised sample $Flrn(\varphi)$. Proposition 5.2.5 (1) below says that the mean of all such samples equals the validity $\omega \models p$.

We illustrate the above definitions (5.4). Consider a three-element set $X = \{a, b, c\}$ with distribution $\omega = \frac{1}{6}|a\rangle + \frac{1}{2}|b\rangle + \frac{1}{3}|c\rangle$ and observable $p = 6 \cdot \mathbf{1}_a + 4 \cdot \mathbf{1}_b + 12 \cdot \mathbf{1}_c$. It is not hard to see that:

$$\omega \models p = 7 \qquad \text{and} \qquad Var(\omega, p) = 13.$$

The distributions of validities is:

$$\begin{aligned}
\textit{mn-val}[2](\omega, p) &= \tfrac{1}{36}\big|\mathit{Flrn}(2|a\rangle) \models p\big\rangle + \tfrac{1}{6}\big|\mathit{Flrn}(1|a\rangle + 1|b\rangle) \models p\big\rangle + \tfrac{1}{4}\big|\mathit{Flrn}(2|b\rangle) \models p\big\rangle \\
&\quad + \tfrac{1}{9}\big|\mathit{Flrn}(1|a\rangle + 1|c\rangle) \models p\big\rangle + \tfrac{1}{3}\big|\mathit{Flrn}(1|b\rangle + 1|c\rangle) \models p\big\rangle + \tfrac{1}{9}\big|\mathit{Flrn}(2|c\rangle) \models p\big\rangle \\
&= \tfrac{1}{36}\big|1|a\rangle \models p\big\rangle + \tfrac{1}{6}\big|\tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle \models p\big\rangle + \tfrac{1}{4}\big|1|b\rangle \models p\big\rangle \\
&\quad + \tfrac{1}{9}\big|\tfrac{1}{2}|a\rangle + \tfrac{1}{2}|c\rangle \models p\big\rangle + \tfrac{1}{3}\big|\tfrac{1}{2}|b\rangle + \tfrac{1}{2}|c\rangle \models p\big\rangle + \tfrac{1}{9}\big|1|c\rangle \models p\big\rangle \\
&= \tfrac{1}{36}\big|6\big\rangle + \tfrac{1}{6}\big|5\big\rangle + \tfrac{1}{4}\big|4\big\rangle + \tfrac{1}{9}\big|9\big\rangle + \tfrac{1}{3}\big|8\big\rangle + \tfrac{1}{9}\big|12\big\rangle
\end{aligned}$$

It is not hard to see that its mean satisfies:

$$\tfrac{1}{36} \cdot 6 + \cdots + \tfrac{1}{9} \cdot 12 \;=\; 7 \;=\; \omega \models p.$$

As to the variance:

$$(\tfrac{1}{36} \cdot 6^2 + \cdots + \tfrac{1}{9} \cdot 12^2) - 7^2 \;=\; \tfrac{13}{2} \;=\; \tfrac{1}{2} \cdot \mathit{Var}(\omega, p).$$

These outcomes are in line with Proposition 5.2.5 (1) (2) below.

Next, the distribution of variances is:

$$\begin{aligned}
\textit{mn-var}[2](\omega, p) &= \tfrac{1}{36}\big|\mathit{Var}(1|a\rangle, p)\big\rangle + \tfrac{1}{6}\big|\mathit{Var}(\tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle, p)\big\rangle + \tfrac{1}{4}\big|\mathit{Var}(1|b\rangle, p)\big\rangle \\
&\quad + \tfrac{1}{9}\big|\mathit{Var}(\tfrac{1}{2}|a\rangle + \tfrac{1}{2}|c\rangle, p)\big\rangle + \tfrac{1}{3}\big|\mathit{Var}(\tfrac{1}{2}|b\rangle + \tfrac{1}{2}|c\rangle, p)\big\rangle + \tfrac{1}{9}\big|\mathit{Var}(1|c\rangle, p)\big\rangle \\
&= \tfrac{1}{36}\big|6^2 - 6^2\big\rangle + \tfrac{1}{6}\big|26 - 5^2\big\rangle + \tfrac{1}{4}\big|4^2 - 4^2\big\rangle + \tfrac{1}{9}\big|90 - 9^2\big\rangle \\
&\quad + \tfrac{1}{3}\big|80 - 8^2\big\rangle + \tfrac{1}{9}\big|12^2 - 12^2\big\rangle \\
&= \tfrac{7}{18}\big|0\big\rangle + \tfrac{1}{6}\big|1\big\rangle + \tfrac{1}{9}\big|9\big\rangle + \tfrac{1}{3}\big|16\big\rangle.
\end{aligned}$$

The mean $\frac{7}{18} \cdot 0 + \cdots + \frac{1}{3} \cdot 16$ of this distribution equals $\frac{13}{2} = \frac{1}{2} \cdot \mathit{Var}(\omega, p)$. This is an instance of Proposition 5.2.5 (3).

Diagrammatically the definitions (5.4) involve the composites:

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "mn[K]"] & \mathcal{D}(\mathcal{M}[K](X)) \arrow[r, "\mathcal{D}(\mathit{Flrn})"] & \mathcal{D}(\mathcal{D}(X)) \arrow[r, bend left, "\mathcal{D}(-\models p)"] \arrow[r, bend right, "\mathcal{D}(\mathit{Var}(-,p))"'] & \mathcal{D}(\mathbb{R})
\end{tikzcd} \tag{5.5}$$

The distribution of validities *mn-val*$[K]$ generalises the *distribution of means* that is often used in hypothesis testing in statistics (see *e.g.* Suhov and Kelbert (2005)). This distribution of means uses the validiy $(-) \models p$ to compute the mean, by taking the inclusion $\mathit{incl} \colon X \hookrightarrow \mathbb{R}$ as observable, as in Definition 4.2.1. This works of course only when X is a set of numbers. The above approach (5.4) with an arbitrary observable p is more general.

Once we have formed a distribution of validities / variances, we can ask what its validity / variance is. It turns out that these can be expressed in terms of validity / variance of the orginal random variable. These results resemble Theorem 3.3.4 which says that transforming a multinomial distribution along frequentist learning yields the orginal (urn) distribution.

Proposition 5.2.5 *Let (ω, p) be a random variable on a set X, with a number $K > 0$.*

1 The mean of the distribution of validities is the validity of the original random variable:

$$\mathit{Mean}\Big(\textit{mn-val}[K](\omega, p)\Big) \;=\; mn[K](\omega) \models \Big(\mathit{Flrn}(-) \models p\Big) \;=\; \omega \models p.$$

2 *The variance of the distribution of validities satisfies:*

$$Var\Big(mn\text{-}val[K](\omega, p)\Big) = \frac{Var(\omega, p)}{K}.$$

3 *The mean of the distribution of variances is:*

$$Mean\Big(mn\text{-}var[K](\omega, p)\Big) = mn[K](\omega) \models Var(Flrn(-), p) = \frac{K-1}{K} \cdot Var(\omega, p).$$

There is a fourth option to consider, namely the variance of the distribution of variances, but that doesn't seem to be very interesting.

Proof. 1 This is Proposition 4.6.1.

2 The second item requires more work. Let's assume $supp(\omega) = \{x_1, \ldots, x_n\}$. We first prove the auxiliary result $(*)$ below, via the Multinomial Theorem (1.43), and via Exercise 3.3.8. Usage of this exercise is denoted below by the marked equation $\overset{(E)}{=}$.

$$\sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \cdot \Big(Flrn(\varphi) \models p\Big)^2 = \frac{(K-1) \cdot (\omega \models p)^2 + (\omega \models p^2)}{K}. \qquad (*)$$

We reason as follows.

$$\begin{aligned}
&\sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \cdot \Big(Flrn(\varphi) \models p\Big)^2 \\
&= \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \Big(\textstyle\sum_i \frac{\varphi(x_i)}{K} \cdot p(x_i)\Big)^2 \\
&\overset{(1.43)}{=} \frac{1}{K^2} \cdot \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \cdot \sum_{\psi \in \mathcal{N}[2](\{1,\ldots,n\})} \binom{}{\psi} \cdot \textstyle\prod_i (\varphi(x_i) \cdot p(x_i))^{\psi(i)} \\
&= \frac{1}{K^2} \cdot \Bigg(2 \sum_{i \neq j} \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x_i) \cdot p(x_i) \cdot \varphi(x_j) \cdot p(x_j) \\
&\qquad + \sum_i \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \cdot \varphi(x_i)^2 \cdot p(x_i)^2 \Bigg) \\
&\overset{(E)}{=} \frac{1}{K^2} \cdot \Bigg(2 \sum_{i \neq j} K \cdot (K-1) \cdot \omega(x_i) \cdot p(x_i) \cdot \omega(x_j) \cdot p(x_j) \\
&\qquad + \sum_i K \cdot (K-1) \cdot \omega(x_i)^2 \cdot p(x_i)^2 + K \cdot \omega(x_i) \cdot p(x_i)^2 \Bigg) \\
&\overset{(1.43)}{=} \frac{K-1}{K} \cdot \big(\textstyle\sum_i \omega(x_i) \cdot p(x_i)\big)^2 + \frac{1}{K} \cdot \big(\textstyle\sum_i \omega(x_i) \cdot p(x_i)^2\big) \\
&= \frac{(K-1) \cdot (\omega \models p)^2 + (\omega \models p^2)}{K}.
\end{aligned}$$

Now we are ready to prove the formula for the variance of the distribution of validities in

item (2) in the proposition. We use item (1) and the auxiliarly equation $(*)$.

$$\begin{aligned}
Var\Big(mn\text{-}val[K](\omega,p)\Big) &= mn\text{-}val[K](\omega,p) \models id^2 - \Big(mn\text{-}val[K](\omega,p) \models id\Big)^2 \\
&= \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \cdot \Big(Flrn(\varphi) \models p\Big)^2 - \Big(Mean\Big(mn\text{-}val[K](\omega,p)\Big)\Big)^2 \\
&\overset{(*)}{=} \frac{(K-1)\cdot(\omega \models p)^2 + (\omega \models p^2)}{K} - (\omega \models p)^2 \\
&= \frac{-(\omega \models p)^2 + (\omega \models p^2)}{K} = \frac{Var(\omega,p)}{K}.
\end{aligned}$$

3 We use item (1) and $(*)$.

$$\begin{aligned}
Mean\Big(mn\text{-}var[K](\omega,p)\Big) &= \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \cdot Var(Flrn(\varphi),p) \\
&= \sum_{\varphi \in \mathcal{M}[K](X)} mn[K](\omega)(\varphi) \cdot \Big(Flrn(\varphi) \models p^2) - (Flrn(\varphi) \models p)^2\Big) \\
&\overset{(*)}{=} (\omega \models p^2) - \frac{(K-1)\cdot(\omega \models p)^2 + (\omega \models p^2)}{K} \\
&= \frac{(K-1)\cdot(\omega \models p^2)}{K} - \frac{(K-1)\cdot(\omega \models p)^2}{K} = \frac{K-1}{K} \cdot Var(\omega,p). \quad \square
\end{aligned}$$

Exercises

5.2.1 In the distribution of validities (5.4) we have used multinomial samples. We can also use hypergeometric ones, for an urn $\upsilon \in \mathcal{N}(X)$, an observable $p\colon X \to \mathbb{R}$ and a number $K \leq L = \|\upsilon\|$, via the definitions:

$$\begin{aligned}
hg\text{-}val[K](\upsilon,p) &:= \mathcal{D}(Flrn(-) \models p)\Big(hg[K](\upsilon)\Big) = \sum_{\varphi \leq_K \upsilon} hg[K](\upsilon)(\varphi) \,\big|\, Flrn(\varphi) \models p\rangle \\
hg\text{-}var[K](\upsilon,p) &:= \mathcal{D}(Var(Flrn(-),p))\Big(hg[K](\upsilon)\Big) = \sum_{\varphi \leq_K \upsilon} hg[K](\upsilon)(\varphi) \,\big|\, Var(Flrn(\varphi),p)\rangle.
\end{aligned}$$

Prove the following analogues of Proposition 5.2.5 (1) using the properties of the hypergeometric distribution from Lemma 3.5.5 (1) and Exercise 3.5.5.

1 $$\begin{aligned} Mean\Big(hg\text{-}val[K](\upsilon,p)\Big) &= hg[K](\upsilon) \models \Big(Flrn(-) \models p\Big) \\ &= Flrn(\upsilon) \models p. \end{aligned}$$

2 $$Var\Big(hg\text{-}val[K](\upsilon,p)\Big) = \frac{L-K}{(L-1)\cdot K} \cdot Var(Flrn(\upsilon),p).$$

This result corresponds to (Rice, 2006, §7.3.1, Thm. B).

3 $$Mean\Big(hg\text{-}var[K](\upsilon,p)\Big) = \frac{L\cdot(K-1)}{(L-1)\cdot K} \cdot Var(Flrn(\upsilon),p).$$

This corresponds to (Rice, 2006, §7.3.2, Thm. A).

5.2.2 The same definitions and similar results apply in the Pólya case. First, for an urn $\upsilon \in$

$\mathcal{N}[L](X)$ with full support and an observable $p\colon X \to \mathbb{R}$, we define:

$$\begin{aligned}
\textit{pl-val}[K](\upsilon, p) &:= \mathcal{D}(\textit{Flrn}(-) \models p)\big(\textit{pl}[K](\upsilon)\big) \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \textit{pl}[K](\upsilon)(\varphi) \,\big|\, \textit{Flrn}(\varphi) \models p \,\rangle \\
\textit{pl-var}[K](\upsilon, p) &:= \mathcal{D}(\textit{Var}(\textit{Flrn}(-), p))\big(\textit{pl}[K](\upsilon)\big) \\
&= \sum_{\varphi \in \mathcal{N}[K](X)} \textit{pl}[K](\upsilon)(\varphi) \,\big|\, \textit{Var}(\textit{Flrn}(\varphi), p) \,\rangle.
\end{aligned}$$

Prove, as in the previous exericse, now via Lemma 3.6.1 (1) and Exercise 3.6.2.

1 $\textit{Mean}\big(\textit{pl-val}[K](\upsilon, p)\big) = \textit{pl}[K](\upsilon) \models \big(\textit{Flrn}(-) \models p\big) = \textit{Flrn}(\upsilon) \models p.$

2 $\textit{Var}\big(\textit{pl-val}[K](\upsilon, p)\big) = \dfrac{L+K}{(L+1)\cdot K} \cdot \textit{Var}(\textit{Flrn}(\upsilon), p).$

3 $\textit{Mean}\big(\textit{pl-var}[K](\upsilon, p)\big) = \dfrac{L\cdot(K-1)}{(L+1)\cdot K} \cdot \textit{Var}(\textit{Flrn}(\upsilon), p).$

5.2.3 Let $p, q\colon X \to \mathbb{R}$ be two observables and let $K \in \mathbb{N}_{>0}$ be a draw size number. Let $\omega \in \mathcal{D}(X)$ be distribution and $\upsilon \in \mathcal{N}(X)$ be an urn of size $L = \|\upsilon\| > 0$. Prove consecutively:

1 $\textit{Cov}\big(\textit{mn}[K](\omega),\ \textit{Flrn}(-) \models p,\ \textit{Flrn}(-) \models q\big) = \dfrac{1}{K} \cdot \textit{Cov}(\omega, p, q).$

2 $\textit{Cov}\big(\textit{hg}[K](\upsilon),\ \textit{Flrn}(-) \models p,\ \textit{Flrn}(-) \models q\big) = \dfrac{(L-K)}{(L-1)\cdot K} \cdot \textit{Cov}(\omega, p, q),$ where $K \leq L$.

3 $\textit{Cov}\big(\textit{pl}[K](\upsilon),\ \textit{Flrn}(-) \models p,\ \textit{Flrn}(-) \models q\big) = \dfrac{(L+K)}{(L+1)\cdot K} \cdot \textit{Cov}(\omega, p, q).$

5.2.4 1 Prove Equation (5.3).

2 Formulate and prove similar equations for the three distributions $\textit{hg}[K](\upsilon)$, $\textit{pl}[K](\upsilon)$ and $\textit{Pmn}[\lambda](\omega)$.

5.2.5 Recall the parallel Poisson distribution $\textit{ppois}[\Lambda]$ from Definition 4.7.4, for a factor $\Lambda\colon X \to \mathbb{R}_{\geq 0}$. Show that (pointwise):

$$\textit{Mean}\big(\textit{ppois}[\Lambda]\big) = \Lambda = \textit{Var}\big(\textit{ppois}[\Lambda]\big).$$

5.3 Joint-state covariance and correlation

Section 5.1 has introduced variance and *shared-state* covariance / correlation. This section looks at a slightly different version, which we call *joint-state* covariance / correlation. It is subtly different from the shared-state version since the observables involved are defined not on the sample space of the shared underlying state, but on the components of a product space.

We will thus first define covariance and correlation for a joint state with observables on its component spaces, as a special case of what we have seen so far.

Definition 5.3.1 Let $\tau \in \mathcal{D}(X_1 \times X_2)$ be a joint state on sets X_1, X_2 and let $q_1 \in \textit{Obs}(X_1)$ and $q_2 \in \textit{Obs}(X_2)$ be two observables on these two sets *separately*. In this situation the *joint covariance* is defined as:

$$\textit{JCov}\big(\tau, q_1, q_2\big) := \textit{Cov}\big(\tau,\ (\pi_1)^*(q_1),\ (\pi_2)^*(q_2)\big).$$

Thus we use the weakenings $(\pi_1)^*(q_1) = q_1 \otimes \mathbf{1}$ and $(\pi_2)^*(q_2) = \mathbf{1} \otimes q_2$ to turn the observables

q_1, q_2 on different sets X_1, X_2 into observables on the same (product) set $X_1 \times X_2$ — so that Definition 5.1.7 applies.

Similarly, the *joint correlation* is:

$$JCor\big(\tau, q_1, q_2\big) \coloneqq Cor\big(\tau,\ (\pi_1)^*(q_1),\ (\pi_2)^*(q_2)\big).$$

In both these cases, if there are inclusions $X_1 \hookrightarrow \mathbb{R}$ and $X_2 \hookrightarrow \mathbb{R}$, then one can use these inclusions as random variables and write just $JCov(\tau)$ and $JCor(\tau)$.

Joint covariance can be reformulated in different ways, including in the style that we have seen before, in Lemma 5.1.3 and 5.1.8.

Lemma 5.3.2 *Joint covariance can be reformulated in terms of the two marginals $\tau[1,0]$ and $\tau[0,1]$ of the joint state $\tau \in \mathcal{D}(X_1 \times X_2)$, namely as:*

$$\begin{aligned} JCov\big(\tau, q_1, q_2\big) &= \tau \models \big(q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1}\big) \otimes \big(q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1}\big). \\ &= \big(\tau \models q_1 \otimes q_2\big) - \big(\tau[1,0] \models q_1\big) \cdot \big(\tau[0,1] \models q_2\big). \end{aligned}$$

As a consequence, when τ is non-entwined, the joint covariance is zero, whatever the obervables q_1, q_2 are:

$$JCov(\tau_1 \otimes \tau_2, q_1, q_2) = 0.$$

The situation with zero (joint) covariance is subtle and will be investigated further in the next section.

Proof. The first equation follows essentially from the relation between sequential and parallel conjunction & and $\otimes$, see Exercise 4.5.9.

$$\begin{aligned} &JCov(\tau, q_1, q_2) \\ &\quad = Cov(\tau,\ (\pi_1)^*(q_1),\ (\pi_2)^*(q_2)) \\ &\quad = \tau \models ((\pi_1)^*(q_1) - (\tau \models (\pi_1)^*(q_1)) \cdot \mathbf{1}) \ \&\ ((\pi_2)^*(q_2) - (\tau \models (\pi_2)^*(q_2)) \cdot \mathbf{1}) \\ &\quad \overset{(4.16)}{=} \tau \models ((\pi_1)^*(q_1) - (\tau[1,0] \models q_1) \cdot (\pi_1)^*(\mathbf{1})) \ \&\ ((\pi_2)^*(q_2) - (\tau[0,1] \models q_2) \cdot (\pi_2)^*(\mathbf{1}))) \\ &\quad = \tau \models \big((\pi_1)^*(q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1})\big) \ \&\ \big((\pi_2)^*(q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1})\big) \\ &\quad = \tau \models (q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1}) \otimes (q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1}). \end{aligned}$$

Via this first equation in Lemma 5.3.2 we prove the second one:

$$\begin{aligned} &JCov(\tau, q_1, q_2) \\ &= \tau \models (q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1}) \otimes (q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1}) \\ &= \sum_{x_1 \in X_1,\, x_2 \in X_2} \tau(x_1, x_2) \cdot \big(q_1 - (\tau[1,0] \models q_1) \cdot \mathbf{1}\big)(x_1) \cdot \big(q_2 - (\tau[0,1] \models q_2) \cdot \mathbf{1}\big)(x_2) \\ &= \sum_{x_1 \in X_1,\, x_2 \in X_2} \tau(x_1, x_2) \cdot \big(q_1(x_1) - (\tau[1,0] \models q_1)\big) \cdot \big(q_2(x_2) - (\tau[0,1] \models q_2)\big) \\ &= \left(\sum_{x_1 \in X_1,\, x_2 \in X_2} \tau(x_1, x_2) \cdot q_1(x_1) \cdot q_2(x_2)\right) - \left(\sum_{x_1 \in X_1,\, x_2 \in X_2} \tau(x_1, x_2) \cdot q_1(x_1) \cdot (\tau[0,1] \models q_2)\right) \\ &\qquad - \left(\sum_{x_1 \in X_1,\, x_2 \in X_2} \tau(x_1, x_2) \cdot (\tau[1,0] \models q_1) \cdot q_2(x_2)\right) + (\tau[1,0] \models q_1) \cdot (\tau[0,1] \models q_2) \\ &= (\tau \models q_1 \otimes q_2) - (\tau[1,0] \models q_1) \cdot (\tau[0,1] \models q_2). \end{aligned}$$

❑

We turn to some illustrations. Many examples of covariance are actually of joint form, especially if the underlying sets are subsets of the real numbers. In the joint case it makes sense to leave these inclusions implicit, as will be illustrated below.

Example 5.3.3 1 Consider sets $X = \{1, 2\}$ and $Y = \{1, 2, 3\}$ as subsets of $\mathbb{R}$, together with a joint distribution $\tau \in \mathcal{D}(X \times Y)$ given by:

$$\tau = \tfrac{1}{4}|1,1\rangle + \tfrac{1}{4}|1,2\rangle + \tfrac{1}{4}|2,2\rangle + \tfrac{1}{4}|2,3\rangle.$$

Its two marginals are:

$$\tau[1,0] = \tfrac{1}{2}|1\rangle + \tfrac{1}{2}|2\rangle \qquad\qquad \tau[0,1] = \tfrac{1}{4}|1\rangle + \tfrac{1}{2}|2\rangle + \tfrac{1}{4}|3\rangle.$$

Since both $X \hookrightarrow \mathbb{R}$ and $Y \hookrightarrow \mathbb{R}$ we get means as validities of the inclusions:

$$Mean(\tau[1,0]) = \tfrac{3}{2} \qquad\qquad Mean(\tau[1,0]) = 2.$$

Now we can compute the joint covariance in the joint state τ as:

$$\begin{aligned} JCov(\tau) &= \tau \models (incl - Mean(\tau[1,0]) \cdot \mathbf{1}) \otimes (incl - Mean(\tau[1,0]) \cdot \mathbf{1}) \\ &= \sum_{x \in X,\, y \in Y} \tau(x,y) \cdot (x - \tfrac{3}{2}) \cdot (y - 2) = \tfrac{1}{4}(-\tfrac{1}{2} \cdot -1 + \tfrac{1}{2} \cdot 1) = \tfrac{1}{4}. \end{aligned}$$

2 In order to calculate the (joint) correlation of τ we first need the variances of its marginals:

$$\begin{aligned} Var(\tau[1,0]) &= \sum_{x \in X} \tau[1,0](x) \cdot (x - \tfrac{3}{2})^2 = \tfrac{1}{4} \\ Var(\tau[0,1]) &= \sum_{y \in Y}, \tau[0,1](y) \cdot (y - 2)^2 = \tfrac{1}{2}. \end{aligned}$$

Then:

$$JCor(\tau) = \frac{JCov(\tau)}{\sqrt{Var(\tau[1,0])} \cdot \sqrt{Var(\tau[0,1])}} = \frac{1/4}{1/2 \cdot 1/\sqrt{2}} = \tfrac{1}{2}\sqrt{2}.$$

There are obvious choices for observables used in (co)variance and correlation when the underlying spaces are numbers. Otherwise, the choice of observable matters — except in the double binary case, considered below.

Proposition 5.3.4 *Let $\tau \in \mathcal{D}(\{a,b\} \times \{c,d\})$ be a binary distribution on a product of binary, two-element sets. The joint correlation is then almost independent of the choice of observable, in the following sense: for non-constant observables $p_1, q_1 \colon \{a,b\} \to \mathbb{R}$ and $p_2, q_2 \colon \{c,d\} \to \mathbb{R}$ one has:*

$$\begin{aligned} JCor\big(\tau, p_1, p_2\big) &= \pm JCor\big(\tau, q_1, q_2\big) \\ &= \begin{cases} JCor\big(\tau, q_1, q_2\big) & \text{if } (q_1 - p_1)(a) \geq (q_1 - p_1)(b) \\ & \qquad \iff (q_2 - p_2)(c) \geq (q_2 - p_2)(d) \\ -JCor\big(\tau, q_1, q_2\big) & \text{otherwise.} \end{cases} \end{aligned}$$

Proof. We define numbers u_1, v_1 as:

$$u_1 := \frac{q_1(a) - q_1(b)}{p_1(a) - p_1(b)} \qquad\qquad v_1 := \frac{q_2(c) - q_2(d)}{p_2(c) - p_2(d)}.$$

And and u_2, v_2 as:

$$\begin{aligned} u_2 &:= q_1(a) - u_1 \cdot p_1(a) = q_1(b) - u_1 \cdot p_1(b) \\ v_2 &:= q_2(c) - v_1 \cdot p_2(c) = q_2(d) - v_1 \cdot p_2(d). \end{aligned}$$

Then: $q_1 = u_1 \cdot p_1 + u_2 \cdot \mathbf{1}$ and $q_2 = v_1 \cdot p_2 + v_2 \cdot \mathbf{1}$. Hence by Theorem 5.1.10 (3):

$$\begin{aligned} JCor\big(\tau,\, q_1,\, q_2\big) &= JCor\big(\tau,\, u_1 \cdot p_1 + u_2 \cdot \mathbf{1},\, v_1 \cdot p_2 + v_2 \cdot \mathbf{1}\big) \\ &= Cor\big(\tau,\, (\pi_1)^*(u_1 \cdot p_1 + u_2 \cdot \mathbf{1}),\, (\pi_2)^*(v_1 \cdot p_2 + v_2 \cdot \mathbf{1})\big) \\ &= Cor\big(\tau,\, u_1 \cdot (\pi_1)^*(p_1) + u_2 \cdot \mathbf{1},\, v_1 \cdot (\pi_2)^*(p_2) + v_2 \cdot \mathbf{1}\big) \\ &= Cor\big(\tau,\, u_1 \cdot (\pi_1)^*(p_1),\, v_1 \cdot (\pi_2)^*(p_2)\big) \\ &= \begin{cases} Cor\big(\tau,\, (\pi_1)^*(p_1),\, (\pi_2)^*(p_2)\big) & \text{if } u_1, v_1 \text{ have the same sign} \\ -Cor\big(\tau,\, (\pi_1)^*(p_1),\, (\pi_2)^*(p_2)\big) & \text{otherwise} \end{cases} \\ &= \begin{cases} JCor\big(\tau,\, p_1,\, p_2\big) & \text{if } u_1, v_1 \text{ have the same sign} \\ -JCor\big(\tau,\, p_1,\, p_2\big) & \text{otherwise.} \end{cases} \end{aligned}$$

The numbers u_1, v_1 have the same sign when $(q_1 - p_1)(a) \geq (q_1 - p_1)(b) \iff (q_2 - p_2)(c) \geq (q_2 - p_2)(d)$. ❑

When one draws two balls from an urn, in multinomial mode, with replacement, two draws are independent, in an inuitive manner, since they are drawn from the same urn. If we draw in hypergeometric (draw-delete) or Pólya (draw-add) mode the first draw influences the second one. This influence can be expressed in terms of covariance, but how to do this precisely is subtle.

In (Rice, 2006, §7.3.1, Lem. B) the phenomenon is articulated via an equation $Cov(X_i, X_j) = -\sigma^2/(L-1)$, for $i \neq j$, where σ^2 is the variance of an implicit distribution (on the real numbers) obtained from a population of size L. There is a side-remark that sampling is done without replacement, where the X_i and X_j denote the two draws. This side-remark refers to another distribution that is implicit, namely in the expression $Cov(X_i, X_j)$. Below we make the equation precise and more general: we make both implicit distributions, on the left and on the right, explicit and generalise to an arbitrary sample space (instead of the reals), with two observables p, q and with the appropriate (joint and shared) state covariance. The variance σ^2 then arises as special case when $p = q$, since $Var(\omega, p) = Cov(\omega, p, p)$. This leads to the statement below, involving hypergeometric draws of size two. The Pólya analogue appears in Exercise 5.3.6.

Proposition 5.3.5 *Let $\upsilon \in \mathcal{N}[L](X)$ be an urn of size $L \geq 2$ and let $p, q \colon X \to \mathbb{R}$ be two observables. Then:*

$$JCov\Big(arr_*(hg[2](\upsilon)),\, p,\, q\Big) = \frac{-Cov(Flrn(\upsilon),\, p,\, q)}{L-1}.$$

The distribution on the left-hand-side arises as pushforward of the hypergeometric distribution $hg[2](\upsilon) \in \mathcal{D}(\mathcal{N}[2](X))$ for draws of size two, along the arrange channel $arr \colon \mathcal{N}[2](X) \rightsquigarrow X \times X$. The result gives sequences, instead of multisets, of hypergeometric draws. This is the 'sequence' version $seqhg[2]$ of the hypergeometric distribution from Exercise 2.6.12, explicitly described for two draws in Exercise 2.2.5.

Notice that if the urn size L goes to infinity, the joint covariance becomes zero. This makes sense, since a first draw from a very large urn hardly influences a second draw.

Proof. Let us abbreviate $\tau = arr_*(hg[2](\upsilon)) \in \mathcal{D}(X \times X)$. It is not hard to see that we can describe it explicitly, like in Theorem 2.6.2, as:

$$\tau = \sum_{x \in supp(\upsilon)} \sum_{y \in supp(\upsilon - 1|x\rangle)} Flrn(\upsilon)(x) \cdot Flrn(\upsilon - 1|x\rangle)(y) \,\big|x, y\rangle.$$

The validities of the weakened observables $p \circ \pi_1, q \circ \pi_2 \colon X \times X \to \mathbb{R}$ are obtained via Exercise 2.2.5 and Lemma 4.1.5:

$$\begin{aligned} \tau \models p \circ \pi_1 &= \mathcal{D}(\pi_1)(\tau) \models p = Flrn(\upsilon) \models p \\ \tau \models q \circ \pi_2 &= \mathcal{D}(\pi_2)(\tau) \models p = Flrn(\upsilon) \models q. \end{aligned}$$

Next we calculate the validity of the parallel product $p \otimes q = (p \circ \pi_1) \,\&\, (q \circ \pi_2)$ of observables:

$$\begin{aligned} \tau \models p \otimes q &= \sum_{x \in supp(\upsilon)} \sum_{y \in supp(\upsilon - 1|x\rangle)} \frac{\upsilon(x)}{L} \cdot \frac{(\upsilon - 1|x\rangle)(y)}{L-1} \cdot p(x) \cdot q(y) \\ &= \sum_{x \in supp(\upsilon)} \frac{\upsilon(x) \cdot p(x)}{L} \cdot \left(\left(\sum_{y \in supp(\upsilon)} \frac{\upsilon(y) \cdot q(y)}{L-1} \right) - \frac{q(x)}{L-1} \right) \\ &= \sum_{x \in supp(\upsilon)} \frac{\upsilon(x) \cdot p(x)}{L} \cdot \left(\frac{L}{L-1} \cdot (Flrn(\upsilon) \models q) - \frac{q(x)}{L-1} \right) \\ &= \frac{L}{L-1} \cdot (Flrn(\upsilon) \models p) \cdot (Flrn(\upsilon) \models q) - \frac{1}{L-1} \cdot (Flrn(\upsilon) \models p \,\&\, q) \end{aligned}$$

We can now piece things together:

$$\begin{aligned} &JCov\big(arr_*(hg[2](\upsilon)),\ p,\ q\big) \\ &= Cov\big(\tau,\ p \circ \pi_1,\ q \circ \pi_2\big) \\ &= \big(\tau \models (p \circ \pi_1) \,\&\, (q \circ \pi_2)\big) - \big(\tau \models p \circ \pi_1\big) \cdot \big(\tau \models q \circ \pi_2\big) \\ &= \left(\frac{L}{L-1} - 1\right) \cdot (Flrn(\upsilon) \models p) \cdot (Flrn(\upsilon) \models q) - \frac{1}{L-1} \cdot (Flrn(\upsilon) \models p \,\&\, q) \\ &= \frac{1}{L-1} \cdot (Flrn(\upsilon) \models p) \cdot (Flrn(\upsilon) \models q) - \frac{1}{L-1} \cdot (Flrn(\upsilon) \models p \,\&\, q) \\ &= \frac{-Cov(Flrn(\upsilon),\ p,\ q)}{L-1}. \end{aligned}$$

❑

In the multinomial version of Proposition 5.3.5 drawing is done with replacement, so that two consecutive draws yield a non-entwined distribution of the form: $arr_*(mn[2](\omega)) = \omega \otimes \omega$, see (3.16). This distribution has (joint state) covariance equal to zero, see Lemma 5.3.2.

We have defined joint covariance as a special case of ordinary covariance. We now show that shared-state covariance can also be seen as joint-state covariance, namely for a copied state. Recall that copying of states is a subtle matter, since $\Delta_*(\omega) \neq \omega \otimes \omega$, in general, see Subsection 2.3.2.

Proposition 5.3.6 *Shared state covariance can be expressed as joint covariance via copying:*

$$Cov\big(\omega, p_1, p_2\big) = JCov\big(\Delta_*(\omega), p_1, p_2\big).$$

More generally, for suitably typed channels c, d,

$$Cov\big(\omega,\ c^*(p),\ d^*(q)\big) = JCov\big(\langle c, d\rangle_*(\omega),\ p,\ q\big).$$

The latter equation is reminiscent of Proposition 4.5.3, with validity $\models$ being maintained under a swap of state / predicate transformation.

Proof. We prove the second equation since the first one is a special case, namely when c, d are identity channels. Via Lemma 5.3.2 we get:

$$\begin{aligned}
JCov(\langle c,d\rangle_*(\omega), p, q) &= \langle c,d\rangle_*(\omega) \models (p - (\langle c,d\rangle_*(\omega)[1,0] \models p)\cdot \mathbf{1}) \\
&\qquad\qquad \otimes (q - (\langle c,d\rangle_*(\omega)[0,1] \models q)\cdot \mathbf{1}) \\
&= \omega \models \langle c,d\rangle^*\big((p - (c_*(\omega) \models p)\cdot \mathbf{1}) \otimes (q - (d_*(\omega) \models q)\cdot \mathbf{1})\big) \\
&= \omega \models \big(c^*(p - (\omega \models c^*(p))\cdot \mathbf{1})\big) \mathbin{\&} \big(d^*(q - (\omega \models d^*(q))\cdot \mathbf{1})\big) \\
&\qquad \text{by Lemma 4.5.2 (7)} \\
&= \omega \models (c^*(p) - (\omega \models c^*(p)\cdot \mathbf{1}) \mathbin{\&} (d^*(q) - (\omega \models d^*(q)\cdot \mathbf{1}) \\
&\qquad \text{since } c^* \text{ is linear and preserves } \mathbf{1} \\
&= Cov(\omega, c^*(p), d^*(q)).
\end{aligned}$$

❑

We started with 'ordinary' covariance in Definition 5.1.7 for two random variables with a shared state. It was subsequently used to define the 'joint' version in Definition 5.3.1, with two random variable defined on the components. The above result shows that we could have done this the other way around too: obtain the ordinary formulation in terms of the joint version via the copied state $\Delta_*(\omega)$. As we shall see below, there are notable differences between shared-state and joint-state versions, see Proposition 5.4.3 and Theorem 5.4.6 in the next section.

The difference between *Cov* and *JCov* is subtle, as the next (folklore) result illustrates. It does involve a joint state, with random variables defined via projection, but with an overlap, so that we still formulate the result as ordinary covariance *Cov*.

Lemma 5.3.7 *Consider three Poisson rates $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}_{\geq 0}$ with associated Poisson distributions $\pi_i := pois[\lambda_i]$, combined in a product $\pi := \pi_1 \otimes \pi_2 \otimes \pi_3 \in \mathcal{D}_\infty(\mathbb{N}^3)$. Consider the two random variables $q_1, q_2 \colon \mathbb{N}^3 \to \mathbb{N}$ defined as:*

$$q_1(n,m,k) := n + m \qquad \textit{and} \qquad q_2(n,m,k) := n + k.$$

Then:

$$Cov(\pi, q_1, q_2) = \lambda_1.$$

Proof. We first calculate the validity below, using the mean of a Poisson distribution, see Example 4.2.2 (3).

$$\begin{aligned}
\pi \models q_1 = \sum_{n,m,k\in\mathbb{N}} \pi_1(n)\cdot\pi_2(m)\cdot\pi_3(k)\cdot(n+m) &= \sum_{n,m\in\mathbb{N}} pois[\lambda_1](n)\cdot pois[\lambda_2](m)\cdot(n+m) \\
&= \left(\sum_{n\in\mathbb{N}} pois[\lambda_1](n)\cdot n\right) + \left(\sum_{m\in\mathbb{N}} pois[\lambda_2](m)\cdot m\right) \\
&= Mean(pois[\lambda_1]) + Mean(pois[\lambda_2]) = \lambda_1 + \lambda_2.
\end{aligned}$$

Similarly we get $\pi \models q_2 = \lambda_2 + \lambda_3$. Next, via Exercise 5.1.8 (2),

$$\begin{aligned}
\pi \models q_1 \,\&\, q_2 &= \sum_{n,m,k\in\mathbb{N}} \pi_1(n)\cdot\pi_2(m)\cdot\pi_3(k)\cdot(n+m)\cdot(n+k) \\
&= \sum_{n,m,k\in\mathbb{N}} pois[\lambda_1](n)\cdot pois[\lambda_2](m)\cdot pois[\lambda_3](k)\cdot(n^2+n\cdot m+n\cdot k+m\cdot k) \\
&= \left(\sum_{n\in\mathbb{N}} pois[\lambda_1](n)\cdot n^2\right) + \left(\sum_{n,m\in\mathbb{N}} pois[\lambda_1](n)\cdot pois[\lambda_2](m)\cdot n\cdot m\right) \\
&\quad + \left(\sum_{n,k\in\mathbb{N}} pois[\lambda_1](n)\cdot pois[\lambda_3](k)\cdot n\cdot k\right) + \left(\sum_{m,k\in\mathbb{N}} pois[\lambda_2](m)\cdot pois[\lambda_3](k)\cdot m\cdot k\right) \\
&= \lambda_1^2 + \lambda_1 + \lambda_1\cdot\lambda_2 + \lambda_1\cdot\lambda_3 + \lambda_2\cdot\lambda_3.
\end{aligned}$$

Finally we can compute:

$$\begin{aligned}
Cov(\pi, q_1, q_2) &= \big(\pi \models q_1 \,\&\, q_2\big) - \big(\pi \models q_1\big)\cdot\big(\pi \models q_2\big) \\
&= \big(\lambda_1^2 + \lambda_1 + \lambda_1\cdot\lambda_2 + \lambda_1\cdot\lambda_3 + \lambda_2\cdot\lambda_3\big) - \big(\lambda_1+\lambda_2\big)\cdot\big(\lambda_1+\lambda_3\big) = \lambda_1. \qquad ❑
\end{aligned}$$

We formulate a joint-state analogue for the linearity properties of Theorem 5.1.10.

Theorem 5.3.8 *Consider a state $\tau \in \mathcal{D}(X_1 \times X_2)$, with observables $q_1, q_3 \in Obs(X_1)$, $q_2 \in Obs(X_2)$ and numbers $r, s \in \mathbb{R}$.*

1 Joint-state covariance satisfies:

$$\begin{aligned}
JCov(\tau, q_1, q_2) &= JCov(\tau, q_2, q_1) \\
JCov(\tau, q_1, \mathbf{1}) &= 0 \\
JCov(\tau, r\cdot q_1, q_2) &= r\cdot JCov(\tau, q_1, q_2) \\
JCov(\omega, q_1 + q_3, q_2) &= JCov(\tau, q_1, q_2) + Cov(\tau, q_3, q_2). \\
JCov(\tau, q_1 + r\cdot\mathbf{1}, q_2 + s\cdot\mathbf{1}) &= JCov(\tau, q_1, q_2).
\end{aligned}$$

2 For joint-state correlation we have:

$$\begin{aligned}
JCor(\tau, r\cdot q_1, s\cdot q_2) &= \begin{cases} JCor(\tau, q_1, q_2) & \text{if } r, s \text{ have the same sign} \\ -JCor(\tau, q_1, q_2) & \text{otherwise.} \end{cases} \\
JCor(\tau, q_1 + r\cdot\mathbf{1}, q_2 + s\cdot\mathbf{1}) &= JCor(\omega, q_1, q_2).
\end{aligned}$$

Proof. This follows directly from Theorem 5.1.10, using that predicate transformation $(\pi_i)^*$ is linear and thus preserves sums and scalar multiplications (and also truth), see Lemma 4.5.2 (2). ❑

We conclude this section with a medical example about the correlation between disease and test.

Example 5.3.9 We start with a space $D = \{d, d^\perp\}$ for occurence of a disease or not (for a particular person) and a space $T = \{p, n\}$ for a positive or negative test outcome. Prevalence is used to indicate the prior likelihood of occurrence of the disease, for instance in the whole population, before a test. It can be described via a flip-like channel:

$$[0,1] \xrightarrow{prev} D \quad \text{with} \quad prev(r) := r|d\rangle + (1-r)|d^\perp\rangle.$$

We assume that there is a test for the disease with the following characteristics.

- ('sensitivity') If someone has the disease, then the test is positive with probability of 90%.
- ('specificity') If someone does not have the disease, there is a 95% chance that the test is negative.

We formalise this via a channel $test\colon D \to\!\!\!\!\!\!\circ\; T$ with:

$$test(d) \;=\; \tfrac{9}{10}|p\rangle + \tfrac{1}{10}|n\rangle \qquad\qquad test(d^\perp) \;=\; \tfrac{1}{20}|p\rangle + \tfrac{19}{20}|n\rangle.$$

We can now form the joint 'graph' state:

$$joint(r) \;\coloneqq\; \langle id, test\rangle_*(prev(r)) \;\in\; \mathcal{D}(D \times T).$$

Exercise 5.3.4 below tells that it does not really matter which observables we choose, so we simply take $\mathbf{1}_d\colon D \to [0,1]$ and $\mathbf{1}_p\colon T \to [0,1]$, mapping d and p to 1 and $d^\perp$ and n to 0. We are thus interested in (joint-state) correlation function:

$$[0,1] \;\ni\; r \longmapsto JCor\Big(joint(r),\, \mathbf{1}_d,\, \mathbf{1}_p\Big) \;\in\; [-1,1].$$

This is plotted in Figure 5.2, on the left. We see that, with the sensitivity and specificity values as given above, there is a clear positive correlation beteen disease and test, but less so in the corner cases with minimal and maximal prevalence.

We now fix a prevalence of 20% and wish to understand correlation as a function of sensitivity and specificity. We thus parameterise the above test channel to $test(se, sp)\colon D \to\!\!\!\!\!\!\circ\; T$ with parameters $se, sp \in [0,1]$.

$$\begin{aligned} test(se,sp)(d) \;&=\; se|p\rangle + (1-se)|n\rangle \\ test(se,sp)(d^\perp) \;&=\; (1-sp)|p\rangle + sp|n\rangle. \end{aligned}$$

As before we form a joint state, but now with different parameters.

$$joint(se, sp) \;\coloneqq\; \langle id, test(se,sp)\rangle_*(prev(\tfrac{1}{5})) \;\in\; \mathcal{D}(D \times T).$$

We are then interested in the function:

$$[0,1]\times[0,1] \;\ni\; (se, sp) \longmapsto JCor\Big(joint(se,sp),\, \mathbf{1}_d,\, \mathbf{1}_p\Big) \;\in\; [-1,1].$$

It is described on the right in Figure 5.2. We see that with maximal sensitivity and specificity (both 1) the correlation between disease and test is also maximal (actually 1), and dually with minimal sensitivity and specificity (both 0) the correlation is minimal (namely -1). These (unrealistic) extremes correspond to an optimal test and an inferior one.

Exercise 5.3.5 makes some intuitive properties of this (parameterised) correlation between disease and test explicit.

Exercises

5.3.1 Find examples of covariance and correlation computations in the literature (or online) and determine if they are of shared-state or joint-state form.

5.3.2 Prove that:

$$JCor(\tau, q_1, q_2) \;=\; \frac{JCov(\tau, q_1, q_2)}{StDev(\tau[1,0], q_1) \cdot StDev(\tau[0,1], q_2)}.$$

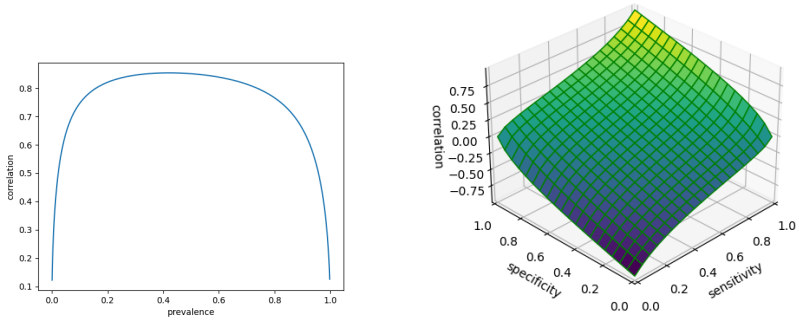


Figure 5.2 Disease-Test correlations, on the left as a function of prevalence (with fixed sensitivity and specificity) and on the right as a function of sensitivity and specificity (with fixed prevalence), see Example 5.3.9 for details.

5.3.3 Consider distributions $\tau_i \in \mathcal{D}(X_i)$ with observables $p_i \in Obs(X_i)$ for $i = 1, 2$. Use Theorem 5.1.10 (2) to prove that:

$$Var\Big(\tau_1 \otimes \tau_2,\ (\pi_1)^*(p_1) + (\pi_2)^*(p_2)\Big) \ = \ Var(\tau_1, p_1) + Var(\tau_2, p_2),$$

where the observable $(\pi_1)^*(p_1) + (\pi_2)^*(p_2)\colon X_1 \times X_2 \to \mathbb{R}$ sends a pair (x_1, x_2) to $p_1(x_1) + p(x_2)$.

5.3.4 Consider two two-element sample spaces $X_1 = \{a_1, b_1\}$ and $X_2 = \{a_2, b_2\}$ with a joint state $\tau \in \mathcal{D}(X_1 \times X_2)$. Prove that in this binary case joint-state covariance and correlation do not depend on the observables, that is:

$$JCor(\tau, p_1, p_2) \ = \ \pm JCor(\tau, q_1, q_2),$$

for all *non-constant* observables $p_1, q_1 \in Obs(X_1)$, $p_2, q_2 \in Obs(X_2)$.

Hint: Use Theorem 5.3.8 to massage p_1, p_2 to the observables which send a_1, b_1 to 0 and a_2, b_2 to 1.

5.3.5 Prove in the context of Example 5.3.9,

1 for each $s \in [0, 1]$ one has:

$$JCor\Big(joint(s, 1 - s,\ \mathbf{1}_d,\ \mathbf{1}_p\Big) \ = \ 0;$$

2 for all $se, sp \in [0, 1]$,

$$se + sp \geq 1 \iff JCor\Big(joint(se, sp),\ \mathbf{1}_d,\ \mathbf{1}_p\Big) \geq 0.$$

5.3.6 Prove the following Pólya analogue of Proposition 5.3.5.

$$JCov\Big(arr_*(pl[2](\upsilon)),\ p,\ q\Big) \ = \ \frac{Cov(Flrn(\upsilon),\ p,\ q)}{L+1}.$$

5.4 Independence for random variables

Earlier we have called a joint state / distribution entwined when it differs from the product of its marginals. This may be called dependence, but the word 'dependence' is standardly used for random variables. Such dependence (or independence) is the topic of this section. We shall follow the approach of the previous two sections and introduce two versions of dependence, also called *shared-state* and *joint-state*. Independence is related to the property 'covariance is zero', but in a subtle manner. This will be elaborated below.

Suppose we have two random variables describing the number of icecream sales and the temperature. Intuitively one expects a dependence between the two, and that the two variables are correlated (in an informal sense). The opposite, namely *independence* is usually formalised as follows. Two random variables p_1, p_2 are called independent if the equation,

$$P[p_1 = a, p_2 = b] \;=\; P[p_1 = a] \cdot P[p_2 = b]. \tag{5.6}$$

holds for all real numbers a, b. In this formulation a distribution is assumed in the background. We like to use it explicitly. How should the above equation (5.6) then be read?

As we described in Section 4.3, the expression $P[p = a]$ can be interpreted as the image distribution of the random variable, applied to $a \in \mathbb{R}$. More explicitly, it is obtained via transformation of the implicit distribution $\omega \in \mathcal{D}(X)$ along the observable $p\colon X \to \mathbb{R}$, considered as deterministic channel:

$$P[p = a] \;:=\; \mathcal{D}(p)(\omega)(a) \;=\; p_*(\omega) \models \mathbf{1}_a \;\overset{(4.15)}{=}\; \omega \models p^*(\mathbf{1}_a).$$

We can then translate the above requirement in Equation (5.6) into the condition:

$$\langle p_1, p_2\rangle_*(\omega) \;=\; \mathcal{D}(\langle p_1, p_2\rangle)(\omega) \;=\; \mathcal{D}(p_1)(\omega) \otimes \mathcal{D}(p_2)(\omega) \;=\; p_{1*}(\omega) \otimes p_{2*}(\omega). \tag{5.7}$$

But this says that the joint state $\langle p_1, p_2\rangle_*(\omega)$ on $\mathbb{R} \times \mathbb{R}$, transformed along $\langle p_1, p_2\rangle\colon X \to \mathbb{R} \times \mathbb{R}$, is non-entwined: it is the product of its marginals. Indeed, its (first) marginal is:

$$\langle p_1, p_2\rangle_*(\omega)[1, 0] \;=\; (\pi_1)_*(\langle p_1, p_2\rangle_*(\omega)) \;=\; (\pi_1 \circ \langle p_1, p_2\rangle)_*(\omega) \;=\; (p_1)_*(\omega).$$

This brings us to the following definition. We formulate it for two random variables, but it can easily be extended to n-ary form.

Definition 5.4.1 Let (ω, p_1) and (ω, p_2) be two random variables with a common, shared state ω. These random variables will be called *independent* if the transformed state $\langle p_1, p_2\rangle_*(\omega)$ on $\mathbb{R} \times \mathbb{R}$ is non-entwined, as in Equation (5.7): it is required to be the product of its marginals, as in:

$$\langle p_1, p_2\rangle_*(\omega) \;=\; p_{1*}(\omega) \otimes p_{2*}(\omega).$$

We sometimes call this the *shared-state* form of independence, in order to distinguish it from a later *joint-state* version.

We give an illustration, of non-independence, that is, of dependence.

Example 5.4.2 Consider a fair coin $\mathit{flip} = \frac{1}{2}|1\rangle + \frac{1}{2}|0\rangle$. We are going to use it twice, first to determine how much we will bet (either €100 or €50), and secondly to determine whether the bet is won or not. Thus we use the distribution $\omega = \mathit{flip} \otimes \mathit{flip}$ with underlying set 2×2, where $2 = \{0, 1\}$. We will define two observables $p_1, p_2\colon 2 \times 2 \to \mathbb{R}$, to be used as random variables for this same, shared distribution ω.

We first define an auxiliary observable $p\colon 2 \to \mathbb{R}$ for the amount of the bet:

$$p(1) = 100 \qquad\qquad p(0) = 50.$$

We then define $p_1 = p \otimes \mathbf{1} = p \circ \pi_1 \colon 2 \times 2 \to \mathbb{R}$, via weakening, as on the left below. The observable p_2 is defined on the right.

$$p_1(x,y) := p(x) = \begin{cases} 100 & \text{if } x = 1 \\ 50 & \text{if } x = 0 \end{cases} \qquad\qquad p_2(x,y) := \begin{cases} p(x) & \text{if } y = 1 \\ -p(x) & \text{if } y = 0. \end{cases}$$

We claim that (ω, p_1) and (ω, p_2) are *not* independent. Intuitively this may be clear, since the observable p forms a connection between p_1 and p_2. Formally, we can see this by doing the calculations. First we find out what the joint state is:

$$\begin{aligned} \langle p_1, p_2\rangle_*(\omega) &= \mathcal{D}(\langle p_1, p_2\rangle)(\mathit{flip} \otimes \mathit{flip}) \\ &= \tfrac{1}{4}\big|p_1(1,1), p_2(1,1)\big\rangle + \tfrac{1}{4}\big|p_1(1,0), p_2(1,0)\big\rangle \\ &\qquad + \tfrac{1}{4}\big|p_1(0,1), p_2(0,1)\big\rangle + \tfrac{1}{4}\big|p_1(0,0), p_2(0,0)\big\rangle \\ &= \tfrac{1}{4}\big|100, 100\big\rangle + \tfrac{1}{4}\big|100, -100\big\rangle + \tfrac{1}{4}\big|50, 50\big\rangle + \tfrac{1}{4}\big|50, -50\big\rangle. \end{aligned}$$

Its two marginals are, in $\mathcal{D}(\mathbb{R})$,

$$\begin{aligned} (p_1)_*(\omega) &= \langle p_1, p_2\rangle_*(\omega)[1,0] = \tfrac{1}{2}|100\rangle + \tfrac{1}{2}|50\rangle \\ (p_2)_*(\omega) &= \langle p_1, p_2\rangle_*(\omega)[0,1] = \tfrac{1}{4}|100\rangle + \tfrac{1}{4}|-100\rangle + \tfrac{1}{4}|50\rangle + \tfrac{1}{4}|-50\rangle. \end{aligned}$$

It is not hard to see that the parallel product $\otimes$ of these two marginal distributions differs from the joint distribution $\langle p_1, p_2\rangle_*(\omega)$ on $\mathbb{R} \times \mathbb{R}$.

Proposition 5.4.3 *The shared-state covariance of shared-state independent random variables is zero: if two random variables (ω, p_1) and (ω, p_2) with the same state are independent, then $\mathit{Cov}(\omega, p_1, p_2) = 0$.*

The converse does not hold.

Proof. If (ω, p_1) and (ω, p_2) are independent, then, by definition, $\langle p_1, p_2\rangle_*(\omega) = p_{1*}(\omega) \otimes p_{2*}(\omega)$. The calculation belows shows that covariance is then zero. It uses multiplication $\&\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as observable. It can also be described as the parallel product $\mathit{id} \otimes \mathit{id}$ of the observable $\mathit{id} : \mathbb{R} \to \mathbb{R}$ with itself.

$$\begin{aligned} \mathit{Cov}(\omega, p_1, p_2) &= (\omega \models p_1 \,\&\, p_2) - (\omega \models p_1) \cdot (\omega \models p_2) && \text{by Lemma 5.1.8} \\ &= (\omega \models \& \circ \langle p_1, p_2\rangle) \\ &\qquad - (p_{1*}(\omega \models \mathit{id})) \cdot (p_{2*}(\omega \models \mathit{id})) && \text{by (4.3)} \\ &= (\langle p_1, p_2\rangle_*(\omega) \models \&) \\ &\qquad - (p_{1*}(\omega) \otimes p_{2*}(\omega) \models \mathit{id} \otimes \mathit{id}) && \text{by Lemma 4.4.8, and (4.3)} \\ &= (\langle p_1, p_2\rangle_*(\omega) \models \&) - (\langle p_1, p_2\rangle_*(\omega) \models \&) && \text{by assumption} \\ &= 0. \end{aligned}$$

The claim that the converse does not hold follows from Example 5.4.4, right below. ❑

Example 5.4.4 We continue Example 5.4.2. The set-up used there involves two dependent random variables (ω, p_1) and (ω, p_2), with shared state $\omega = \mathit{flip} \otimes \mathit{flip}$. We show here that they (nevertheless) have covariance zero. This proves the second claim of Proposition 5.4.3, namely that zero-covariance need not imply independence, in the shared-state context.

We first compute the validities (expected values):

$$\begin{aligned}
\omega \models p_1 &= \tfrac{1}{4} \cdot p_1(1,1) + \tfrac{1}{4} \cdot p_1(1,0) + \tfrac{1}{4} \cdot p_1(0,1) + \tfrac{1}{4} \cdot p_1(0,0) \\
&= \tfrac{1}{4} \cdot 100 + \tfrac{1}{4} \cdot 100 + \tfrac{1}{4} \cdot 50 + \tfrac{1}{4} \cdot 50 \;=\; 75 \\
\omega \models p_2 &= \tfrac{1}{4} \cdot 100 + \tfrac{1}{4} \cdot -100 + \tfrac{1}{4} \cdot 50 + \tfrac{1}{4} \cdot -50 \;=\; 0
\end{aligned}$$

Then:

$$\begin{aligned}
Cov(\omega, p_1, p_2) &= \omega \models (p_1 - (\omega \models p_1) \cdot \mathbf{1}) \;\&\; (p_2 - (\omega \models p_2) \cdot \mathbf{1}) \\
&= \tfrac{1}{4}\big((100-75) \cdot 100 + (100-75) \cdot -100 + (50-75) \cdot 50 + (50-75) \cdot -50\big) \\
&= 0.
\end{aligned}$$

We now turn to independence in *joint-state* form, in analogy with joint-state covariance in Definition 5.3.1.

Definition 5.4.5 Let $\tau \in \mathcal{D}(X_1 \times X_2)$ be a joint state and with two observables $q_1 \in Obs(X_1)$ and $q_2 \in Obs(X_2)$. We say that there is *joint-state* independence of q_1, q_2 if the two random variables $(\tau, (\pi_1)^*(q_1))$ and $(\tau, (\pi_2)^*(q_2))$ are (shared-state) independent, as described in Definition 5.4.1.

Concretely, this means that:

$$(q_1 \otimes q_2)_*(\tau) \;=\; (q_1)_*(\tau[1,0]) \otimes (q_2)_*(\tau[0,1])). \tag{5.8}$$

Equation (5.8) is an instance of the formulation (5.7) used in Definition 5.4.1 since $(\pi_i)^*(q_i) = (q_i)^*(\pi_i)$ and:

$$\begin{aligned}
\langle q_1 \circ \pi_1,\, q_2 \circ \pi_2 \rangle_*(\tau) &= \mathcal{D}(\langle q_1 \circ \pi_1,\, q_2 \circ \pi_2 \rangle)(\tau) \;=\; \mathcal{D}(q_1 \times q_2)(\tau) \overset{(2.36)}{=} (q_1 \otimes q_2)_*(\tau) \\
(q_1 \circ \pi_1)_*(\tau) \otimes (q_2 \circ \pi_2)_*(\tau) &= q_{1*}((\pi_1)_*(\tau)) \otimes q_{2*}((\pi_2)_*(\tau)) \;=\; (q_1)_*(\tau[1,0]) \otimes (q_2)_*(\tau[0,1]).
\end{aligned}$$

In the joint-state case — unlike in the shared-state situation — there is a tight connection between non-entwinedness, independence and covariance being zero.

Theorem 5.4.6 *For a joint state $\tau \in \mathcal{D}(X_1 \times X_2)$ the following three statements are equivalent.*

1. *The joint state τ is non-entwined, i.e. τ is the product of its marginals.*
2. *Any two observables $q_i \in Obs(X_i)$ are joint-state independent with respect to τ.*
3. *The joint-state covariance $JCov(\tau, q_1, q_2)$ is zero, for all observables $q_i \in Obs(X_i)$ — or equivalently, all correlations $JCor(\tau, q_1, q_2)$ are zero.*

Proof. Let joint state $\tau \in \mathcal{D}(X_1 \times X_2)$ be given. We write $\tau_i := (\pi_i)_*(\tau) \in \mathcal{D}(X_i)$ for its marginals.

(1) $\Rightarrow$ (2). If τ is non-entwined, then $\tau = \tau_1 \otimes \tau_2$. Hence for all observables $q_1 \in Obs(X_1)$ and $q_2 \in Obs(X_2)$ we have that $\sigma := (q_1 \otimes q_2)_*(\tau)$ is non-entwined. To see this, first note that $(\pi_i)_*(\sigma) = (q_i)_*(\tau_i)$. Then, by Exercise 2.4.10:

$$(\pi_1)_*(\sigma) \otimes (\pi_2)_*(\sigma) \;=\; (q_1)_*(\tau_1) \otimes (q_2)_*(\tau_2) \;=\; (q_1 \otimes q_2)_*(\tau_1 \otimes \tau_2) \;=\; (q_1 \otimes q_2)_*(\tau) \;=\; \sigma.$$

(2) $\Rightarrow$ (3). Let $q_1 \in Obs(X_1)$ and $q_2 \in Obs(X_2)$ be two observables. We may assume that q_1, q_2 are independent with respect to τ, that is, $(q_1 \otimes q_2)_*(\tau) = (q_1)_*(\tau_1) \otimes (q_2)_*(\tau_2)$ as in (5.8). We must prove $JCov(\tau, q_1, q_2) = 0$. Consider, like in the proof of Proposition 5.4.3, the multiplication map

$\&\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$, given by $\&(r_1,r_2) = r_1\cdot r_2$, as an observable on $\mathbb{R}\times\mathbb{R}$. We consider its validity:

$$\begin{aligned}
(q_1\otimes q_2)_*(\tau)\models\& &= \sum_{r_1,r_2\in\mathbb{R}} (q_1\otimes q_2)_*(\tau))(r_1,r_2)\cdot\&(r_1,r_2)\\
&= \sum_{r_1,r_2\in\mathbb{R}} \mathcal{D}(q_1\times q_2)(\tau)(r_1,r_2)\cdot r_1\cdot r_2\\
&= \sum_{r_1,r_2\in\mathbb{R}} \left(\sum_{(x_1,x_2)\in(q_1\times q_2)^{-1}(r,s)} \tau(x_1,x_2)\right)\cdot r\cdot s\\
&= \sum_{r_1,r_2\in\mathbb{R}} \sum_{x_1\in q_1^{-1}(r),x_2\in q_2^{-1}(s)} \tau(x_1,x_2)\cdot q_1(x_1)\cdot q_2(x_2)\\
&= \sum_{x_1,x_2\in X} \tau(x_1,x_2)\cdot(q_1\otimes q_2)(x_1,x_2) \;=\; \tau\models q_1\otimes q_2.
\end{aligned}$$

In the same way one proves $(q_1)_*(\tau_1)\otimes(q_2)_*(\tau_2)\models\& = (\tau_1\models q_1)\cdot(\tau_2\models q_2)$. Now we are done via the formulation of binary covariance from Lemma 5.3.2:

$$\begin{aligned}
JCov(\tau,q_1,q_2) &= (\tau\models q_1\otimes q_2) - (\tau_1\models q_1)\cdot(\tau_2\models q_2)\\
&= ((q_1\otimes q_2)_*(\tau)\models\&) - ((q_1)_*(\tau_1)\otimes(q_2)_*(\tau_2)\models\&)\\
&= ((q_1\otimes q_2)_*(\tau)\models\&) - ((q_1\otimes q_2)_*(\tau)\models\&) \;=\; 0.
\end{aligned}$$

(3) $\Rightarrow$ (1). Let joint-state covariance $JCov(\tau,q_1,q_2)$ be zero for all observables q_1,q_2. In order to prove that τ is non-entwined, we have to show $\tau(x,y) = \tau_1(x)\cdot\tau_2(y)$ for all $(x,y)\in X_1\times X_2$. We choose as random variables the observables $\mathbf{1}_x$ and $\mathbf{1}_y$ and use again the formulation of covariance from Lemma 5.3.2. Then, since, by assumption, the binary covariance is zero, so that:

$$\tau(x,y) \;=\; \tau\models\mathbf{1}_x\otimes\mathbf{1}_y \;=\; (\tau_1\models\mathbf{1}_x)\cdot(\tau_2\models\mathbf{1}_y) \;=\; \tau_1(x)\cdot\tau_2(y). \qquad \square$$

In essence this result says that joint-state independence and joint-state covariance being zero are not properties of observables, but of joint states.

Exercises

5.4.1 Prove, in Definition 5.4.5, that the first marginal $(q_1\otimes q_2)_*(\tau)[1,0]$ of the transformed state along $q_1\otimes q_2$ is equal to the transformed marginal $q_{1*}(\tau[1,0])$.

5.5 Moments

This section concentrates on moments of random variables. A moment forms an infinite sequence of representational real number values. These numbers generalise validity and also variance. It is shown that the moment sequence fully characterises a distribution, in the sense that the distribution can be reconstructed from the moment sequence. This happens via so-called moment generating functions. Limits of such functions are of especial interest, since they can describe how sequences of distributions behave in the limit. This phenomenon, captured abstractly in the famous Central Limit Theorem is illustrated.

As an aside, what we describe below is sometimes called a 'raw' moment. There are also 'centralised' and 'standardised' moments where one first centralises or standardises the random variable before computing moments.

Definition 5.5.1 Let (ω, p) be a random variable, consisting of a distribution $\omega \in \mathcal{D}(X)$ and an observable $p\colon X \to \mathbb{R}$, and let $n \in \mathbb{N}$ be an arbitrary number. The n-th moment $mom[n](\omega, p) \in \mathbb{R}$ is the number defined as:

$$mom[n](\omega, p) \coloneqq \omega \models p^n = \omega \models p \,\&\, \cdots \,\&\, p.$$

Thus, explicitly, $mom[n](\omega) = \sum_x \omega(x) \cdot p(x)^n$.

When ω is a distribution on (a subset of) the real numbers, we simply write $mom[n](\omega) = mom[n](\omega, id)$, where we leave the identity observable $id : \mathbb{R} \to \mathbb{R}$ implicit.

By definition, $mom[0](\omega, p) = 1$, since $p^0 = \mathbf{1}$, and $mom[1](\omega, p) = \omega \models p$, since $p^1 = p$. Further, $Var(\omega, p) = mom[2](\omega, p) - mom[1](\omega, p)^2$, see Lemma 5.1.3. The inequality $(\omega \models p^2) \geq (\omega \models p)^2)$ in Corollary 5.1.4 (1) will later, in Proposition 6.1.8 (1), be generalised to arbitrary powers and yields an inequality, for any $n \in \mathbb{N}$,

$$mom[n](\omega, p) = \omega \models p^n \geq (\omega \models p)^n = \Big(mom[1](\omega, p)\Big)^n.$$

One of the main results will be that a distribution on the reals is completely determined by its moments, see Proposition 5.5.8 (2). But first we illustrate moments in concrete cases.

Example 5.5.2 We fix a probability $r \in [0, 1]$.

1 We first look at the coin distribution $flip(r) = r|1\rangle + (1 - r)|0\rangle$, as distribution on $\mathbb{R}$. It easy to see that all moments are the same:

$$mom[n](flip(r)) = flip(r) \models id^n = r \cdot 1^n + (1 - r) \cdot 0^n = r.$$

2 The situation is considerably more complex for the binomial distribution $bn[K](r) \in \mathcal{D}(\{0, 1, \ldots, K\})$ from Example 2.1.2 (2). The first three moments are easy:

$$\begin{aligned}
mom[0](bn[K](r)) &= 1 \\
mom[1](bn[K](r)) &= Mean(bn[K](r)) = K \cdot r, \qquad \text{see Example 4.2.2 (2)} \\
mom[2](bn[K](r)) &= Var(bn[K](r)) + Mean(bn[K](r))^2 \\
&= K \cdot r \cdot (1 - r) + (K \cdot r)^2, \qquad \text{see Exercise 5.1.15} \\
&= K \cdot (K - 1) \cdot r^2 + K \cdot r.
\end{aligned}$$

There is a formula for the general case with $K \geq n$, see (Skorski, 2022, Thm. 1). It uses Stirling numbers of the second kind and falling factorials, see Section 1.2 for details. It reads:

$$mom[n](bn[K](r)) = \sum_{0 \leq i \leq n} \left\{ {n \atop i} \right\} \cdot r^i \cdot (K)_i.$$

3 Let $\omega = \frac{1}{12}|0\rangle + \frac{1}{6}|1\rangle + \frac{1}{4}|2\rangle + \frac{1}{2}|3\rangle \in \mathcal{D}(\mathbb{R})$ be a concrete distribution. Then:

$$mom[n](\omega) = \sum_{0 \leq i \leq 3} \omega(i) \cdot i^n = \begin{cases} 1 & \text{if } n = 0 \\ \frac{1}{6} + \frac{2^n}{4} + \frac{3^n}{2} & \text{if } n > 0. \end{cases}$$

Lemma 5.5.3 *Let $\omega, \rho \in \mathcal{D}(\mathbb{R})$ be distributions on the reals and let $s \in \mathbb{R}$ be scalar.*

1 Moments commute in the following way with scalar multiplication.

$$mom[n](s * \omega) = s^n \cdot mom[n](\omega)$$

2 The moment of a convolution sum satisfies:

$$mom[n](\omega + \rho) \;=\; \sum_{0 \le i \le n} \binom{n}{i} \cdot mom[i](\omega) \cdot mom[n-i](\rho).$$

3 For $K \in \mathbb{N}_{>0}$ the K-fold convolution $K \cdot \omega = \omega + \cdots + \omega \in \mathcal{D}(\mathbb{R})$ satisfies:

$$mom[n](K \cdot \omega) \;=\; \sum_{\varphi \in \mathcal{N}[n](\{1,\ldots,K\})} (\varphi) \cdot \prod_{1 \le i \le K} mom[\varphi(i)](\omega).$$

Proof. 1 Easy, since:

$$\begin{aligned} mom[n](s * \omega) \;=\; s * \omega \models id^n \;&=\; \sum_{x \in \mathbb{R}} \omega(x) \cdot (s \cdot x)^n \\ &=\; \sum_{x \in \mathbb{R}} \omega(x) \cdot s^n \cdot x^n \;=\; s^n \cdot \sum_{x \in \mathbb{R}} \omega(x) \cdot x^n \;=\; s^n \cdot mom[n](\omega). \end{aligned}$$

2 By the Binomial Theorem:

$$\begin{aligned} mom[n](\omega + \rho) \;=\; \sum_{z \in \mathbb{R}} (\omega + \rho)(z) \cdot z^n \;&=\; \sum_{x,y \in \mathbb{R}} \omega(x) \cdot \rho(y) \cdot (x+y)^n \\ &=\; \sum_{x,y \in \mathbb{R}} \omega(x) \cdot \rho(y) \cdot \sum_{0 \le i \le n} \binom{n}{i} \cdot x^i \cdot y^{n-i} \\ &=\; \sum_{0 \le i \le n} \binom{n}{i} \cdot \left(\sum_{x \in \mathbb{R}} \omega(x) \cdot x^i \right) \cdot \left(\sum_{y \in \mathbb{R}} \rho(y) \cdot y^{n-i} \right) \\ &=\; \sum_{0 \le i \le n} \binom{n}{i} \cdot mom[i](\omega) \cdot mom[n-i](\rho). \end{aligned}$$

3 We now need the Multinomial Theorem (1.43):

$$\begin{aligned} mom[n](K \cdot \omega) \;&=\; \sum_{\vec{x} \in \mathbb{R}^K} \prod_{1 \le i \le K} \omega(x_i) \cdot (x_1 + \cdots + x_K)^n \\ &=\; \sum_{\vec{x} \in \mathbb{R}^K} \prod_{1 \le i \le K} \omega(x_i) \cdot \sum_{\varphi \in \mathcal{N}[n](\{1,\ldots,K\})} (\varphi) \cdot \prod_{1 \le i \le K} x_i^{\varphi(i)} \\ &=\; \sum_{\varphi \in \mathcal{N}[n](\{1,\ldots,K\})} (\varphi) \cdot \prod_{1 \le i \le K} \sum_{x_j \in \mathbb{R}} \omega(x_j) \cdot x_i^{\varphi(i)} \\ &=\; \sum_{\varphi \in \mathcal{N}[n](\{1,\ldots,K\})} (\varphi) \cdot \prod_{1 \le i \le K} mom[\varphi(i)](\omega). \end{aligned}$$

❑

In the next result we assume a distribution with mean zero. This can always be achieved via centralisation, see Definition 4.2.1.

Proposition 5.5.4 *Let $\omega \in \mathcal{D}(\mathbb{R})$ be a distribution satisfying $Mean(\omega) = mom[1](\omega) = 0$. For $K \in \mathbb{N}_{>0}$ we abbreviate:*

$$\omega_K \;:=\; \tfrac{1}{\sqrt{K}} * K \cdot \omega \;\in\; \mathcal{D}(\mathbb{R}). \tag{5.9}$$

1 The mean and variance of these distributions ω_K satisfy:

$$Mean(\omega_K) \;=\; \sqrt{K} \cdot Mean(\omega) \;=\; 0 \qquad\qquad Var(\omega_K) \;=\; Var(\omega).$$

2 *For each* $n \in \mathbb{N}$,

$$\lim_{K\to\infty} mom[n](\omega_K) = \begin{cases} \dfrac{(2m)!}{m!\cdot 2^m}\cdot Var(\omega)^m & \text{if } n \text{ is even, with } n = 2m. \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

The latter equation shows that in the limit, the moments of ω_K and thereby ω_K itself, are determined by one number, namely by the variance of ω. These equations for the moments of $\omega_K = \frac{1}{\sqrt{K}} * K\cdot\omega$ correspond to the moments of the continuous normal distribution with mean zero and variance equal to $Var(\omega)$, see *e.g.* (Rice, 2006, Ch. 4, Exc. 90). They may thus be interpreted that in the limit, ω_K approaches the normal distribution, as illustrated in Figure 5.3 below. This general approximation result is known as the Central Limit Theorem, see *e.g.* Billingsley (1995); Feller (1970); Grinstead and Snell (1998); Rice (2006).

Proof. 1 By Exercise 5.1.14 and Lemma 4.2.4 we have:

$$Mean\Big(\tfrac{1}{\sqrt{K}} * K\cdot\omega\Big) = \tfrac{1}{\sqrt{K}}\cdot Mean(K\cdot\omega) = \tfrac{1}{\sqrt{K}}\cdot K\cdot Mean(\omega) = \sqrt{K}\cdot Mean(\omega).$$

Additionally, by Exercise 5.1.12,

$$Var\Big(\tfrac{1}{\sqrt{K}} * K\cdot\omega\Big) = \Big(\tfrac{1}{\sqrt{K}}\Big)^2\cdot Var(K\cdot\omega) = \tfrac{1}{K}\cdot K\cdot Var(\omega) = Var(\omega).$$

2 We first use the formula in Lemma 5.5.3 (3) to find some regularities for the moments $mom[n](K\cdot\omega)$, without scalar multiplication yet. We use that the mean of ω is zero. This means that whenever there is a multiset $\varphi \in \mathcal{N}[n](\{1,\dots,K\})$ containing at least one element occurring once, we get $\prod_i mom[\varphi(i)](\omega) = 0$. Explicitly, for $n = 3,4,5,6$ where $K \geq n$, we have:

$$\begin{aligned}
mom[3](K\cdot\omega) &= K\cdot mom[3](\omega) \\
mom[4](K\cdot\omega) &= K\cdot mom[4](\omega) + 3\cdot K\cdot(K-1)\cdot mom[2](\omega)^2 \\
mom[5](K\cdot\omega) &= K\cdot mom[5](\omega) \\
&\quad + 10\cdot K\cdot(K-1)\cdot mom[2](\omega)\cdot mom[3](\omega) \\
mom[6](K\cdot\omega) &= K\cdot mom[6](\omega) + 10\cdot K\cdot(K-1)\cdot mom[3](\omega)^2 \\
&\quad + 15\cdot K\cdot(K-1)\cdot mom[2](\omega)\cdot mom[4](\omega) \\
&\quad + 15\cdot K\cdot(K-1)\cdot(K-2)\cdot mom[2](\omega)^3
\end{aligned}$$

We wish to determine the limit of $mom[n](\omega_K) = \frac{1}{(\sqrt{K})^n}\cdot mom[n](K\cdot\omega)$ for K going to infinity. All of the summands $(\varphi)\cdot\prod_i mom[\varphi(i)](\omega)$, divided by $(\sqrt{K})^n$ go to zero, if they do not contain enough factors K, $K-1$, $K-2$, etc. For instance for $n = 3,4$, we get:

$$\begin{aligned}
\frac{mom[3](K\cdot\omega)}{(\sqrt{K})^3} &= \frac{K}{K^{3/2}}\cdot mom[3](\omega) \\
&\longrightarrow 0 \\
\frac{mom[4](K\cdot\omega)}{(\sqrt{K})^4} &= \frac{K}{K^2}\cdot mom[4](\omega) + \frac{3\cdot K\cdot(K-1)}{K^2}\cdot mom[2](\omega)^2 \\
&\longrightarrow 3\cdot mom[2](\omega)^2.
\end{aligned}$$

In this same way one sees:

$$\frac{mom[5](K\cdot\omega)}{(\sqrt{K})^5} \longrightarrow 0 \qquad \frac{mom[6](K\cdot\omega)}{(\sqrt{K})^6} \longrightarrow 15\cdot mom[2](\omega)^3.$$

Thus, we need to concentrate on those multisets $\varphi \in \mathcal{N}[n](\{1,\ldots,K\})$ with maximal coefficient value $\mathbf{(}\varphi\mathbf{)}$, more specifically, with the maximum number of factors K, $K-1$, $K-2$ etc. in $\mathbf{(}\varphi\mathbf{)}$. These maxima are reached when φ contains the highest number of lowest multiplicities. As mentioned, the φ with $\varphi(i) = 1$, for some i, can be ignored, so we have to focus on those φ with as many cases $\varphi(i) = 2$ as possible. At this stage, we have to distinguish whether the number n is even or not.

- When n is even, say $n = 2m$ we can have minimal multisets φ of the form $2|i_1\rangle + \cdots + 2|i_m\rangle$, with size $\|\varphi\| = n$, as required. When K is big enough, the number of such φ, times the relevant factor is:
$$\begin{aligned}\binom{K}{m} \cdot \frac{\mathbf{(}\varphi\mathbf{)}}{(\sqrt{K})^n} \cdot mom[2](\omega)^m &= \frac{K}{K} \cdot \frac{K-1}{K} \cdot \ldots \cdot \frac{K-m}{K} \cdot \frac{n!}{m! \cdot (2!)^m} \cdot mom[2](\omega)^m \\ &\longrightarrow \frac{(2m)!}{m! \cdot 2^m} \cdot mom[2](\omega)^m.\end{aligned}$$
- When n is odd, say $n = 2m+1$ we concentrate on minimal multisets φ of the form $2|i_1\rangle + \cdots + 2|i_{m-1}\rangle + 3|i_m\rangle$, with one multiplicity of 3, again with size $\|\varphi\| = n$. We now get:
$$\begin{aligned}&\binom{K}{m} \cdot \frac{\mathbf{(}\varphi\mathbf{)}}{(\sqrt{K})^n} \cdot mom[2](\omega)^m \\ &= \frac{K}{K} \cdot \frac{K-1}{K} \cdot \ldots \cdot \frac{K-m}{K} \cdot \frac{n!}{m! \cdot (2!)^{m-1} \cdot 3! \cdot \sqrt{K}} \cdot mom[2](\omega)^m \\ &\longrightarrow 0.\end{aligned}$$
The latter expression goes to zero since there is one factor $\frac{1}{\sqrt{K}}$ left.

Finally, we are done by noting that $mom[2](\omega) = Var(\omega)$ since, by assumption, $Mean(\omega) = 0$. ❑

Example 5.5.5 We include an illustration with $\omega = \frac{1}{10}|1\rangle + \frac{1}{5}|2\rangle + \frac{7}{10}|3\rangle$, see the bar plot on the upper-left in Figure 5.3. Its mean is:
$$Mean(\omega) = \omega \models id = \sum_{1 \leq i \leq 3} \omega(i) \cdot i = \tfrac{1}{10} \cdot 1 + \tfrac{1}{5} \cdot 2 + \tfrac{7}{10} \cdot 3 = \tfrac{13}{5}$$
The variance is $Var(\omega) = \frac{11}{25}$.

The resulting centralised distribution ω' is:
$$\omega' = shift(-\tfrac{13}{5}, \omega) = \mathcal{D}((-) - \tfrac{13}{5})(\omega) = \tfrac{1}{10}\big|-\tfrac{8}{5}\big\rangle + \tfrac{1}{5}\big|-\tfrac{3}{5}\big\rangle + \tfrac{7}{10}\big|\tfrac{2}{5}\big\rangle.$$
It is plotted on the upper-right in Figure 5.3, as a shift to the left, of the distribution plotted on its left.

Next we form the scaled K-fold sums $\omega'_K = \frac{1}{\sqrt{K}} * K \cdot \omega'$ from (5.9), for $K = 5$ and for $K = 50$. The results are plotted at the bottom of Figure 5.3, giving a Bell curve. The picture contains as continous red line the (scaled) probability density function of the normal distribution with mean zero and variance given by $Var(\omega) = \frac{11}{25}$.

We turn to another construction that looks different from moments but turns out to involve them implicitly. We show how a random variable can be turned into what is called a moment generating function. We describe the construction for discrete distributions. But it also works for continuous distributions and thus provides a setting where the two kinds of distributions — discrete and continuous — can be combined and compared.

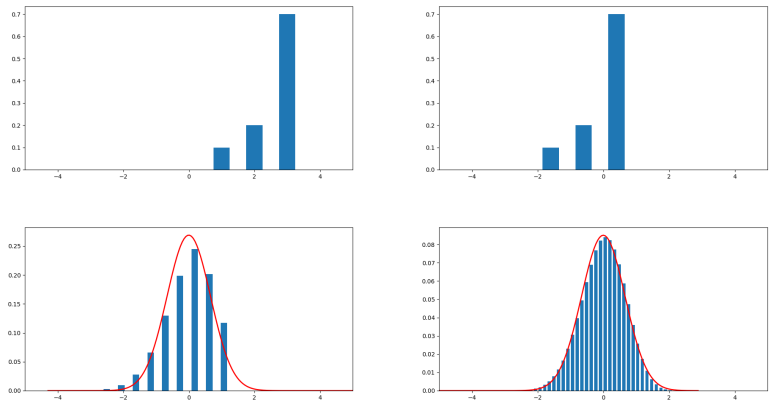


Figure 5.3 Illustration from Example 5.5.5, of a discrete distribution ω, on the upper-left, its centralisation ω' on the upper-right. The scaled iterated convolution sums $\omega_K = \frac{1}{\sqrt{K}} * K \cdot \omega'$ from (5.9) are described by the bar charts at the bottom, with $K = 5$ iterations on the left and $K = 50$ iterations on the right. These iterations lead to a Bell curve and are approached by the probability density function, as red line, of the (scaled) normal distribution, with mean zero and standard deviation equal to the standard deviation of the original discrete distribution ω. Such a Bell curve arises for arbitrary discrete distributions on the reals.

Definition 5.5.6 Let the pair $\omega \in \mathcal{D}(X)$ and $p\colon X \to \mathbb{R}$ form a random variable (ω, p). It gives rise to a *moment generating function* $mgf(\omega, p)\colon \mathbb{R} \to \mathbb{R}_{>0}$ defined as:

$$mgf(\omega, p)(t) \coloneqq \omega \models e^{p(-)\cdot t} = \sum_{x \in supp(\omega)} \omega(x) \cdot e^{p(x)\cdot t}.$$

When $X = \mathbb{R}$ we write $mgf(\omega)\colon \mathbb{R} \to \mathbb{R}_{>0}$ for $mgf(\omega, id)$.

These moment generating functions satisfy some useful properties.

Lemma 5.5.7 *Consider the moment gerating function $mgf(\omega, p)\colon X \to \mathbb{R}_{>0}$ from Definition 5.5.6.*

1 *Using the familiar Taylor series expansion $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$ we can write:*

$$mgf(\omega, p)(t) = \sum_{n \in \mathbb{N}} \frac{mom[n](\omega, p)}{n!} \cdot t^n.$$

Thus, the n-th moment $mom[n](\omega, p)$ can be retrieved from the moment generating function as the n-th derivative at 0, that is, as $mgf(\omega, p)^{(n)}(0)$.

2 *We turn to distributions on $\mathbb{R}$, with identity function as observation. Since $(\mathbb{R}_{>0}, \cdot, 1)$ is a commutative monoid, so is the set of functions $\mathbb{R}^{\mathbb{R}}_{>0}$, via pointwise multiplications. We now consider the moment generating function as acting on distributions on $\mathbb{R}$, in:*

$$\mathcal{D}(\mathbb{R}) \xrightarrow{mgf} \mathbb{R}^{\mathbb{R}}_{>0} \tag{5.10}$$

In this form mgf *is a map of commutative monoids — where we use the convolution sum on* $\mathcal{D}(\mathbb{R})$ *from Proposition 2.7.2. This map also preserves scalar multiplication. Explicitly:*

$$\begin{aligned} mgf(\omega + \rho) &= mgf(\omega) \cdot mgf(\rho) \\ mgf(1|0\rangle) &= \mathbf{1} \\ mgf(s * \omega) &= s \cdot mgf(\omega). \end{aligned}$$

Proof. 1 Since:

$$\begin{aligned} mgf(\omega, p)(t) &= \sum_{x \in supp(\omega)} \omega(x) \cdot e^{p(x) \cdot t} = \sum_{x \in supp(\omega)} \omega(x) \cdot \sum_{n \in \mathbb{N}} \frac{(p(x) \cdot t)^n}{n!} \\ &= \sum_{n \in \mathbb{N}} \sum_{x \in supp(\omega)} \omega(x) \cdot \frac{p(x)^n}{n!} \cdot t^n \\ &= \sum_{n \in \mathbb{N}} \frac{\omega \models p^n}{n!} \cdot t^n = \sum_{n \in \mathbb{N}} \frac{mom[n](\omega, p)}{n!} \cdot t^n. \end{aligned}$$

2 The zero element $1|0\rangle$ of the vector space $\mathcal{D}(\mathbb{R})$ is sent to the unit element $t \mapsto 1$ in $\mathbb{R}^{\mathbb{R}}_{>0}$ since:

$$mgf(1|0\rangle)(t) = 1 \cdot e^{0 \cdot t} = 1.$$

Convolutions sums are sent to pointwise products:

$$\begin{aligned} mgf(\omega + \rho)(t) &= \sum_{z \in \mathbb{R}} (\omega + \rho)(z) \cdot e^{z \cdot t} \overset{(2.60)}{=} \sum_{x,y \in \mathbb{R}} \omega(x) \cdot \rho(y) \cdot e^{(x+y) \cdot t} \\ &= \sum_{x,y \in \mathbb{R}} \omega(x) \cdot \rho(y) \cdot e^{x \cdot t} \cdot e^{y \cdot t} \\ &= \left(\sum_{x \in \mathbb{R}} \omega(x) \cdot e^{x \cdot t} \right) \cdot \left(\sum_{y \in \mathbb{R}} \rho(y) \cdot e^{y \cdot t} \right) \\ &= mgf(\omega)(t) \cdot mgf(\rho)(t) = \Big(mgf(\omega) \cdot mgf(\rho) \Big)(t). \end{aligned}$$

Further, the scalar multiplication of distributions (2.64) satisfies:

$$\begin{aligned} mgf(s * \omega)(t) &= \sum_{z \in \mathbb{R}} (s * \omega)(z) \cdot e^{z \cdot t} \overset{(2.64)}{=} \sum_{x \in \mathbb{R}} \omega(x) \cdot e^{(s \cdot x) \cdot t} \\ &= \sum_{x \in \mathbb{R}} \omega(x) \cdot e^{x \cdot (s \cdot t)} \\ &= mgf(\omega)(s \cdot t) = \Big(s \cdot mgf(\omega) \Big)(t). \end{aligned}$$

❑

The next result shows that a distribution (on the reals) can be reconstruced from its moment generation function. We use the proof from (Grinstead and Snell, 1998, Thm. 10.1.1). It yields an answer to what is called the moment problem, that is, the reconstruction of a distribution from its moment sequence.

Proposition 5.5.8 *Let* $f := mgf(\omega) \colon \mathbb{R} \to \mathbb{R}_{>0}$, *for* $\omega \in \mathcal{D}(\mathbb{R})$. *Then one can reconstruct the distribution* ω *from* f.

As a result,

1 *the moment generation function* mgf *in Diagram (5.10) is injective;*
2 *a distribution* $\omega \in \mathcal{D}(\mathbb{R})$ *can be reconstructed from its moment sequence* $(\omega \models id^n)_{n \in \mathbb{N}}$.

Proof. Let $\omega \in \mathcal{D}(\mathbb{R})$ and let the associated moment generating function $f := mgf(\omega)\colon \mathbb{R} \to \mathbb{R}_{>0}$ be available. We show how to obtain ω by using f alone.

Since $\mathbb{R}$ is totally ordered, we can write the support of ω as $supp(\omega) = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}$ with $x_1 < x_2 < \cdots < x_m$. Then we can write f and its derivative f' as:

$$\begin{aligned} f(t) &= \omega(x_1) \cdot e^{x_1 \cdot t} + \cdots + \omega(x_m) \cdot e^{x_m \cdot t} \\ f'(t) &= \omega(x_1) \cdot x_1 \cdot e^{x_1 \cdot t} + \cdots + \omega(x_m) \cdot x_m \cdot e^{x_m \cdot t}. \end{aligned}$$

We claim that the largest number x_m in the support can be extracted from f as $\lim_{t\to\infty} \frac{f'(t)}{f(t)}$. Indeed:

$$\begin{aligned} \lim_{t\to\infty} \frac{f'(t)}{f(t)} &= \lim_{t\to\infty} \frac{f'(t)/e^{x_m \cdot t}}{f(t)/e^{x_m \cdot t}} \\ &= \frac{\lim_{t\to\infty} \omega(x_1) \cdot x_1 \cdot e^{(x_1 - x_m)\cdot t} + \cdots + \omega(x_{m-1}) \cdot x_{m-1} \cdot e^{(x_{m-1} - x_m)\cdot t} + \omega(x_m) \cdot x_m}{\lim_{t\to\infty} \omega(x_1) \cdot e^{(x_1 - x_m)\cdot t} + \cdots + \omega(x_{m-1}) \cdot e^{(x_{m-1} - x_m)\cdot t} + \omega(x_m)} \\ &= \frac{\omega(x_m) \cdot x_m}{\omega(x_m)} \qquad \text{since } x_i - x_m < 0 \\ &= x_m. \end{aligned}$$

Once this largest element $x_m \in \mathbb{R}$ in the support is found, we can obtain the associated probability $\omega(x_m)$ from f as:

$$\lim_{t\to\infty} \frac{f(t)}{e^{x_m \cdot t}} = \lim_{t\to\infty} \omega(x_1) \cdot e^{(x_1 - x_m)\cdot t} + \cdots + \omega(x_{m-1}) \cdot e^{(x_{m-1} - x_m)\cdot t} + \omega(x_m) = \omega(x_m).$$

At this stage we replace f by the function $t \mapsto f(t) - \omega(x_m) \cdot e^{x_m \cdot t}$. We can now obtain the second largest element in the support, x_{m-1}, and its probability $\omega(x_{m-1})$, via the same steps as described above, but this time using the replacement function. We continue finding elements in the support and their probabilities until we end up with the zero function. At that stage we have reconstructed the entire distribution ω.

Clearly, the above method to retrieve ω from the function $mgf(\omega)\colon \mathbb{R} \to \mathbb{R}_{>0}$ shows that mgf in (5.10) is injective.

Also, if two distributions $\omega, \rho \in \mathcal{D}(\mathbb{R})$ have the same moments $\omega \models id^n = m_n = \rho \models id^n$, then they have the same moment generating function $mgf(\omega) = f = mgf(\rho)$, where $f(t) = \sum_n \frac{m_n}{n!} \cdot t^n$, by Lemma 5.5.7 (1). Hence $\omega = \rho$, since mgf is injective. ❑

In Proposition 5.5.4 (2) we have seen that for a distribution on the reals with mean zero, its moments of iterated convolutions, in the limit, are determined by the variance of the distribution. Moreover, these moments turn out to be the same of a (continuous) normal distribution. Below we shall see a similar situation in terms of moment generating functions. Again, this is known as the Central Limit Theorem, see *e.g.* Billingsley (1995); Feller (1970); Grinstead and Snell (1998); Rice (2006).

Proposition 5.5.9 *Let $\omega \in \mathcal{D}(\mathbb{R})$ be a discrete distribution with mean zero and let $\omega_K := \frac{1}{\sqrt{K}} * K \cdot \omega$ as in (5.9). Then:*

$$\left(\lim_{K\to\infty} mgf(\omega_K)\right)(t) = e^{\frac{Var(\omega)}{2} \cdot t^2}.$$

We see that in the limit the moment generating function of ω_K is entirely determined by a single number, namely by the variance of ω. In general, the function $t \mapsto e^{\mu \cdot t + \frac{\nu}{2} \cdot t^2}$ is the moment

generating function of the normal distribution with mean μ and variance ν, see *e.g.* (Blitzstein and Hwang, 2019, Ex. 6.4.12).

Proof. We write $f_K = mgf(\omega_K)$ for the moment generating function of $\omega_K \in \mathcal{D}(\mathbb{R})$. We also abbreviate the moments of ω as $m_n := mom[n](\omega)$, so that $m_0 = 1$, $m_1 = 0$ and $m_2 = Var(\omega)$. By Lemma 5.5.7 we have:

$$\begin{aligned} f_K(t) &= \left(mgf(\omega)(\tfrac{1}{\sqrt{K}} \cdot t)\right)^K = \left(1 + 0 + \frac{Var(\omega)}{2K} \cdot t^2 + \sum_{n \geq 3} \frac{m_n}{n!} \cdot \frac{t^n}{K^{n/2}}\right)^K \\ &= \left(1 + \frac{g_K(t)}{K}\right)^K \qquad \text{for } g_K(t) := \frac{Var(\omega) \cdot t^2}{2} + \sum_{n \geq 3} \frac{m_n}{n!} \cdot \frac{t^n}{K^{n/2-1}}. \end{aligned}$$

We claim that $\lim_{K \to \infty} g_K(t) = \frac{Var(\omega)}{2} \cdot t^2$. This follows from Lemma 5.5.7 (1):

$$\lim_{K \to \infty} \sum_{n \geq 3} \frac{m_n}{n!} \cdot \frac{t^n}{K^{n/2-1}} \leq \lim_{K \to \infty} \frac{mgf(\omega)(t) - 1 - m_1 \cdot t - \frac{m_2}{2} \cdot t^2}{\sqrt{K}} = 0.$$

The inequality follows from the fact that $K^{n/2-1} \geq K^{1/2} = \sqrt{K}$ for $n \geq 3$ and $K \geq 1$. In a final step we use the familiar fact that e-powers can be obtained from limits in the following way.

$$\lim_{n \to \infty} \left(1 + \frac{a_n}{n}\right)^n = e^a \qquad \text{when} \qquad \lim_{n \to \infty} a_n = a.$$

With this equation we can conclude:

$$\lim_{K \to \infty} mgf(\omega_K) = \lim_{K \to \infty} \left(1 + \frac{g_K(t)}{K}\right)^K = e^{\lim_{K \to \infty} g_K(t)} = e^{\frac{Var(\omega)}{2} \cdot t^2}. \qquad \square$$

Exercises

5.5.1 Consider a function $f \colon X \to Y$ with a distribution $\omega \in \mathcal{D}(X)$ on its domain and an observable $q \colon Y \to \mathbb{R}$ on its codomain. Show that:

$$mom[n](\mathcal{D}(f)(\omega), q) = mom[n](\omega, q \circ f).$$

This resembles Exercise 5.1.17.

5.5.2 Recall the multiplicative convolution $\omega \star \rho = \mathcal{D}(\cdot)(\omega \otimes \rho) \in \mathcal{D}(\mathbb{R})$ for $\omega, \rho \in \mathcal{D}(\mathbb{R})$ from Exercise 2.7.4, obtained by using the commutative monoid $(\mathbb{R}, \cdot, 1)$ of the reals with multiplication. Show that this multiplicative convolution satisfies:

$$mom[n](\omega \star \rho) = mom[n](\omega) \cdot mom[n](\rho).$$

5.5.3 Check that:

$$mgf(flip(r))(t) = r \cdot e^t + 1 - r.$$

Conclude that:

$$mgf(bn[K](r))(t) = (r \cdot e^t + 1 - r)^K.$$

5.5.4 Consider the following variation on Example 5.1.5. Fix $N \in \mathbb{N}_{>0}$ and let $\upsilon = \sum_{-N \leq i \leq N} \frac{1}{2N+1} |i\rangle$ be the uniform distribution on the set $\{-N, -N+1 \ldots, 0, \ldots N-1, N\} \subseteq \mathbb{R}$.

1 Check that $Mean(\upsilon) = 0$ and $Var(\upsilon) = \frac{N(N+1)}{3}$, and thus $mom[2](\upsilon) = \frac{N(N+1)}{3}$.

2 Define $\upsilon_K := \frac{1}{\sqrt{K}} * (K \cdot \upsilon)$ and show that:

$$\lim_{K\to\infty} mgf(\upsilon_K)(t) \;=\; e^{\frac{N(N+1)}{6}\cdot t^2}.$$

5.5.5 Use Proposition 2.7.10 to show that for a distribution $\omega \in \mathcal{D}(\mathbb{R})$ and a number $K \in \mathbb{N}_{>0}$,

$$\begin{aligned}\tfrac{1}{\sqrt{K}} * K \cdot \omega \;&=\; \sum_{\varphi\in\mathcal{M}[K](\mathbb{R})} mn[K](\omega)(\varphi)\,\Big|\, \tfrac{1}{\sqrt{K}} \cdot sum(\varphi)\Big\rangle \\ &=\; \sum_{\varphi\in\mathcal{M}[K](\mathbb{R})} mn[K](\omega)(\varphi)\,\Big|\, \sqrt{K} \cdot Mean(Flrn(\varphi))\Big\rangle.\end{aligned}$$

Thus, the Central Limit Theorem also says that they means of (normalised) large multinomial draws approximate a Bell curve.

5.6 The laws of large draws and of large numbers

Intuitively, when we sample many elements from a distribution ω, the distribution of samples will be much like ω. In a bit more precise form, for a large multiset φ drawn from ω, the distribution $Flrn(\varphi)$ obtained by normalisation will be close to ω. We can express this closeness in terms of a small (total variation) distance $d(\omega, Flrn(\varphi))$. There is one subtlety involved: this distance going to zero happens 'in probability', that is, one has to take the multinomial probability $mn[K](\omega)(\varphi)$ of drawing φ (of size K) from ω into account. For instance, the draw φ may be of the form $K|x\rangle$, so that the element x is drawn K-times. When ω is not the point distribution $1|x\rangle$, the distance between ω and $Flrn(K|x\rangle) = 1|x\rangle$ does not go to zero. However, the probability that $K|x\rangle$ is drawn is very small, for large K. We thus arrive at the following statement:

$$\lim_{K\to\infty} mn[K](\omega) \models d(\omega, Flrn(-)) \;=\; 0,$$

where $mn[K](\omega) \models d(\omega, Flrn(-)) = \sum_\varphi mn[K](\omega)(\varphi) \cdot d(\omega, Flrn(\varphi))$.

This limit statement expresses that when the draw size K goes to infinity, the distance between ω and the frequentist learning $Flrn(\varphi)$ of a draw φ goes to zero, 'in probability', by taking the multinomial probabilities $mn[K](\omega)(\varphi)$ of the draws into account. In Jacobs (2025c) the above equation is called the law of large draws. It expresses in a direct manner the main idea that large draws are much like the original distribution. Below, we shall first prove this law of large draws and then use it to derive results that resemble what are commonly called the laws of large numbers. They exist in several forms. We present these laws of large numbers also as limits of validities going to zero, without using any small number ε, as in traditional form. In the end it is briefly discussed that the formulation that we use, in terms of validities going to zero, is stronger than the formulation in terms of ε.

Theorem 5.6.1 *For each distribution $\omega \in \mathcal{D}(X)$,*

$$\lim_{K\to\infty} mn[K](\omega) \models d(\omega, Flrn(-)) \;=\; 0.$$

Proof. We use Lemma 3.3.3 and follow the proof of Proposition 5.2.1 (1) (using Exercise 3.3.8)

in:

$$
\begin{aligned}
&\lim_{K\to\infty} mn[K](\omega) \models d(\omega, Flrn(-)) \\
&= \lim_{K\to\infty} \sum_{\varphi\in\mathcal{N}[K](X)} mn[K](\omega)(\varphi)\cdot\tfrac{1}{2}\cdot \sum_{y\in supp(\omega)} \big|Flrn(\varphi)(y) - \omega(y)\big| \\
&= \tfrac{1}{2} \sum_{y\in supp(\omega)} \lim_{K\to\infty} mn[K](\omega) \models \big|Flrn(-)(y) - (mn[K](\omega) \models Flrn(-)(y))\big| \\
&\leq \tfrac{1}{2} \sum_{y\in supp(\omega)} \lim_{K\to\infty} \sqrt{Var\big(mn[K](\omega),\, Flrn(-)(y)\big)} \qquad \text{by Corollary 5.1.4 (2)} \\
&= \tfrac{1}{2} \sum_{y\in supp(\omega)} \lim_{K\to\infty} \sqrt{\big(mn[K](\omega) \models Flrn(-)(y)^2\big) - \big(mn[K](\omega) \models Flrn(-)(y)\big)^2} \\
&= \tfrac{1}{2} \sum_{y\in supp(\omega)} \lim_{K\to\infty} \sqrt{\frac{K\cdot(K-1)\cdot\omega(y)^2 + K\cdot\omega(y)}{K^2} - \omega(y)^2} \\
&= \tfrac{1}{2} \sum_{y\in supp(\omega)} \lim_{K\to\infty} \frac{\sqrt{\omega(y)\cdot(1-\omega(y))}}{\sqrt{K}} \\
&= 0.
\end{aligned}
$$

❑

The law of large numbers is sometimes, like in (Grinstead and Snell, 1998, §8.1), formulated in terms of iterated convolution sums. We show how this fits in the formulation of Theorem 5.6.1, via the relation between iterated convolutions and multinomials in Proposition 2.7.10. We use this for distributions on the real numbers $\mathbb{R}$, as commutative (additive) monoid.

Corollary 5.6.2 *For a distribution $\omega \in \mathcal{D}(\mathbb{R})$, write $\mu := Mean(\omega) = \omega \models id$. The iterated convolution $K\cdot\omega = \omega + \cdots + \omega \in \mathcal{D}(\mathbb{R})$ satisfies:*

$$\lim_{K\to\infty} K\cdot\omega \models \big|\mu - \tfrac{(-)}{K}\big| = \lim_{K\to\infty} \omega^K \models \big|\mu - \tfrac{sum(-)}{K}\big| = 0.$$

The observable $\big|\mu - \frac{(-)}{K}\big| : \mathbb{R} \to \mathbb{R}$ sends $x \in \mathbb{R}$ to $\big|\mu - \frac{x}{K}\big|$. Similarly, the observable $\big|\mu - \frac{sum(-)}{K}\big| : \mathbb{R}^K \to \mathbb{R}$ maps $\vec{x}$ to $\big|\mu - \frac{sum(\vec{x})}{K}\big|$.

Notice that 'in distribution' now refers to convolutions $K\cdot\omega = \omega + \cdots + \omega$ and to product distributions $\omega^K = \omega \otimes \cdots \otimes \omega$, instead of to multinomials in Theorem 5.6.1.

Proof. To prove that the first limit goes to zero, we use the equation $K\cdot\omega = \mathcal{D}(sum)(mn[K](\omega))$ from Proposition 2.7.10. We use the following observation. Let $M \in \mathbb{R}_{>0}$ be such that $supp(\omega) \subseteq [-M, M] \subseteq \mathbb{R}$. For a multiset $\varphi \in supp(mn[K](\omega))$ we have $supp(\varphi) \subseteq supp(\omega)$. Thus:

$$
\begin{aligned}
\big|Mean(\omega) - Mean(Flrn(\varphi))\big| &= \Big|\sum_{x\in supp(\omega)} \omega(x)\cdot x - \sum_{x\in supp(\omega)} Flrn(\varphi)(x)\cdot x\Big| \\
&\leq \sum_{x\in supp(\omega)} |x|\cdot\big|\omega(x) - Flrn(\varphi)(x)\big| \\
&\leq \sum_{x\in supp(\omega)} M\cdot\big|\omega(x) - Flrn(\varphi)(x)\big| \\
&= 2M\cdot d(\omega, Flrn(\varphi)), \qquad \text{see Proposition 4.8.1.}
\end{aligned}
\tag{$*$}
$$

We can now prove that the first limit in the theorem goes to zero, via Theorem 5.6.1.

$$\begin{aligned}
\lim_{K\to\infty} K\cdot\omega \models \left|\mu - \tfrac{(-)}{K}\right| &= \lim_{K\to\infty} \mathcal{D}(sum)(mn[K](\omega)) \models \left|Mean(\omega) - \tfrac{(-)}{K}\right| \\
&\overset{(4.3)}{=} \lim_{K\to\infty} mn[K](\omega) \models \left|Mean(\omega) - \tfrac{sum(-)}{K}\right| \\
&\overset{(4.21)}{=} \lim_{K\to\infty} mn[K](\omega) \models \left|Mean(\omega) - Mean(Flrn(-))\right| \\
&\leq \lim_{K\to\infty} mn[K](\omega) \models 2M\cdot d(\omega, Flrn(-)) \qquad \text{by } (*) \\
&= 2M\cdot \lim_{K\to\infty} mn[K](\omega) \models d(\omega, Flrn(-)) \\
&= 0.
\end{aligned}$$

Now that we know that the first limit in the theorem goes to zero, we can conclude that the second limit goes to zero too. We exploit the defining equation $K\cdot\omega = \mathcal{D}(sum)(\omega^K)$, see Definition 2.7.1. We are done by another application of Lemma 4.1.5. ❑

The previous result applies to distributions on the reals. We like to extend this to arbitrary distributions, with a function to the reals, that is, to random variables. We also like to extend from mean to variance. This is done next.

Corollary 5.6.3 *Let (ω, p) be a random variable, where $\omega \in \mathcal{D}(X)$ and $p\colon X \to \mathbb{R}$.*

1 Using $\mu := \omega \models p$,

$$\lim_{K\to\infty} \omega^K \models \left|\mu - (\tfrac{sum}{K}\circ p^K)\right| = 0.$$

The expression $\frac{sum}{K}\circ p^K$ describes the function $X^K \to \mathbb{R}$ given by $\vec{x} \mapsto \frac{1}{K}\cdot(p(x_1)+\cdots+p(x_K))$.

2 Similarly, for $\nu := Var(\omega, p)$,

$$\lim_{K\to\infty} \omega^K \models \left|\nu - \left(\tfrac{sum}{K}\circ((p^K - \mu\cdot\mathbf{1})^2)^K\right)\right| = 0.$$

The function $\frac{sum}{K}\circ((p^K - \mu\cdot\mathbf{1})^2)^K$ sends $\vec{x}\in X^K$ to $\frac{1}{K}\cdot((p(x_1)-\mu)^2+\cdots+(p(x_K)-\mu)^2)$.

The first equation justifies the sampling described in Remark 4.6.8: the validity $\omega \models p$ can be obtained as $\frac{1}{K}\cdot(p(x_1)+\cdots+p(x_K))$, for K-many samples x_i from ω, for large K. Similarly, the second point says that the variance $Var(\omega, p)$ can be obtained from samples x_i and computing $\frac{1}{K}\cdot\sum_i (p(x_i) - (\omega\models p))^2$.

Proof. 1 We apply Corollary 5.6.2 to the pushforward distribution $\mathcal{D}(p)(\omega) \in \mathcal{D}(\mathbb{R})$ satisfying $Mean(\mathcal{D}(p)(\omega)) = (\omega \models p) = \mu$, see Exercise 5.1.2. Then:

$$\begin{aligned}
0 &= \lim_{K\to\infty} \mathcal{D}(p)(\omega)^K \models \left|Mean(\mathcal{D}(p)(\omega)) - \tfrac{sum(-)}{K}\right| \\
&= \lim_{K\to\infty} \mathcal{D}(p^K)(\omega^K) \models \left|\mu - \tfrac{sum(-)}{K}\right| \qquad \text{by Lemma 2.3.2 (3)} \\
&\overset{(4.3)}{=} \lim_{K\to\infty} \omega^K \models \left|\mu - (\tfrac{sum}{K}\circ p^K)\right|
\end{aligned}$$

2 We now use the function $f := (p - \mu\cdot\mathbf{1})^2\colon X \to \mathbb{R}$ so that $\mathcal{D}(f)(\omega) \in \mathcal{D}(\mathbb{R})$ satisfies: $Mean(\mathcal{D}(f)(\omega)) = Var(\omega, p) = \nu$, see Exercise 5.1.2. We proceed as before:

$$\begin{aligned}
0 &= \lim_{K\to\infty} \mathcal{D}(f)(\omega)^K \models \left|Mean(\mathcal{D}(f)(\omega)) - \tfrac{sum(-)}{K}\right| \\
&= \lim_{K\to\infty} \omega^K \models \left|\nu - (\tfrac{sum}{K}\circ f^K)\right| \\
&= \lim_{K\to\infty} \omega^K \models \left|\nu - \left(\tfrac{sum}{K}\circ((p^K - \mu\cdot\mathbf{1})^2)^K\right)\right|.
\end{aligned}$$

❑

We conclude this section by relating the above formulations using a limit of validities to traditional formulations.

Remark 5.6.4 Let $\omega \in \mathcal{D}(X)$ with $p\colon X \to \mathbb{R}$ be a random variable, with validity $\mu := \omega \models p$. The traditional formulation of the law of large numbers says that for each $\varepsilon > 0$ one has:

$$\lim_{K\to\infty} P\Big(\big|\mu - (\tfrac{sum}{K} \circ p^K)\big| \geq \varepsilon\Big) \;=\; 0. \tag{5.11}$$

This $P(-)$ notation leaves the distribution implicit but should be read as validity $\omega^K \models (-)$. We elaborate on the meaning of the inequality $\geq \varepsilon$.

In general, if we have an observable $p\colon X \to \mathbb{R}$ and a number $a \in \mathbb{R}$ we introduce a sharp predicate $[p \geq a]$ on X, defined in an obvious way as:

$$[p \geq a](x) := \begin{cases} 1 & \text{if } p(x) \geq a \\ 0 & \text{otherwise.} \end{cases} \tag{5.12}$$

Then, in presence of a distribution $\omega \in \mathcal{D}(X)$, Markov's inequality holds:

$$a \cdot \Big(\omega \models [p \geq a]\Big) \;\leq\; \omega \models p. \tag{5.13}$$

The proof is easy:

$$\begin{aligned} \omega \models p \;\geq\; \omega \models p \;\&\; [p \geq a] \;\geq\; \omega \models (a \cdot \mathbf{1}) \;\&\; [p \geq a] &= \omega \models a \cdot (\mathbf{1} \;\&\; [p \geq a]) \\ &= \omega \models a \cdot [p \geq a] \\ &= a \cdot (\omega \models [p \geq a]). \end{aligned}$$

Suppose now we have a sequence of random variables (ω_K, p_K), all on the same set X. Then:

$$\lim_{K\to\infty} \omega_K \models p_K \;=\; 0 \;\Longrightarrow\; \lim_{K\to\infty} \omega_K \models [p_K \geq \varepsilon] \;=\; 0, \quad \text{for each } \varepsilon > 0. \tag{5.14}$$

In the results earlier in this section we have used formulations like on left of the implication $\Longrightarrow$. They are thus stronger than the formulations with an ε on the right. We claim that the formulations on the left are not only stronger but also easier and clearer.

The proof of the implication (5.14) is easy. Fix $\varepsilon > 0$ and let $\delta > 0$ be given; our aim is to force $\omega_K \models [p_K \geq \varepsilon]$ below δ. Since $\lim_{K\to\infty} \omega_K \models p_K = 0$ we can find and $N \in \mathbb{N}$ such that for all $K \geq N$ one has $\omega_K \models p_K \leq \varepsilon \cdot \delta$. Then also for $K \geq N$, by Markov's inequality (5.13):

$$\omega_K \models [p_K \geq \varepsilon] \;\leq\; \tfrac{1}{\varepsilon} \cdot (\omega_K \models p_K) \;\leq\; \tfrac{1}{\varepsilon} \cdot \varepsilon \cdot \delta \;=\; \delta.$$

We can now conclude that the traditional formulation (5.11) of the law of large numbers is a consequence of the formulation in Corollary 5.6.3 (1), via the implication (5.14).

Exercises

5.6.1 Check that Theorem 5.6.1 can equivalently be formulated as:

$$\lim_{K\to\infty} \mathcal{D}(Flrn)(mn[K](\omega)) \models d(\omega, -) \;=\; 0.$$

5.6.2 Derive from Theorem 5.6.1 that for distributions $\omega, \rho \in \mathcal{D}(X)$ one has:

$$\lim_{K\to\infty} mn[K](\omega) \models d(Flrn(-), \rho) \;=\; d(\omega, \rho).$$

5.6.3 Let (ω, p) be a random variable. Show that the n-th moment can is approximated as follows.

$$\lim_{K\to\infty} \omega^K \models \Big| \mathit{mom}[n](\omega, p) - \Big(\tfrac{\mathit{sum}}{K} \circ (p \,\&\, \cdots \,\&\, p)^K\Big)\Big| \;=\; 0.$$

5.6.4 For a number $a \in \mathbb{R}$, write $\mathbf{1}_{\geq a} \colon \mathbb{R} \to [0,1]$ for the sharp predicate with $\mathbf{1}_{\geq a}(r) = 1$ iff $r \geq a$. Consider an observable $p \colon X \to \mathbb{R}$ as a deterministic channel. Check that the sharp predicate $[p \geq a]$ from (5.12) can also be described as predicate transformation $p^*(\mathbf{1}_{\geq a})$.

5.6.5 There is another famous inequality for the kind of sharp predicates introduced in (5.12), besides the one of Markov in (5.13), namely Chebyshev's inequality. It says, for a random variable (ω, p) and a number $a \in \mathbb{R}$,

$$a^2 \cdot \Big(\omega \models \big[\,|\, p - (\omega \models p) \cdot \mathbf{1} \,| \;\geq a\big]\Big) \;\leq\; \mathit{Var}(\omega, p).$$

Derive this inequality of Chebyshev from Markov's inequality.

6

Updating distributions

One of the most interesting and magical aspects of probability distributions (states) is that they can be 'updated'. Informally, this means that in the light of evidence (new information), one can revise a distribution to a new distribution, so that the new distribution better matches the evidence. This updating is also called belief update, conditioning, revision, learning or inference.

Less informally, for a factor (or predicate or event) p and a distribution ω, both on the same sample space, one can form a new updated (conditioned) distribution, written as $\omega|_p$. It absorps the evidence p into ω. This updating satisfies various properties, including the famous rule of Thomas Bayes, and also a less well known but also important validity-increase property:

$$\omega|_p \models p \;\geq\; \omega \models p.$$

It states that after absorbing the evidence p the validity of p increases. This makes perfect good sense and is crucial in probabilistic learning. We will say informally: learning makes you wiser. After learning about p, you expect it more.

Updating is particularly interesting for joint distributions. A theme that will run through this chapter is that probabilistic updating (conditioning) has 'crossover' influence. This means that if we have a joint distribution on a product sample space, then updating in one component typically changes the distribution in the other component (the marginal). This crossover influence is another magical aspect in a probabilistic setting and depends on correlation between the two components. As we shall see, channels play an important role in this phenomenon.

The combination of updating and transformation along a channel adds its own dynamics to the topic. An update $\omega|_p$ involves both a distribution ω and a factor p. In presence of a channel we can first transform the state ω or the factor p and then update. But we can also transform an updated state $\omega|_p$. How are all these related? The power of conditioning becomes apparent when it is combined with transformation, especially for inference in probabilistic reasoning. We shall see two forms of inference: *forward* inference involves conditioning followed by state transformation, as $c_*(\omega|_p)$. It is commonly called causal reasoning. There is also *backward* inference, as in $\omega|_{c^*(q)}$, which is observable transformation followed by conditioning, and is known as evidential reasoning. We illustrate the usefulness of both these inference techniques in many examples in Sections 6.2 and 6.9, but also for Bayesian networks, in Sections 6.4 and 6.7. In the latter two sections we illustrate how a representation as a string diagram guides the reasoning in a network, up and down channels, via forward and backward inference, involving (forward) state transformation and (backward) predicate transformation. These string diagrams are used both for Bayesian networks and for (hidden) Markov models. Inbetween, in Sections 6.5 and 6.6 we sketch how conditioning works in the probabilistic programming languages ProbLog and WebPPL.

In Section 6.9 we take a first look at the topic of parameter learning, especially in the form of

learning the value of a bias $r \in [0,1]$ in a distribution $\mathit{flip}(r)$, when there is coin flip evidence. More concretely, if we have a coin of unknown bias r, and we have evidence of so-and-so many of its head and tail flips, what can we deduce about this bias r? Since the bias has a continuous range $[0,1]$, we should use a continuous distribution over $[0,1]$, that gets updated with every piece of evidence. So far we have considered only discrete distributions. We illustrate how chopping up the unit interval into discrete parts, with a discrete distribution on these parts, still gives reasonably good results.

Section 6.10 is of a more fundamental nature. Technically, it gives an isomorphism between the set $\mathcal{D}(X)$ of distributions on a finite set X, and the set of functions $\mathit{Pred}(X) \to [0,1]$ that preserve the effect module structure. We argue that this isomorphism relates the frequentist view on probabilities, using distributions $\omega \in \mathcal{D}(X)$ as assignments of probabilities, to the Bayesian view, using belief evaluations of predicates.

The final two section 6.11 and 6.12 of this chapter collect miscellanious results about updating, involving (co)variances and also e-powers e^p of observables p.

6.1 Update basics

We shall use updating and conditioning synonymously. These terms refer to the incorporation of evidence into a distribution, where the evidence is given by a predicate (or more generally by a factor). At this stage we use the term 'evidence' in an informal sense, but later on, in Section 8.3, we give it a precise meaning. This section describes the definition and basic results, including Bayes' theorem and validity-increase. The relevance of conditioning in probabilistic reasoning will be demonstrated in many examples later on in this chapter.

Definition 6.1.1 Let $\omega \in \mathcal{D}(X)$ be a distribution on a sample space X and let $p \in \mathit{Fact}(X) = (\mathbb{R}_{\geq 0})^X$ be a factor, on the same space X.

1 If the validity $\omega \models p$ is non-zero, we define a new distribution $\omega|_p \in \mathcal{D}(X)$ as normalised pointwise product of ω and p:

$$\omega|_p(x) := \frac{\omega(x) \cdot p(x)}{\omega \models p} \qquad \textit{i.e.} \qquad \omega|_p = \sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \models p} \big|x\rangle. \tag{6.1}$$

This new distribution $\omega|_p$ may be pronounced as: ω, given p. We call it the *Bayesian update* of ω, to distinguish it from other forms of updating that we shall see later on.

2 The *conditional expectation* or *conditional validity* of an observable q on X, given p and ω, is the validity:

$$\omega|_p \models q.$$

3 For a channel $c \colon X \rightsquigarrow Y$ and a factor q on Y we define the updated channel $c|_q \colon X \rightsquigarrow Y$ via pointwise updating:

$$c|_q(x) := c(x)|_q.$$

In writing $c|_q$ we assume that validity $c(x) \models q = c^*(q)(x)$ is non-zero, for each $x \in X$.

Often, the distribution ω before updating is called the *prior* or the *a priori* distribution. It captures the situation before we learn about the evidence. In (Jaynes, 2003, §1.1) the prior is

compared to common sense. The updated distribution $\omega|_p$ is called the *posterior* or the *a posteriori* distribution. The posterior incorporates the evidence given by the factor p. The task of computing the posterior from the prior is called *inference*. In principle, one can also reconstruct the prior from the posterior, see Exercise 6.9.4, but this is less common. One may expect that in the posterior $\omega|_p$ the factor p is more true than in the prior ω. This is indeed the case, as will be shown in Theorem 6.1.6 below.

In Equation (6.1) we see that the validity $\omega \models p$ is used for normalisation. When writing $\omega|_p$ we tacitly assume that the validity $\omega \models p$ is non-zero. Computing this validity may be computationally expensive, when the distribution ω has a very large support. We shall describe how to sample an updated distribution in Subsection 6.8.1.

Notice that we define updating $\omega|_p$ only for factor $p\colon X \to \mathbb{R}_{\geq 0}$, with non-negative outcomes, and not for arbitrary observables $X \to \mathbb{R}$. The latter lead to negative probabilities, which are excluded in our setting.

The conditioning $c|_q$ of a channel in item (3) is in fact a generalisation of the conditioning of a state $\omega|_p$ in item (1), since the state $\omega \in \mathcal{D}(X)$ can be seen as a channel $\omega\colon \mathbf{1} \rightsquigarrow X$ with a one-element set $\mathbf{1} = \{0\}$ as domain. We shall demonstrate the usefulness of conditioning of channels in Section 6.3, but we focus on updating of distributions first.

The standard conditional probability notation is: $P(E \mid D)$ for events $E, D \subseteq X$, where the distribution involved is left implicitly. If ω is this implicit distribution, then $P(E \mid D)$ corresponds to the conditional expectation expressed by the validity $\omega|_{\mathbf{1}_D} \models \mathbf{1}_E$ of the sharp predicate $\mathbf{1}_E$, in the state ω updated with the sharp predicate $\mathbf{1}_D$. Indeed,

$$\begin{aligned}
\omega|_{\mathbf{1}_D} \models \mathbf{1}_E \;=\; \sum_{x \in X} \omega|_{\mathbf{1}_D}(x) \cdot \mathbf{1}_E(x) &\overset{(6.1)}{=} \sum_{x \in E} \frac{\omega(x) \cdot \mathbf{1}_D(x)}{\omega \models \mathbf{1}_D} \\
&= \frac{\sum_{x \in E \cap D} \omega(x)}{\sum_{x \in D} \omega(x)} \;=\; \frac{P(E \cap D)}{P(D)} \;=\; P(E \mid D).
\end{aligned}$$

The formulation $\omega|_p$ of conditioning that is used above is not restricted to sharp predicates, but works much more generally for fuzzy predicates / factors p. This is sometimes called updating with *soft* or *uncertain* evidence Chan and Darwiche (2005); Darwiche (2009); Jacobs (2019). It is what we use as default form. Another difference with the traditional notation $P(E \mid D)$ is that in our notation $\omega|_p$ both the prior ω and the posterior $\omega|_p$ are written explicitly, so that one can easily make clear which distribution is used, for instance in a random variable or in an expected value calculation.

Examples 6.1.2 1 Let's take the numbers of a dice as sample space: $\mathit{pips} = \{1, 2, 3, 4, 5, 6\}$, with a fair / uniform dice distribution $\mathit{dice} = \mathit{unif}_{\mathit{pips}} = \frac{1}{6}|1\rangle + \frac{1}{6}|2\rangle + \frac{1}{6}|3\rangle + \frac{1}{6}|4\rangle + \frac{1}{6}|5\rangle + \frac{1}{6}|6\rangle$. We again use a predicate $\mathit{evenish} \in \mathit{Pred}(\mathit{pips}) = [0, 1]^{\mathit{pips}}$ but now with slightly different probabilities than in Example 4.1.3 (3).

$$\begin{array}{lll}
\mathit{evenish}(1) = \frac{1}{5} & \mathit{evenish}(3) = \frac{1}{10} & \mathit{evenish}(5) = \frac{1}{10} \\
\mathit{evenish}(2) = \frac{9}{10} & \mathit{evenish}(4) = \frac{9}{10} & \mathit{evenish}(6) = \frac{4}{5}
\end{array}$$

It is not hard to see that $\mathit{dice} \models \mathit{evenish} = \frac{1}{2}$.

In a next step we take the predicate *evenish* as evidence and update our *dice* state, giving as

new distribution:

$$\begin{aligned} dice\big|_{evenish} &= \sum_{x \in pips} \frac{dice(x) \cdot evenish(x)}{dice \models evenish} \big|x\rangle \\ &= \frac{1/6 \cdot 1/5}{1/2}\big|1\rangle + \frac{1/6 \cdot 9/10}{1/2}\big|2\rangle + \frac{1/6 \cdot 1/10}{1/2}\big|3\rangle \\ &\qquad + \frac{1/6 \cdot 9/10}{1/2}\big|4\rangle + \frac{1/6 \cdot 1/10}{1/2}\big|5\rangle + \frac{1/6 \cdot 4/5}{1/2}\big|6\rangle \\ &= \tfrac{1}{15}\big|1\rangle + \tfrac{3}{10}\big|2\rangle + \tfrac{1}{30}\big|3\rangle + \tfrac{3}{10}\big|4\rangle + \tfrac{1}{30}\big|5\rangle + \tfrac{4}{15}\big|6\rangle. \end{aligned}$$

As expected, the probabilities for the even pips are now, in the posterior, higher than for the odd ones: the *evenish* evidence has been incorporated.

This illustration is elaborated in ProbLog and in WebPPL in the beginnings of Sections 6.5 and 6.6.

2 The following alarm example is due to Pearl Pearl (1988). It involves an 'alarm' set $A = \{a, a^\perp\}$ and a 'burglary' set $B = \{b, b^\perp\}$, with the following a priori joint distribution $\omega \in \mathcal{D}(A \times B)$.

$$0.000095|a, b\rangle + 0.009999|a, b^\perp\rangle + 0.000005|a^\perp, b\rangle + 0.989901|a^\perp, b^\perp\rangle.$$

Apparently, alarms are very rare, with or without burglary, and the most common situation is described by the last summand: no alarm, no burglary.

The *a priori* burglary distribution is the second marginal:

$$\omega[0, 1] = 0.0001|b\rangle + 0.9999|b^\perp\rangle.$$

Someone reports that the alarm went off, but with only 80% certainty because of deafness. This can be described as a predicate $p\colon A \to [0, 1]$ with $p(a) = 0.8$ and $p(a^\perp) = 0.2$. We can also write this predicate in terms of weighted point predicates: $p = 0.8 \cdot \mathbf{1}_a + 0.2 \cdot \mathbf{1}_{a^\perp}$. There is a 'type' mismatch between p and ω, since p is a predicate on A and ω is a (joint) distribution on the product set $A \times B$. This mismatch can be solved via weakening p to $p \otimes \mathbf{1} = (\pi_1)^*(p)$, so that it becomes a predicate on $A \times B$. Then we can do the update of the joint distribution:

$$\omega|_{p \otimes \mathbf{1}} \in \mathcal{D}(A \times B).$$

In order understand it in detail we first compute the validity:

$$\omega \models p \otimes \mathbf{1} = \sum_{x \in A,\, y \in B} \omega(x, y) \cdot p(x) = 0.206.$$

We can then compute the updated joint state as:

$$\begin{aligned} \omega|_{p \otimes \mathbf{1}} &= \sum_{x \in A,\, y \in B} \frac{\omega(x, y) \cdot p(x)}{\omega \models p} \big|x, y\rangle \\ &= 0.0003688|a, b\rangle + 0.03882|a, b^\perp\rangle + 0.000004853|a^\perp, b\rangle + 0.9608|a^\perp, b^\perp\rangle. \end{aligned}$$

The resulting posterior burglary distribution — with the alarm evidence taken into account — is obtained by taking the second marginal of the updated distribution:

$$(\omega|_{p \otimes \mathbf{1}})[0, 1] = 0.0004|b\rangle + 0.9996|b^\perp\rangle.$$

We see that the burglary probability is four times higher in the posterior than in the prior. What happens is noteworthy: evidence about one component A changes the probabilities in another component B. This 'crossover influence' (in the terminology of Jacobs and Zanasi (2017)) or

'crossover inference' happens precisely because the joint distribution ω is entwined, so that the different parts can influence each other.

3 Updating plays a role in various forms in statistical mechanics. Here we quote an illustration from as standard textbook (Callen, 1985, §16.1) about dices — used as analogy for coupled physical systems.

> A simple analogy is instructive. Consider a set of three dice, one of which is red (the remaining two being white). The three dice have been "thrown" many thousands of times. Whenever the sum of the numbers on the three dice has been 12 (and only then), the number on the red die has been recorded. In what fraction of these recorded throws has the red die shown a one, a two, ..., a six?
>
> The result, left to the reader, is that the red die has shown a one in $\frac{2}{25}$ of the throws, a two $\frac{3}{25}$, ..., a five in $\frac{6}{25}$, and a six in $\frac{5}{25}$ of the recorded throws.

Let's try to model this in the currrent set-up. The throwing of three dices is the parallel product $dice^3 = dice \otimes dice \otimes dice$, where $dice = \sum_{1 \leq i \leq 6} \frac{1}{6}|i\rangle$. The above description says that only the cases where the sum of the three outcomes are relevant. Thus we update this product distribution with the sharp 'twelve' predicate $tw\colon \{1, \ldots, 6\}^3 \to \{0, 1\}$, via:

$$tw(i, j, k) = 1 \iff i + j + k = 12.$$

The above question asks for the outcomes of the red dice. Let's say that the first dice in our product is red. Then, taking the first marginal of the updated distribution yields a distribution that matches the partial answer given in the quote.

$$\Big((dice \otimes dice \otimes dice)\big|_{tw}\Big)[1, 0, 0] = \tfrac{2}{25}|1\rangle + \tfrac{3}{25}|2\rangle + \tfrac{4}{25}|3\rangle + \tfrac{1}{5}|4\rangle + \tfrac{6}{25}|5\rangle + \tfrac{1}{5}|6\rangle.$$

The expression on the left matches the above description of the distribution of interest in a reasonably straightforward manner.

Updating allows us to express new properties of distributions, such as memorylessness of the geometric distribution. This new property informally says that the probability of some future event occurring is not affected by the occurrence of past events. We give a precise formulation below in terms of updating (to events happing later on) and shifting.

The geometric distribution $geo[r] \in \mathcal{D}_\infty(\mathbb{N})$ has infinite support, see Exercise 2.1.15. Updating, as in Definition 6.1.1, also works for such distributions with infinite support.

Proposition 6.1.3 *Let $r \in (0, 1]$ and $n \geq 1$. Write $[- > n]$ for the sharp predicate $\mathbb{N} \to \{0, 1\}$ which with $[m > n] = 1$ iff $m > n$. Then:*

$$geo[r]\big|_{[->n]} = shift(n, geo[r]) = \mathcal{D}_\infty((-) + n)\Big(geo[r]\Big).$$

Proof. We first compute the validity:

$$\begin{aligned} geo[r] \models [- > n] &= \sum_{m>0} geo[r](m) \cdot [m > n] \\ &= \sum_{m>n} r \cdot (1 - r)^{m-1} = (1 - r)^n \cdot \sum_{m>0} r \cdot (1 - r)^{m-1} = (1 - r)^n. \end{aligned}$$

Thus:

$$\begin{aligned}
geo[r]\big|_{[->n]} &= \sum_{m>0} \frac{geo[r](m) \cdot [m > n]}{geo[r] \models [- > n]} \,\big|m\rangle \\
&= \sum_{m>n} \frac{r \cdot (1-r)^{m-1}}{(1-r)^n} \,\big|m\rangle \\
&= \sum_{m>n} r \cdot (1-r)^{m-n-1} \,\big|m\rangle \\
&= \sum_{m>0} r \cdot (1-r)^{m-1} \,\big|m+n\rangle \;=\; \mathcal{D}_\infty((-)+n)\big(geo[r]\big).
\end{aligned}$$

❑

One of the main results about conditioning is Bayes' theorem. We present it here for factors, and not just for sharp predicates (events), as is common.

Theorem 6.1.4 *Let ω be distribution on a sample space X, and let p, q be factors on X.*

1 The product rule holds for conditional validity:

$$\omega|_p \models q \;=\; \frac{\omega \models p \,\&\, q}{\omega \models p}.$$

2 Bayes' rule holds:

$$\omega|_p \models q \;=\; \frac{(\omega|_q \models p) \cdot (\omega \models q)}{\omega \models p}.$$

This result carefully distinguishes a product rule, in item 1, and Bayes' rule, in item (2). This distinction is not always made, since the rules are closely related, and the product rule is sometimes also called Bayes' rule. The latter involves an 'inversion' of updates.

Proof. 1 Assuming that the validity $\omega \models p$ is non-zero, we straightforwardly compute:

$$\begin{aligned}
\omega|_p \models q \;=\; \sum_{x \in X} \omega|_p(x) \cdot q(x) &\overset{(6.1)}{=} \sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \models p} \cdot q(x) \\
&= \frac{\sum_{x \in X} \omega(x) \cdot p(x) \cdot q(x)}{\omega \models p} \\
&= \frac{\sum_{x \in X} \omega(x) \cdot (p \,\&\, q)(x)}{\omega \models p} \;=\; \frac{\omega \models p \,\&\, q}{\omega \models p}.
\end{aligned}$$

2 This follows directly by using the previous item twice, in combination with the commutativity of conjunction &, in:

$$\omega|_p \models q \overset{(1)}{=} \frac{\omega \models p \,\&\, q}{\omega \models p} = \frac{\omega \models q \,\&\, p}{\omega \models p} \overset{(1)}{=} \frac{(\omega|_q \models p) \cdot (\omega \models q)}{\omega \models p}.$$

❑

Example 6.1.5 We instantiate Proposition 6.1.4 with sharp predicates $\mathbf{1}_E, \mathbf{1}_D$ for subsets / events $E, D \subseteq X$. Then the familiar formulations of the product / Bayes rule appear.

1 The product rule specialises to the definition of conditional probability:

$$P(E \mid D) \;=\; \omega|_{\mathbf{1}_D} \models \mathbf{1}_E \;=\; \frac{\omega \models \mathbf{1}_D \,\&\, \mathbf{1}_E}{\omega \models \mathbf{1}_D} \;=\; \frac{\omega \models \mathbf{1}_{D \cap E}}{P(D)} \;=\; \frac{P(D \cap E)}{P(D)}.$$

2 Bayes rule, in the general formulation of Proposition 6.1.4 (2) specialises to the well known inversion property of conditional probabilities:

$$P(E \mid D) = \omega|_{\mathbf{1}_D} \models \mathbf{1}_E = \frac{(\omega|_{\mathbf{1}_E} \models \mathbf{1}_D) \cdot (\omega \models \mathbf{1}_D)}{\omega \models \mathbf{1}_E} = \frac{P(D \mid E) \cdot P(E)}{P(D)}.$$

We have explained updating $\omega|_p$ as incorporating the evidence p into the distribution ω. Thus, one expects p to be 'more true' in $\omega|_p$ than in ω. The next result shows that this is indeed the case. It plays an important role in learning see Chapter 8.

Theorem 6.1.6 *Let ω be a distribution and p a factor on the same set, with non-zero validity $\omega \models p$. Updating with p then gives an increase of validity:*

$$\omega|_p \models p \;\geq\; \omega \models p. \tag{6.2}$$

If p is a predicate and the validity $\omega \models p$ is not one, updating with the orthosupplement $p^\perp$ gives a decrease of validity:

$$\omega|_{p^\perp} \models p \;\leq\; \omega \models p.$$

The restriction from factors to predicates is required for the second inequality because the orthosupplement is defined only for predicates. This restriction is not essential, since a factor can always be scaled down to a predicate, at least on the support of the distribution at hand, without affecting the update, see the discussion before Example 6.1.9 below.

Proof. We recall the inequality $\omega \models p \;\&\; p \;\geq\; (\omega \models p)^2$ from Corollary 5.1.4 (1), or from Exercise 4.4.6 (2). Then, by the product rule from Theorem 6.1.4 (1),

$$\omega|_p \models p = \frac{\omega \models p \;\&\; p}{\omega \models p} \geq \frac{(\omega \models p)^2}{\omega \models p} = \omega \models p.$$

We use this increase of validity to prove the decrease through updating with $p^\perp$. It follows from:

$$1 - \big(\omega|_{p^\perp} \models p\big) = \omega|_{p^\perp} \models p^\perp \geq \omega \models p^\perp = 1 - \big(\omega \models p\big). \qquad \square$$

We consider the validity increase in (6.2) of p after Bayesian updating with p to be an essential and characteristic feature of updating. Astonishingly, the two inequalities in the above theorem do not (seem to) occur in the literature on probability theory. The cause of this absence may be the (bad) habit of leaving distributions implicit. Then the first inequality in the above theorem reduces to $\mathbb{E}[p] \leq \mathbb{E}[p]$ and becomes meaningless. It could have emerged in a more meaningful form as:

$$\mathop{\mathbb{E}}_{x \sim \omega|_p} p(x) \;\geq\; \mathop{\mathbb{E}}_{x \sim \omega} p(x).$$

But this formulation only makes sense if one has explicit notation for the updated distribution $\omega|_p$. Another form would be $\mathbb{E}[p^2] \geq \mathbb{E}[p]^2$ where the same distribution is used implicitly on both sides. The latter inequality is more common but it lacks the connection to updating. Thus, we view the update inequalities in the previous theorem as a motivation to update notation.

Another reason for the absence of the inequality (6.2) may be that it is trivial for the special case when the predicate p is sharp, see Exercise 6.1.2 below. If one restricts oneself to sharp predicates, the increase may not be seen as a special / relevant result.

We add a few more basic facts about conditioning.

Lemma 6.1.7 *Let ω be distribution on X, with factors $p, q \in \mathrm{Fact}(X)$.*

1 Conditioning with truth has no effect:

$$\omega|_{\mathbf{1}} = \omega.$$

2 Conditioning with a point predicate gives a point state: for $a \in X$,

$$\omega|_{\mathbf{1}_a} = 1|a\rangle, \qquad \textit{assuming } \omega(a) \neq 0.$$

3 Iterated conditionings commute and can be performed via conjunction:

$$(\omega|_p)|_q = \omega|_{p\&q} = (\omega|_q)|_p.$$

4 Conditioning is stable under multiplication of the factor with a scalar $s > 0$:

$$\omega|_{s\cdot p} = \omega|_p.$$

5 Conditioning can be done component-wise, for product states and factors:

$$(\sigma \otimes \tau)\big|_{(p\otimes q)} = (\sigma|_q) \otimes (\tau|_p).$$

6 Marginalisation of a conditioning with a (similarly) weakened predicate is conditioning of the marginalised state:

$$\omega|_{\mathbf{1}\otimes q}[0,1] = \omega[0,1]\big|_q.$$

7 For a function $f\colon X \to Y$, used as deterministic channel,

$$(f_*(\omega))\big|_q = f_*(\omega\big|_{q\circ f}).$$

When we ignore undefinedness issues, we see that items (1) and (3) show that conditioning is an action on distributions, namely of the monoid of factors with conjunction $(\mathbf{1}, \&)$, see Definition 1.4.4. The fact that updates commute, in item (3) shows that when one updates with multiple factors, the order of these factors does not matter, so that these factors, when take together, are best organised as a multiset of factors. This will be elaborated further in Chapter 8. The human mind, in contrast, is highly sensitive to the order in which it is updated (or 'primed'), see *e.g.* Jacobs (2025a) for a discussion.

Notice the difference between updating and marginalisation $\omega|_{\mathbf{1}\otimes q}[0,1]$ in the *same* component in item (6), and updating and marginalisation $\omega|_{p\otimes\mathbf{1}}[0,1]$ in *different* components in Example 6.1.5. The latter gives rise to crossover effects.

Proof. 1 Trivial since $\omega \models \mathbf{1} = 1$.

2 Assuming $\omega(a) \neq 0$ we get for each $x \in X$,

$$\omega|_{\mathbf{1}_a}(x) = \frac{\omega(x)\cdot \mathbf{1}_a(x)}{\omega \models \mathbf{1}_a} = \frac{\omega(a)\cdot 1|a\rangle(x)}{\omega(a)} = 1|a\rangle(x).$$

3 It suffices to prove:

$$\begin{aligned}(\omega|_p)|_q(x) &= \frac{\omega|_p(x)\cdot q(x)}{\omega|_p \models q} = \frac{\omega(x)\cdot p(x)/\omega\models p \cdot q(x)}{\omega\models p\&q/\omega\models p} \quad \text{by Proposition 6.1.4 (1)}\\ &= \frac{\omega(x)\cdot (p \,\&\, q)(x)}{\omega \models p \,\&\, q} = \omega|_{p\&q}(x).\end{aligned}$$

4 First we have:

$$\omega \models s \cdot p \;=\; \sum_{x \in X} \omega(x) \cdot (s \cdot p)(x) \;=\; \sum_{x \in X} \omega(x) \cdot s \cdot p(x) \;=\; s \cdot \left(\sum_{x \in X} \omega(x) \cdot p(x) \right) \;=\; s \cdot (\omega \models p).$$

Next:

$$\omega|_{s \cdot p}(x) \;=\; \frac{\omega(x) \cdot (s \cdot p)(x)}{\omega \models s \cdot p} \;=\; \frac{\omega(x) \cdot s \cdot p(x)}{s \cdot (\omega \models p)} \;=\; \frac{\omega(x) \cdot p(x)}{\omega \models p} \;=\; \omega|_p(x).$$

5 For states $\sigma \in \mathcal{D}(X)$, $\tau \in \mathcal{D}(Y)$ and factors p on X and q on Y one has:

$$\begin{aligned}
\Big((\sigma \otimes \tau)\big|_{(p \otimes q)}\Big)(x, y) \;=\; \frac{(\sigma \otimes \tau)(x, y) \cdot (p \otimes q)(x, y)}{(\sigma \otimes \tau) \models (p \otimes q)} &= \frac{\sigma(x) \cdot \tau(y) \cdot p(x) \cdot q(y)}{(\sigma \models p) \cdot (\tau \models q)} \qquad \text{by Lemma 4.4.8} \\
&= \frac{\sigma(x) \cdot p(x)}{\sigma \models p} \cdot \frac{\tau(y) \cdot q(y)}{\tau \models q} \\
&= (\sigma|_p)(x) \cdot (\tau|_q)(y) \\
&= \Big((\sigma|_p) \otimes (\tau|_q)\Big)(x, y).
\end{aligned}$$

6 Let $\omega \in \mathcal{D}(X \times Y)$ and q be a factor on Y; then for an element $y \in Y$,

$$\begin{aligned}
(\omega|_{\mathbf{1} \otimes q}[0, 1])(y) \;=\; \sum_{x \in X} \omega|_{\mathbf{1} \otimes q}(x, y) \;&=\; \sum_{x \in X} \frac{\omega(x, y) \cdot (\mathbf{1} \otimes q)(x, y)}{\omega \models \mathbf{1} \otimes q} \\
&\overset{(4.11)}{=} \frac{(\omega[0, 1])(y) \cdot q(y)}{\omega[0, 1] \models q} \;=\; (\omega[0, 1]\big|_q)(y).
\end{aligned}$$

7 For $f \colon X \to Y$, $\omega \in \mathcal{D}(X)$ and $q \in Fact(Y)$,

$$\begin{aligned}
(f_*(\omega)|_q)(y) \;=\; \mathcal{D}(f)(\omega)|_q(y) \;&=\; \frac{\mathcal{D}(f)(\omega)(y) \cdot q(y)}{\mathcal{D}(f)(\omega) \models q} \\
&= \sum_{x \in f^{-1}(y)} \frac{\omega(x) \cdot q(y)}{\sum_{x \in f^{-1}(y)} \omega(x) \cdot p(y)} \\
&= \sum_{x \in f^{-1}(y)} \frac{\omega(x) \cdot q(f(x))}{\sum_x \omega(x) \cdot q(f(x))} \\
&= \sum_{x \in f^{-1}(y)} \omega|_{q \circ f}(x) \;=\; \mathcal{D}(f)(\omega|_{q \circ f})(y) \;=\; f_*(\omega|_{q \circ f})(y). \qquad \square
\end{aligned}$$

We can now generalise some earlier results.

Proposition 6.1.8 *Let $\omega \in \mathcal{D}(X)$ and $p \in Fact(X)$. We recall that for $n \in \mathbb{N}$ we write $p^n = p \,\&\, \cdots \,\&\, p$, with $p^0 = \mathbf{1}$.*

1 For all $n \in \mathbb{N}$,

$$\omega \models p^n \;\geq\; (\omega \models p)^n.$$

2 For all $n, m, k \in \mathbb{N}$ with $n \geq m \geq 0$,

$$\omega|_{p^n} \models p^k \;\geq\; \omega|_{p^m} \models p^k.$$

3 For all $n, m \in \mathbb{N}$,

$$\omega \models p^{n+m} \;\geq\; (\omega \models p^n) \cdot (\omega \models p^m).$$

4 *When p is a predicate,*

$$\omega \models p^{\perp} \mathbin{\&} p^n \;\leq\; (\omega \models p^{\perp}) \cdot (\omega \models p^n).$$

The first point generalises Corollary 5.1.4 (1), which covers the special case where $n = 2$. The second point generalises Theorem 6.1.6, which deals with the special case where $n = k = 1$ and $m = 0$. The inequality in point (4) is trivial when p is a sharp predicate, but not for a general, fuzzy predicate.

Proof. 1 We use induction on $n \in \mathbb{N}$. The cases for $n = 0$ and $n = 1$ are trivial and the case $n = 2$ is covered by Corollary 5.1.4 (1). Next, for $n \geq 2$,

$$\begin{aligned}
\omega \models p^{n+1} \;=\; \omega \models p \mathbin{\&} p^n \;&=\; (\omega|_p \models p^n) \cdot (\omega \models p) && \text{by Theorem 6.1.4 (1)} \\
&\geq\; (\omega|_p \models p)^n \cdot (\omega \models p) && \text{by induction hypothesis} \\
&\geq\; (\omega \models p)^n \cdot (\omega \models p) && \text{by Theorem 6.1.6} \\
&=\; (\omega \models p)^{n+1}.
\end{aligned}$$

2 We first prove the special case for $n = 1$ and $m = 0$, of the form:

$$\omega|_p \models p^k \;\geq\; \omega \models p^k. \tag{$*$}$$

This can be done via induction on $k \in \mathbb{N}$, where the base case $k = 0$ is trivial. Next:

$$\begin{aligned}
\omega|_p \models p^{k+1} \;=\; \omega|_p \models p \mathbin{\&} p^k \;&=\; (\omega|_p \models p) \cdot (\omega|_p|_p \models p^k) && \text{by Theorem 6.1.4 (1)} \\
&\geq\; (\omega \models p) \cdot (\omega|_p \models p^k) && \text{by Theorem 6.1.6 and IH} \\
&=\; \omega \models p^{k+1} && \text{by Theorem 6.1.4 (1) again.}
\end{aligned}$$

We now prove the inequality $\omega|_{p^n} \models p^k \;\geq\; \omega|_{p^m} \models p^k$ in item (2), We use induction on n. In the base case, when $n = 0$ we must have $m = 0$, so the statement is trivial.

For the induction step, the special case $m = 0$ is handled by repeated application of $(*)$, using that $\omega|_{p^0} = \omega|_{\mathbf{1}} = \omega$ and $\omega|_{p^{n+1}} = \omega|_p \ldots|_p$, see Lemma 6.1.7 (1), (3). Thus we assume $0 < m \leq n + 1$ in:

$$\begin{aligned}
\omega|_{p^{n+1}} \models p^k \;=\; \omega|_{p \& p^n} \models p^k \;=\; \omega|_p|_{p^n} \models p^k \;&\geq\; \omega|_p|_{p^{m-1}} \models p^k && \text{by IH, with } 0 \leq m-1 \leq n \\
&=\; \omega|_{p^m} \models p^k.
\end{aligned}$$

3 When the validity $\omega \models p^n$ is zero the result immediately holds. If this validity is non-zero, we can reason as follows.

$$\begin{aligned}
\frac{\omega \models p^{n+m}}{\omega \models p^n} \;=\; \frac{\omega \models p^n \mathbin{\&} p^m}{\omega \models p^n} \;&=\; \omega|_{p^n} \models p^m && \text{by Theorem 6.1.4 (1)} \\
&\geq\; \omega \models p^m && \text{by the previous point.}
\end{aligned}$$

4 We use item (3) in:

$$\begin{aligned}
\omega \models p^{\perp} \mathbin{\&} p^n \;=\; \omega \models (\mathbf{1} - p) \mathbin{\&} p^n \;&=\; (\omega \models p^n) - (\omega \models p^{n+1}) \\
&\leq\; (\omega \models p^n) - (\omega \models p^n) \cdot (\omega \models p) \\
&=\; (\omega \models p^n) \cdot \big(1 - (\omega \models p)\big) \;=\; (\omega \models p^n) \cdot (\omega \models p^{\perp}). \qquad \square
\end{aligned}$$

In the beginning of this section we have defined updating $\omega|_p$ for a state $\omega \in \mathcal{D}(X)$ and a *factor* $p \colon X \to \mathbb{R}_{\geq 0}$. Now let's assume that this factor p is bounded: there is a bound $B \in \mathbb{R}_{>0}$ such that $p(x) \leq B$ for all $x \in X$. The rescaled factor $\frac{1}{B} \cdot p$ is then a predicate. Proposition 6.1.7 (4) shows

that updating with the factor p is the same as updating with the predicate $\frac{1}{B} \cdot p$. An additional observation is that when we restrict the factor p to the (finite) support $supp(\omega) \subseteq X$ of the distribution at hand, then it is bounded, for instance with bound $B = \max\{p(x) \mid x \in supp(\omega)\}$. Hence we do not loose much if we restrict updating to predicates. Nevertheless it is most convenient to define updating for factors so that we do not have to bother about any rescaling.

We include one more example in this section.

Example 6.1.9 Recall the simple question we had in Remark 2.2.1: two urns of the same size K, are filled with red (R) and green (G) balls, where the only thing that we know is that the first urn has more red balls. The aim is to show that the probability of drawing a red ball is higher from the first urn than from the second urn.

Let's see how our update mechanism handles this situation. The number of red balls in an urn is in the set $X := \{0, 1, \ldots, K\}$. Since this number is unknown, we will work with the uniform distribution $unif_X = \sum_{0 \leq i \leq K} \frac{1}{K+1}|i\rangle \in \mathcal{D}(X)$, for both urns. There is a predicate $red\colon X \to [0, 1]$, namely $red(i) = \frac{i}{K}$. It thus gives the likelihood of a red ball. It is not hard to see that *a priori* we have:

$$unif_X \models red \;=\; \tfrac{1}{2}.$$

We use the order $\geq$ on X as a sharp predicate $geq\colon X \times X \to [0, 1]$, where $geq(x, y) = 1$ iff $x \geq y$. We can then form the updated joint state:

$$(unif_X \otimes unif_X)\big|_{geq} \in \mathcal{D}(X \times X).$$

It incorporates the given information that the first urn contains more red balls than the second one. By taking the first and second marginals we obtain the updated orginal urns, for which we can ask the expectation of a red ball. Independently of the size K of the urns we get:

$$\Big((unif_X \otimes unif_X)\big|_{geq}\Big)[1, 0] \models red \;=\; \tfrac{2}{3}$$
$$\Big((unif_X \otimes unif_X)\big|_{geq}\Big)[0, 1] \models red \;=\; \tfrac{1}{3}.$$

Drawing red from the first urn now clearly has a higher probability. Details of the verification are left as an exercise below.

We conclude with a general observation. Updating $\omega|_p$ is a partial operation, since it is undefined when the validity $\omega \models p$ of the factor p is zero. One can circumvent these partiality problems by restricting to *positive* (or *non-zero*) factors, taking values in the set $\mathbb{R}_{>0}$ of positive real numbers. We shall write $Fact_*(X)$ for the set of positive factors on X. Explicitly:

$$Fact_*(X) := \{p \in Fact(X) \mid \forall x \in X.\, p(x) > 0\}. \tag{6.3}$$

The multiplicative monoid $(Fact_*(X), \&, \mathbf{1})$ of positive factors is actually a group, with the reciprocal $p \mapsto \frac{1}{p}$ as inverse operation. This multiplicative group of positive factors is isomorphic to the additive group $(Obs(X), +, 0, -)$ of observables, for instance via pointwise e-powers and natural exponents:

$$\begin{tikzcd} Fact_*(X) \arrow[r, bend left, "\ln"] \arrow[r, phantom, "\cong"] & Obs(X) \arrow[l, bend left, "e^{(-)}"] \end{tikzcd} \tag{6.4}$$

This is a pointwise extension of the group isomorphism of Exercise 1.4.2. But there is a bit more to say.

Proposition 6.1.10 *The update mapping* $(p, \omega) \mapsto \omega|_p$ *forms a (multiplicative) group action* $upd\colon Fact_*(X) \times \mathcal{D}(X) \to \mathcal{D}(X)$.

Proof. The mapping $upd(p, \omega) := \omega|_p$ is total since the update $\omega|_p$ is always defined for a positive factor $p \in Fact_*(X)$, because $\omega \models p > 0$. The monoid / group action equations from Definition 1.4.4 hold for *upd* by Lemma 6.1.7 (1) and (3). ❑

Interestingly, updating with a positive factor $p \in Fact_*(X)$ can be undone via updating with the reciprocal $\frac{1}{p}$ since, again by Lemma 6.1.7 (3) and (1),

$$\big(\omega|_p\big)\big|_{\frac{1}{p}} = \omega|_{p \,\&\, \frac{1}{p}} = \omega|_{\mathbf{1}} = \omega.$$

Exercises

6.1.1 Consider the following girls / boys riddle: given that a family with two children has a boy, what is the probability that the other child is a girl? Take as space $\{G, B\}$. On it we use the uniform distribution $unif = \frac{1}{2}|G\rangle + \frac{1}{2}|B\rangle$ since there is no prior knowledge about the gender of the chidren.

1 Take as 'at least one girl' and 'at least one boy' predicates on $\{G, B\} \times \{G, B\}$:

$$\begin{aligned} g &:= (\mathbf{1}_B \otimes \mathbf{1}_B)^\perp = (\mathbf{1}_G \otimes \mathbf{1}) \ovee (\mathbf{1}_B \otimes \mathbf{1}_G) \\ b &:= (\mathbf{1}_G \otimes \mathbf{1}_G)^\perp = (\mathbf{1}_B \otimes \mathbf{1}) \ovee (\mathbf{1}_G \otimes \mathbf{1}_B). \end{aligned}$$

Compute $unif \otimes unif \models g$ and $unif \otimes unif \models b$.

2 Check that $unif \otimes unif\big|_b \models g = \frac{2}{3}$.

(For a description and solution of this problem in a special library for probabilistic programming of the functional programming language Haskell, see Erwig and Walkingshaw (2009).)

6.1.2 Let $\omega \in \mathcal{D}(X)$ be a distribution with a non-empty subset $U \subseteq supp(\omega)$. Show that:

$$\omega|_{\mathbf{1}_U} \models \mathbf{1}_U = 1.$$

One can prove this via unpacking the relevant definitions, or via Theorem 6.1.4 (1).

Conclude from this equation that the validity increase (6.2) is trivial for a sharp predicate.

6.1.3 Check that $\omega|_p$ can be described as $Flrn(\omega \bullet p)$, using the action $\bullet$ from Exercise 4.4.14.

6.1.4 Check that the memorylessness of Proposition 6.1.3 essentially involves having the same rates between successive probabilities in the geometric distribution:

$$\frac{geo[r](n)}{geo[r](n+1)} = \frac{geo[r](n+1)}{geo[r](n+2)}.$$

6.1.5 In the setting of Example 6.1.2 (1) define a new predicate $oddish = evenish^\perp = \mathbf{1} - evenish$.

1 Compute $dice|_{oddish}$

2 Prove the equation below, involving a convex sum of states on the left-hand side.

$$(dice \models evenish) \cdot \omega|_{evenish} + (dice \models oddish) \cdot \omega|_{oddish} = dice.$$

6.1.6 We generalise the situation of the previous exercise. Let $p_1, \dots, p_n \in Pred(X)$ be a test, *i.e.* an n-tuple of predicates on X with $p_1 \owedge \cdots \owedge p_n = \mathbf{1}$. Let $\omega \in \mathcal{D}(X)$ and $q \in Fact(X)$.

1 Check that $1 = \sum_i (\omega \models p_i)$.

2 Prove what is called the *law of total probability*:

$$\omega \;=\; \sum_{1 \leq i \leq n} (\omega \models p_i) \cdot \omega|_{p_i}. \tag{6.5}$$

What happens to the expression on the right-hand side if one of the p_i has validity zero? Check that this equation generalises Exercise 6.1.5.

(The expression on the right-hand side in (6.5) is used to turn a test into a 'denotation' function $\mathcal{D}(X) \to \mathcal{D}(\mathcal{D}(X))$ in McIver et al. (2015, 2014), namely as $\omega \mapsto \sum_i (\omega \models p_i) \big| \omega|_{p_i} \rangle$. This proces is described more abstractly in terms of 'hypernormalisation' in Jacobs (2017a); Garner (2023).)

3 Show that:

$$\omega \models q \;=\; \sum_{1 \leq i \leq n} \omega \models q \,\&\, p_i.$$

4 Prove now:

$$\omega|_q \models p_i \;=\; \frac{\omega \models q \,\&\, p_i}{\sum_j \omega \models q \,\&\, p_j}.$$

5 And also:

$$\omega|_q \models p_i \;=\; \frac{(\omega|_{p_i} \models q) \cdot (\omega \models p_i)}{\sum_j (\omega|_{p_j} \models q) \cdot (\omega \models p_j)}.$$

This last equation if called Bayes' law in (Rice, 2006, §1.5).

6.1.7 Let $p \colon X \to \mathbb{R}_{\geq 0}$ be a non-zero factor, on a finite set X. Check that updating the uniform distribution on X with p, as in:

$$unif_X|_p,$$

is a way of turning the factor p into a distribution on X via normalisation.

6.1.8 Let $\omega \in \mathcal{D}(X)$ be a distribution and let p_1, p_2 be factors on X, both with non-zero validity. Show that:

$$\tfrac{1}{2} \cdot \omega|_{p_1} + \tfrac{1}{2} \cdot \omega|_{p_2} \;=\; \omega|_q \qquad \text{for} \qquad q(x) \;=\; (\omega \models p_2) \cdot p_1(x) + (\omega \models p_1) \cdot p_2(x).$$

6.1.9 Show also that updating a state with a linear combination of factors produces a convex sum of updated states, as in:

$$\omega|_{r \cdot p + s \cdot q} \;=\; \frac{r \cdot (\omega \models p)}{r \cdot (\omega \models p) + s \cdot (\omega \models q)} \cdot \omega|_p + \frac{s \cdot (\omega \models q)}{r \cdot (\omega \models p) + s \cdot (\omega \models q)} \cdot \omega|_q,$$

for a state $\omega \in \mathcal{D}(X)$, factors $p, q \colon X \to \mathbb{R}_{\geq 0}$ and scalars $r, s \in \mathbb{R}_{\geq 0}$.

Check that the previous exercise is a special case.

6.1.10 Show that updating a convex sum of states yields a convex sum of updated states: for $\sigma, \tau \in \mathcal{D}(X)$, $p \in Fact(X)$ and $r, s \in [0, 1]$ with $r + s = 1$,

$$\begin{aligned}
(r \cdot \sigma + s \cdot \tau)\big|_p \;&=\; \frac{r \cdot (\sigma \models p)}{r \cdot (\sigma \models p) + s \cdot (\tau \models p)} \cdot \sigma|_p \;+\; \frac{s \cdot (\tau \models p)}{r \cdot (\sigma \models p) + s \cdot (\tau \models p)} \cdot \tau|_p \\
&=\; \frac{r \cdot (\sigma \models p)}{(r \cdot \sigma + s \cdot \tau) \models p} \cdot \sigma|_p \;+\; \frac{s \cdot (\tau \models p)}{(r \cdot \sigma + s \cdot \tau) \models p} \cdot \tau|_p.
\end{aligned}$$

6.1.11 Consider $\omega \in \mathcal{D}(X)$ with a *sharp* predicate $p\colon X \to \{0,1\}$.

1 Check that $p^2 = p$;

2 Assume $\omega \models p$ is non-zero, for the remainder of this exercise; show that:

$$\omega|_p \models p^\perp \;=\; 0.$$

3 Show also that the variance satisfies:

$$\mathit{Var}(\omega, p) \;=\; (\omega \models p) \cdot (\omega \models p^\perp).$$

4 Conclude that in this sharp case: $\mathit{Var}(\omega|_p, p) = 0$.

6.1.12 We have seen the validity increase $\omega|_p \models p \geq \omega \models p$ via updating, in Theorem 6.1.6. This does not work 'in context', in presence of another factor q, that is:

$$\textit{not} \text{ for all } p, q, \quad \omega|_p \models p \,\&\, q \;\geq\; \omega \models p \,\&\, q.$$

Check this, for instance for $\omega = \frac{1}{2}|a\rangle + \frac{1}{3}|b\rangle + \frac{1}{6}|c\rangle$ with $p = \frac{3}{5} \cdot \mathbf{1}_a + \frac{4}{5} \cdot \mathbf{1}_b + \frac{1}{2} \cdot \mathbf{1}_c$ and $q = \frac{2}{5} \cdot \mathbf{1}_a + \frac{1}{4} \cdot \mathbf{1}_b + \frac{3}{4} \cdot \mathbf{1}_c$.

6.1.13 Updating with an arbitrary predicate may both increase and decrease variance, as this exercise illustrates. Let $X = \{a, b, c\}$.

1 Take $\omega_1 = \frac{2}{5}|a\rangle + \frac{3}{10}|b\rangle + \frac{3}{10}|c\rangle$ and $p_1 = \frac{1}{2} \cdot \mathbf{1}_a + \frac{7}{10} \cdot \mathbf{1}_b + \frac{4}{5} \cdot \mathbf{1}_c$. Show that $\mathit{Var}(\omega_1|_{p_1}, p_1) < \mathit{Var}(\omega_1, p_1)$.

2 Now take $\omega_2 = \frac{1}{4}|a\rangle + \frac{1}{2}|b\rangle + \frac{1}{4}|c\rangle$ and $p_2 = \frac{1}{10} \cdot \mathbf{1}_a + \frac{1}{5} \cdot \mathbf{1}_b + \frac{4}{5} \cdot \mathbf{1}_c$. Show that $\mathit{Var}(\omega_2|_{p_2}, p_2) > \mathit{Var}(\omega_2, p_2)$.

6.1.14 This exercise will demonstrate that conditioning may both create and remove entwinedness of distributions.

1 Write $\mathit{yes} = \mathbf{1}_1 \colon \mathbf{2} \to [0,1]$, where $\mathbf{2} = \{0,1\}$, and $\mathit{no} = \mathit{yes}^\perp = \mathbf{1}_0$. Prove that the following conditioning of a non-entwined state,

$$\tau \coloneqq (\mathit{flip} \otimes \mathit{flip})\big|_{(\mathit{yes}\otimes\mathit{yes})\heartsuit(\mathit{no}\otimes\mathit{no})}$$

is entwined.

2 Consider the state $\omega \in \mathcal{D}(\mathbf{2} \times \mathbf{2} \times \mathbf{2})$ given by:

$$\omega \;=\; \tfrac{1}{18}|111\rangle + \tfrac{1}{9}|110\rangle + \tfrac{2}{9}|101\rangle + \tfrac{1}{9}|100\rangle + \tfrac{1}{9}|011\rangle + \tfrac{2}{9}|010\rangle + \tfrac{1}{9}|001\rangle + \tfrac{1}{18}|000\rangle.$$

Prove that ω's first and third component are entwined:

$$\omega[1,0,1] \;\neq\; \omega[1,0,0] \otimes \omega[0,0,1].$$

3 Now let ρ be the following conditioning of ω:

$$\rho \;\coloneqq\; \omega|_{\mathbf{1}\otimes\mathit{yes}\otimes\mathbf{1}}.$$

Prove that ρ's first and third component are non-entwined.

The phenomenon that entwined states become non-entwined via conditioning is called *screening-off*, whereas the opposite, non-entwined states becoming entwined via conditioning, is called *explaining away*.

6.1.15 Show that for $\omega \in \mathcal{D}(X)$ and $p_1, p_2 \in \mathit{Fact}(X)$ one has:

$$(\Delta_*(\omega))\big|_{p_1 \otimes p_2} = \Delta_*(\omega|_{p_1 \& p_2}).$$

Note that this is a consequence of Lemma 6.1.7 (7).

6.1.16 We have mentioned (right after Definition 6.1.1) that updating of the identity channel has no effect. Prove more generally that for an ordinary function $f\colon X \to Y$, updating the associated deterministic channel $\langle f \rangle\colon X \rightsquigarrow Y$ has no effect:

$$\langle f \rangle|_q = \langle f \rangle.$$

6.1.17 Let $c\colon Z \rightsquigarrow X$ and $d\colon Z \rightsquigarrow Y$ be two channels with a common domain Z, and with factors $p \in \mathit{Fact}(X)$ and $q \in \mathit{Fact}(Y)$ on their codomains. Prove that the update of a tuple channel is the tuple of the updates:

$$\langle c, d \rangle|_{p \otimes q} = \langle c|_p, d|_q \rangle.$$

Prove also that for $e\colon U \rightsquigarrow X$ and $f\colon V \rightsquigarrow Y$,

$$(e \otimes f)|_{p \otimes q} = (e|_p) \otimes (f|_q).$$

6.1.18 In Jacobs and Zanasi (2017) the influence of a predicate p on a state ω is measured via the total variation distance $d(\omega, \omega|_p)$. This influence can be zero, for the truth predicate $p = \mathbf{1}$.

Consider the set $\{H, T\}$ with state $\mathit{flip}(r) = r|H\rangle + (1-r)|T\rangle$, and with predicate $p = \mathbf{1}_H$. Prove that $d(\mathit{flip}(r), \mathit{flip}(r)|_p) \to 1$ as $r \to 0$.

6.1.19 Consider the situation in Example 6.1.9 and prove consecutively:

1 $(\mathit{unif}_X \otimes \mathit{unif}_X)\big|_{\mathit{geq}} = \displaystyle\sum_{i,j \in X, i \geq j} \frac{2}{(K+1)(K+2)} \big|i, j\rangle$;

2 the first and second marginals are:

$$\Big((\mathit{unif}_X \otimes \mathit{unif}_X)\big|_{\mathit{geq}}\Big)[1,0] = \sum_{0 \leq i \leq K} \frac{2(i+1)}{(K+1)(K+2)} \big|i\rangle$$
$$\Big((\mathit{unif}_X \otimes \mathit{unif}_X)\big|_{\mathit{geq}}\Big)[0,1] = \sum_{0 \leq i \leq K} \frac{2(K+1-i)}{(K+1)(K+2)} \big|i\rangle$$

3 the validities of the predicate *red* in these marginals are:

$$\Big((\mathit{unif}_X \otimes \mathit{unif}_X)\big|_{\mathit{geq}}\Big)[1,0] \models \mathit{red} = \tfrac{2}{3}$$
$$\Big((\mathit{unif}_X \otimes \mathit{unif}_X)\big|_{\mathit{geq}}\Big)[0,1] \models \mathit{red} = \tfrac{1}{3}.$$

Hint: Recall Proposition 1.2.6 (1), and (2).

6.1.20 Check that Lemma 6.1.7 (7) implies that the update group action $\mathit{upd}\colon \mathit{Fact}_*(X) \times \mathcal{D}(X) \to \mathcal{D}(X)$ from Proposition 6.1.10 is a 'dinatural transformation', from the functor $\mathit{Fact}_*(-) \times \mathcal{D}(-)\colon \mathbf{Sets}^{\mathrm{op}} \times \mathbf{Sets} \to \mathbf{Sets}$ to the second projection $\mathit{proj}_2\colon \mathbf{Sets}^{\mathrm{op}} \times \mathbf{Sets} \to \mathbf{Sets}$. This means that for a function $f\colon X \to Y$ the following diagram commutes.

$$\begin{tikzcd}
 & \mathit{Fact}_*(X) \times \mathcal{D}(X) \arrow[r, "\mathit{upd}"] & \mathcal{D}(X) \arrow[dr, "\mathcal{D}(f)"] & \\
\mathit{Fact}_*(Y) \times \mathcal{D}(X) \arrow[ur, bend left, "((-)\circ f)\times \mathit{id}"] \arrow[dr, bend right, "\mathit{id} \times \mathcal{D}(f)"'] & & & \mathcal{D}(Y) \\
 & \mathit{Fact}_*(Y) \times \mathcal{D}(Y) \arrow[r, "\mathit{upd}"] & \mathcal{D}(Y) \arrow[ur, "\mathit{id}"'] &
\end{tikzcd}$$

6.2 Examples of forward and backward inference

Forward transformation of states / distributions and backward transformation of observables can be combined with updating of states. This combination gives rise to the powerful techniques of forward and backward inference. The current section defines these forms of inference and then elaborates many illustrations. The mathematical analysis of forward and backward inference is postponed to the next section.

The next definition captures the two basic patterns — first formulated in this form in Jacobs and Zanasi (2016). We shall refer to them jointly as channel-based inference, or as reasoning along channels.

Definition 6.2.1 Let $\omega \in \mathcal{D}(X)$ be a distribution on the domain of a channel $c\colon X \rightsquigarrow Y$.

1 For a factor $p \in Fact(X)$, we define *forward inference* as transformation along c of the state ω updated with p, as in:

$$c_*(\omega|_p) \in \mathcal{D}(Y).$$

This is also called *causal reasoning*.

2 For a factor $q \in Fact(Y)$, *backward inference* is updating of ω with the transformed factor:

$$\omega|_{c^*(q)} \in \mathcal{D}(X).$$

This is sometimes called *explanation* or *evidential reasoning*. We shall also refer this operation as *Pearl's update rule*, in contrast with Jeffrey's update rule, to be discussed in Section 7.9.

In both cases the distribution ω is often called the *prior distribution* or simply the *prior*. Similarly, $c_*(\omega|_p)$ and $\omega|_{c^*(q)}$ are called *posterior distributions* or just *posteriors*.

Thus, with forward inference one first conditions and then performs (forward, state) transformation, whereas for backward inference one first performs (backward, factor) transformation, and then one conditions. The next result shows that backward inference produces a validity increase.

In the setting of the second item (2), *abductive inference*, see *e.g.* Dice et al. (2021), involves computing argmax $(\omega|_{c^*(q)})$, which yields the element in X that best explains the evidence q, along the channel c with prior ω.

Theorem 6.2.2 *The validity of a factor q in a predicted state $c_*(\omega)$ is increased when ω is replaced by $\omega|_{c^*(q)}$, in:*

$$c_*\big(\omega|_{c^*(q)}\big) \models q \;\geq\; c_*(\omega) \models q,$$

for $\omega \in \mathcal{D}(X)$, $c\colon X \rightsquigarrow Y$ and $q \in Fact(Y)$.

Proof. Via the back-and-forth transformation of Proposition 4.5.3 and the validity increase of Theorem 6.1.6:

$$c_*\big(\omega|_{c^*(q)}\big) \models q \;=\; \omega|_{c^*(q)} \models c^*(q) \;\geq\; \omega \models c^*(q) \;=\; c_*(\omega) \models q. \qquad \square$$

In the remainder of this section we illustrate the forward and backward inferences mechanisms in multiple examples. They mostly involve backward inference, since that is the more useful technique. An important first step in these examples is to recognise the channel that is hidden

in the description of the problem. It is instructive to try and do this oneself, before reading the analysis and the solution.

Example 6.2.3 We start with the following question from (Ross, 2007, Example 1.12).

Consider two urns. The first contains two white and seven black balls, and the second contains five white and six black balls. We flip a coin and then draw a ball from the first urn or the second urn depending on whether the outcome was heads or tails. What is the conditional probability that the outcome of the toss was heads given that a white ball was selected?

Our analysis involves two sample spaces $\{H, T\}$ for the sides of the coin and $\{W, B\}$ for the colours of the balls in the urns. The coin distribution is uniform: $\mathit{unif} = \frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle$. The above description implicitly contains a channel $c\colon \{H, T\} \rightsquigarrow \{W, B\}$, determined by the two urns:

$$\begin{aligned} c(H) &= \mathit{Flrn}(2|W\rangle + 7|B\rangle) \\ &= \tfrac{2}{9}|W\rangle + \tfrac{7}{9}|B\rangle \end{aligned} \qquad \begin{aligned} c(T) &= \mathit{Flrn}(5|W\rangle + 6|B\rangle) \\ &= \tfrac{5}{11}|W\rangle + \tfrac{6}{11}|B\rangle. \end{aligned}$$

As in the above quote, the first urn is associated with heads and the second one with tails.

The evidence that we have is described in the quote after the word 'given'. It is captured by the point predicate $\mathbf{1}_W$ on the set of colours $\{W, B\}$, indicating that a white ball was selected. This evidence can be pulled back (transformed) along the channel c, to a predicate $c^*(\mathbf{1}_W)$ on the sample space $\{H, T\}$. It is given by:

$$c^*(\mathbf{1}_W)(H) = \sum_{x \in \{W,B\}} c(H)(x) \cdot \mathbf{1}_W(x) = c(H)(W) = \tfrac{2}{9}.$$

Similarly we get $c^*(\mathbf{1}_W)(T) = c(T)(W) = \frac{5}{11}$.

The answer that we are interested in is obtained by updating the prior unif with the transformed evidence $c^*(\mathbf{1}_W)$, as given by $\mathit{unif}|_{c^*(\mathbf{1}_W)}$. This is an instance of backward inference.

In order to obtain the answer, we first we compute the validity:

$$\begin{aligned} \mathit{unif} \models c^*(\mathbf{1}_W) &= \mathit{unif}(H) \cdot c^*(\mathbf{1}_W)(H) + \mathit{unif}(T) \cdot c^*(\mathbf{1}_W)(T) \\ &= \tfrac{1}{2} \cdot \tfrac{2}{9} + \tfrac{1}{2} \cdot \tfrac{5}{11} = \tfrac{1}{9} + \tfrac{5}{22} = \tfrac{67}{198}. \end{aligned}$$

Then:

$$\mathit{unif}|_{c^*(\mathbf{1}_W)} = \frac{1/2 \cdot 2/9}{67/198}|H\rangle + \frac{1/2 \cdot 5/11}{67/198}|T\rangle = \tfrac{22}{67}|H\rangle + \tfrac{45}{67}|T\rangle.$$

Thus, the conditional probability of heads is $\frac{22}{67}$. The same outcome is obtained in Ross (2007), of course, but there via an application of Bayes' rule. A formalisation in ProbLog appears in Example 6.5.1 and in WebPPL in Example 6.6.1 — also with the same outcome.

Example 6.2.4 Consider the following classical question from Tversky and Kahneman (1982).

A cab was involved in a hit and run accident at night. Two cab companies, Green and Blue, operate in the city. You are given the following data:

- 85% of the cabs in the city are Green and 15% are Blue
- A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident, and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

What is the probability that the cab involved in the accident was Blue rather than Green?

We use as colour set $C = \{G, B\}$ for Green and Blue. There is a prior 'base rate' distribution $\omega = \frac{17}{20}|G\rangle + \frac{3}{20}|B\rangle \in \mathcal{D}(C)$, as in the first bullet above. The reliability information in the second bullet translates into a 'correctness' channel $c\colon \{G, B\} \rightsquigarrow \{G, B\}$ given by:

$$c(G) = \tfrac{4}{5}|G\rangle + \tfrac{1}{5}|B\rangle \qquad\qquad c(B) = \tfrac{1}{5}|G\rangle + \tfrac{4}{5}|B\rangle.$$

The second bullet also gives evidence of a Blue car. It translates into a point predicate $\mathbf{1}_B$ on $\{G, B\}$. It can be used for backward inference, giving the answer to the query, as posterior:

$$\omega|_{c^*(\mathbf{1}_B)} = \tfrac{17}{29}|G\rangle + \tfrac{12}{29}|B\rangle \approx 0.5862|G\rangle + 0.4138|B\rangle.$$

Thus the probability that the Blue car was actually involved in the incident is a bit more that 41%. This may seem like a relatively low probability, given that the evidence says 'Blue taxicab' and that observations are 80% accurate. But this low percentage is explained by the fact that there are relatively few Bleu taxicabes in the first place, namely only 15%. This is in the prior, base rate distribution ω. It is argued in Tversky and Kahneman (1982) that humans find it difficult to take such base rates (or priors) into account. This phenomenon is called base rate neglect, see also Gigerenzer and Hoffrage (1995); Stein (2018).

Example 6.2.5 We continue in the setting of Example 2.4.3, with a teacher in a certain mood — pessimistic (p), neutral (n) or optimistic (o) — making predictions about pupils' performances depending on the mood. We assume that the pupils have done rather poorly, with no-one scoring above 5, as described by the following evidence / predicate q on the set of grades $Y = \{1, 2, \ldots, 10\}$.

$$q = \tfrac{1}{10} \cdot \mathbf{1}_1 + \tfrac{3}{10} \cdot \mathbf{1}_2 + \tfrac{3}{10} \cdot \mathbf{1}_3 + \tfrac{2}{10} \cdot \mathbf{1}_4 + \tfrac{1}{10} \cdot \mathbf{1}_5.$$

Using the original mood distribution $\omega = \frac{1}{8}|p\rangle + \frac{3}{8}|n\rangle + \frac{1}{2}|o\rangle$ and channel $c\colon X \rightsquigarrow Y$ from Example 2.4.3, we can compute the validity of this predicate q in the predicted state $c_*(\omega)$ as:

$$c_*(\omega) \models q \;=\; \omega \models c^*(q) \;=\; \tfrac{299}{4000} \;=\; 0.07475.$$

One can check that the updated state $\omega' = \omega|_{c^*(q)}$ obtained via backward inference is:

$$\omega' = \tfrac{77}{299}|p\rangle + \tfrac{162}{299}|n\rangle + \tfrac{60}{299}|o\rangle \approx 0.2575|p\rangle + 0.5418|n\rangle + 0.2007|o\rangle.$$

Interestingly, after updating, the teacher has more realistic view in the sense that the validity of the predicate q has risen to $c_*(\omega') \models q = \frac{15577}{149500} \approx 0.1042$. This validity increase — see Theorem 6.2.2 — is one way how the mind can adapt to external evidence: seeing the poor results leads to more pessimism.

Example 6.2.6 Recall the Medicine-Blood Table (1.31) with data on different types of medicine via a set $M = \{0, 1, 2\}$ and blood pressure via the set $B = \{H, L\}$. From the table we can extract a channel $b\colon M \rightsquigarrow B$ describing the blood pressure distribution for each medicine type. This channel is obtained by column-wise frequentist learning:

$$b(0) = \tfrac{2}{3}|H\rangle + \tfrac{1}{3}|L\rangle \qquad b(1) = \tfrac{7}{9}|H\rangle + \tfrac{2}{9}|L\rangle \qquad b(2) = \tfrac{5}{8}|H\rangle + \tfrac{3}{8}|L\rangle.$$

The prior medicine distribution $\omega = \frac{3}{20}|0\rangle + \frac{9}{20}|1\rangle + \frac{2}{5}|2\rangle$ is obtained from the totals row in the table.

The predicted state $b_*(\omega)$ is $\frac{7}{10}|H\rangle + \frac{3}{10}|L\rangle$. It is the distribution that is learnt from the totals column in Table (1.31). Suppose we wish to focus on the people that take either medicine 1

or 2. We do so by conditioning, via the subset $E = \{1, 2\} \subseteq B$, with associated sharp predicate $\mathbf{1}_E \colon B \to [0, 1]$. Then:

$$\omega \models \mathbf{1}_E \;=\; \tfrac{9}{20} + \tfrac{2}{5} \;=\; \tfrac{17}{20} \qquad \text{so} \qquad \omega|_{\mathbf{1}_E} \;=\; \frac{9/20}{17/20}|1\rangle + \frac{2/5}{17/20}|2\rangle \;=\; \tfrac{9}{17}|1\rangle + \tfrac{8}{17}|2\rangle.$$

Forward reasoning, precisely as in Definition 6.2.1 (1), gives:

$$\begin{aligned} b_*(\omega|_{\mathbf{1}_E}) &= b_*(\tfrac{9}{17}|1\rangle + \tfrac{8}{17}|2\rangle) \\ &= (\tfrac{9}{17} \cdot \tfrac{7}{9} + \tfrac{8}{17} \cdot \tfrac{5}{8})|H\rangle + (\tfrac{9}{17} \cdot \tfrac{2}{9} + \tfrac{8}{17} \cdot \tfrac{3}{8})|H\rangle \;=\; \tfrac{12}{17}|H\rangle + \tfrac{5}{17}|L\rangle. \end{aligned}$$

This shows the distribution of high and low blood pressure among people using medicine 1 or 2.

We turn to backward reasoning. Suppose that we have evidence $\mathbf{1}_H$ on $\{H, L\}$ of high blood pressure. What is then the associated distribution of medicine usage? It is obtained in several steps.

$$\begin{aligned} b^*(\mathbf{1}_H)(x) &= \sum_{y \in B} b(x)(y) \cdot \mathbf{1}_H(y) \;=\; b(x)(H) \\ \omega \models b^*(\mathbf{1}_H) &= \sum_{x \in M} \omega(x) \cdot b^*(\mathbf{1}_H)(x) \;=\; \sum_{x \in M} \omega(x) \cdot b(x)(H) \\ &= \tfrac{3}{20} \cdot \tfrac{2}{3} + \tfrac{9}{20} \cdot \tfrac{7}{9} + \tfrac{2}{5} \cdot \tfrac{5}{8} \;=\; \tfrac{7}{10} \\ \omega|_{b^*(\mathbf{1}_H)} &= \sum_{x \in M} \frac{\omega(x) \cdot b^*(\mathbf{1}_H)(x)}{\omega \models b^*(\mathbf{1}_H)} \,|x\rangle \\ &= \frac{3/20 \cdot 2/3}{7/10}|0\rangle + \frac{9/20 \cdot 7/9}{7/10}|1\rangle + \frac{2/5 \cdot 5/8}{7/10}|2\rangle \\ &= \tfrac{1}{7}|0\rangle + \tfrac{1}{2}|1\rangle + \tfrac{5}{14}|2\rangle \;\approx\; 0.1429|0\rangle + 0.5|1\rangle + 0.3571|2\rangle. \end{aligned}$$

We can also reason with 'soft' evidence, using the full power of fuzzy predicates. Suppose we are only 95% sure that the blood pressure is high, due to some measurement uncertainty. Then we can use as evidence the predicate $q \colon B \to [0, 1]$ given by $q(H) = \frac{19}{20}$ and $q(L) = \frac{1}{20}$, that is, as $q = \frac{19}{20} \cdot \mathbf{1}_H + \frac{1}{20} \cdot \mathbf{1}_L$. It yields:

$$\omega|_{b^*(q)} \;=\; \tfrac{39}{272}|0\rangle + \tfrac{135}{272}|1\rangle + \tfrac{49}{136}|2\rangle \;\approx\; 0.1434|0\rangle + 0.4963|1\rangle + 0.3603|2\rangle.$$

This slightly differs from the outcome with sharp evidence.

ProbLog and WebPPL versions of the updating in this illustration can be found in Examples 6.5.2 and 6.6.2.

We illustrate how to solve a famous riddle via backward inference.

Example 6.2.7 The so-called Monty Hall problem is a famous riddle in probability theory, due to Selvin (1975), see also *e.g.* Snell and Vanderbei (1995); Halpern (2003); Tijms (2007):

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

We formalise the example with a set of doors $D = \{L, M, R\}$, for left, middle and right. We use elements $(x, y) \in D \times D$ to represent your choice of door x and the location of the car y. Since we

know nothing in the beginning, we use a uniform distribution as prior:

$$\upsilon = unif_{D\times D} = \tfrac{1}{9}\big|L,L\rangle + \tfrac{1}{9}\big|L,M\rangle + \tfrac{1}{9}\big|L,R\rangle + \tfrac{1}{9}\big|M,L\rangle + \tfrac{1}{9}\big|M,M\rangle + \tfrac{1}{9}\big|M,R\rangle + \tfrac{1}{9}\big|R,L\rangle + \tfrac{1}{9}\big|R,M\rangle + \tfrac{1}{9}\big|R,R\rangle.$$

The host has full knowledge of the situation, since he sees your choice of door and knows where the car is. He opens a door different from your choice where there is no car. In some cases there is no choice, but when you choose the room where the car is, the host can open two doors. This will be represented below by a uniform distribution over these two choices.

The action of the host is thus represented by a channel $h\colon D\times D \rightsquigarrow D$, given explicitly by:

$$\begin{array}{lll} h(L,L) = \tfrac{1}{2}|M\rangle + \tfrac{1}{2}|R\rangle & h(L,M) = 1|R\rangle & h(L,R) = 1|M\rangle \\ h(M,L) = 1|R\rangle & h(M,M) = \tfrac{1}{2}|L\rangle + \tfrac{1}{2}|R\rangle & h(M,R) = 1|L\rangle \\ h(R,L) = 1|M\rangle & h(R,M) = 1|L\rangle & h(R,R) = \tfrac{1}{2}|L\rangle + \tfrac{1}{2}|M\rangle. \end{array}$$

We will reason backwards. Suppose that the hosts opens the left door. This is represented by the point predicate $\mathbf{1}_L$ on D. Pulling it back along the channel h gives the predicate:

$$h^*(\mathbf{1}_L) = \tfrac{1}{2}\cdot\mathbf{1}_{(M,M)} + \mathbf{1}_{(M,R)} + \mathbf{1}_{(R,M)} + \tfrac{1}{2}\cdot\mathbf{1}_{(R,R)}.$$

Updating with this pulled back predicate yields a new, posterior distribution υ' on $D\times D$.

$$\upsilon' := \upsilon|_{h^*(\mathbf{1}_L)} = \tfrac{1}{6}\big|M,M\rangle + \tfrac{1}{3}\big|M,R\rangle + \tfrac{1}{3}\big|R,M\rangle + \tfrac{1}{6}\big|R,R\rangle.$$

The probability that you win the car is $\upsilon'(M,M) + \upsilon'(R,R) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. This covers the case that you do not change your choice. But if you do change, the chance of winning is $\upsilon'(M,R) + \upsilon'(R,M) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$. Thus, if you change, your chance of winning is twice as big, so that it is wise to change your choice after the host opens a door.

The same reasoning applies when the middle or right door is opened by the host.

Here is a simlar example, involving prisoners.

Example 6.2.8 We quote from (Casella and Berger, 2002, Ex. 1.3.4).

Three prisoners A, B, and C, are on death row. The governor decides to pardon one of the three and chooses at random the prisoner to pardon. He informs the warden of his choice but requests that the name be kept secret for a few days.

The next day, A tries to get the warden to tell him who had been pardoned. The warden refuses. A then asks which of B or C will be executed. The warden thinks for a while, then tells A that B is to be executed.

This reply makes A happy, where he reasons as follows. Given that B will be executed, then either I or C will be pardoned; my change of being pardoned has thus risen from $\frac{1}{3}$ to $\frac{1}{2}$.

This reasoning is not correct: the warden's reply keeps A's chance of being pardoned at $\frac{1}{3}$. One has to use that the warden is truthful but does not disclose, as asked, any information about A. As we shall see, the warden does disclose information about C. We formalise this below.

The starting point is the set of prisoners $P = \{A,B,C\}$, with a uniform distribution $\upsilon = unif_P = \frac{1}{3}|A\rangle + \frac{1}{3}|B\rangle + \frac{1}{3}|C\rangle$, of the pardon probabilities. The warden can tell A that B or C is to be executed, but not A, since that would give A information about himself. Thus, the warden's answer about execution is a channel $e\colon P \rightsquigarrow \{B,C\}$. It takes the following form.

$$e(A) = \tfrac{1}{2}|B\rangle + \tfrac{1}{2}|C\rangle \qquad e(B) = 1|C\rangle \qquad e(C) = 1|B\rangle.$$

In the first case, when A will be pardoned, the warden can give two answers, which are each

represented with probability $\frac{1}{2}$. When B is to be pardoned, the warden cannot tell that B will be executed — since that is not true — so he mentions C. Similarly, for the third case.

The actual reply of the warden, that B will be executed, is represented as the point predicate $\mathbf{1}_B$ on the channel's codomain $\{B, C\}$. When pulled back along the channel e, it gives the following predicate on P.

$$e^*(\mathbf{1}_B) \;=\; \tfrac{1}{2} \cdot \mathbf{1}_A + \mathbf{1}_C.$$

The updated, posterior pardon distribution is:

$$\upsilon|_{e^*(\mathbf{1}_B)} \;=\; \tfrac{1}{3}|A\rangle + \tfrac{2}{3}|C\rangle.$$

Hence the pardon probability for A does not change with the warden's reply that B will be executed. It does, however, change the pardon probability of C, from $\frac{1}{3}$ to $\frac{2}{3}$. So C is the one who should be happy with the warden's reply. See Examples 6.5.3 and 6.6.3 for ProbLog and WebPPL versions of this three-prisoner puzzle.

Example 6.2.9 In (Russell and Norvig, 2003, §20.1) a situation is described with five different bags, numbered 1, ..., 5, each containing its own mixture of cherry (C) and lime (L) candies. This situation can be described via a candy channel:

$$B \xrightarrow{c} \{C, L\} \qquad \text{where} \qquad B \;=\; \{1, 2, 3, 4, 5\} \qquad \text{and} \qquad \begin{cases} c(1) \;=\; 1|C\rangle \\ c(2) \;=\; \frac{3}{4}|C\rangle + \frac{1}{4}|L\rangle \\ c(3) \;=\; \frac{1}{2}|C\rangle + \frac{1}{2}|L\rangle \\ c(4) \;=\; \frac{1}{4}|C\rangle + \frac{3}{4}|L\rangle \\ c(5) \;=\; 1|L\rangle. \end{cases}$$

The initial bag distribution is $\omega = \frac{1}{10}|1\rangle + \frac{1}{5}|2\rangle + \frac{2}{5}|3\rangle + \frac{1}{5}|4\rangle + \frac{1}{10}|5\rangle$.

In the situation described in (Russell and Norvig, 2003, §20.1) the sample space of bags B is regarded as hidden (not directly observable), in a scenario where a new bag $i \in B$ is given and candies are drawn from it, inspected and returned. It turns out that 10 consecutive draws yield a lime candy; what can we then infer about the distribution of bags?

Transforming the lime point predicate $\mathbf{1}_L$ along channel c yields the fuzzy predicate $c^*(\mathbf{1}_L)\colon B \to [0, 1]$ given by:

$$c^*(\mathbf{1}_L) \;=\; \bigovee_i c(i)(L) \cdot \mathbf{1}_i \;=\; \tfrac{1}{4} \cdot \mathbf{1}_2 \ovee \tfrac{1}{2} \cdot \mathbf{1}_3 \ovee \tfrac{3}{4} \cdot \mathbf{1}_4 \ovee 1 \cdot \mathbf{1}_5.$$

The question is what we learn about the bag distribution after observing this predicate 10 consecutive times? This involves computing:

$$\begin{aligned}
\omega|_{c^*(\mathbf{1}_L)} &= \tfrac{1}{10}|2\rangle + \tfrac{2}{5}|3\rangle + \tfrac{3}{10}|4\rangle + \tfrac{1}{5}|5\rangle \\
\omega|_{c^*(\mathbf{1}_L)}|_{c^*(\mathbf{1}_L)} &= \omega|_{c^*(\mathbf{1}_L)\&(c^*(\mathbf{1}_L))} \;=\; \omega|_{c^*(\mathbf{1}_L)^2} \\
&= \tfrac{1}{26}|2\rangle + \tfrac{4}{13}|3\rangle + \tfrac{9}{26}|4\rangle + \tfrac{4}{13}|5\rangle \\
&\approx 0.0385|2\rangle + 0.308|3\rangle + 0.346|4\rangle + 0.308|5\rangle \\
\omega|_{c^*(\mathbf{1}_L)}|_{c^*(\mathbf{1}_L)}|_{c^*(\mathbf{1}_L)} &= \omega|_{c^*(\mathbf{1}_L)\&c^*(\mathbf{1}_L)\&c^*(\mathbf{1}_L)} \;=\; \omega|_{c^*(\mathbf{1}_L)^3} \\
&= \tfrac{1}{76}|2\rangle + \tfrac{4}{19}|3\rangle + \tfrac{27}{76}|4\rangle + \tfrac{8}{19}|5\rangle \\
&\approx 0.0132|2\rangle + 0.211|3\rangle + 0.355|4\rangle + 0.421|5\rangle \quad \ldots
\end{aligned}$$

Figure 20.1 in Russell and Norvig (2003) gives a plot of these distributions; it is reconstructed

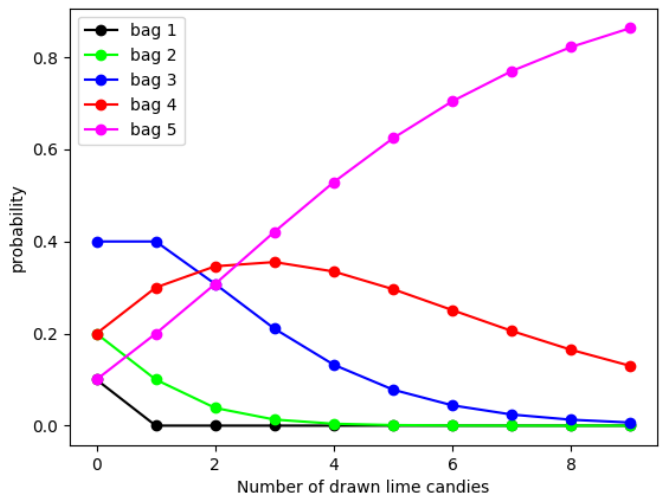


Figure 6.1 Posterior, updated bag distributions $\omega|_{(c^*(\mathbf{1}_L))^n}$ for $n = 0, 1, \ldots, 10$, aligned vertically, after n candy draws that all happen to be lime.

here in Figure 6.1 via the above formulas. It shows that bag 5 quickly becomes most likely — as expected since it contains most lime candies — and that bag 1 is impossible after drawing the first lime. A ProbLog version of this iterated updating appears in Example 6.5.4.

Example 6.2.10 Medical tests form standard examples of Bayesian reasoning, via backward inference, see *e.g.* Exercises 6.2.1 and 6.2.2 below. Here we look at Covid-19 which is interesting because its standard PCR-test has low false positives but high false negatives, and moreover these false negatives depend on the day after infection.

The Covid-19 PCR-test, that was used during the pandemic around 2020, has almost no false positives. This means that if you do not have the disease, then the likelihood of a (false) postive test is very low. This means that the specificity of the test is very high, see Exercises 6.2.1 and 6.2.2 below. In our calculations below we put it at 1%, independently of the day that you get tested. In contrast, the PCR-test has considerable false negative rates, which depend on the day after infection. The plot at the top in Figure 6.2 gives an indication; it does not precisely reflect the medical reality, but it provides a reasonable approximation, good enough for our calculation. This plot shows that if you are infected at day 0, then a test at this day or the day after (day 1) will surely be negative. On the second day after your infection a PCR-test might start to detect, but still there is only a 20% chance of a positive outcome. This probability increases and after on day 6 the likelihood of a positive test has risen to 80%.

How to formalise this situation? We use the following three sample spaces, for Covid (C), days after infection (D), and test outcome (T).

$$C = \{c, c^\perp\} \qquad D = \{0, 1, 2, 3, 4, 5, 6\} \qquad T = \{p, n\}.$$

The test probabilities are then captured via a test channel $t\colon C \times D \rightsquigarrow T$, given in the following

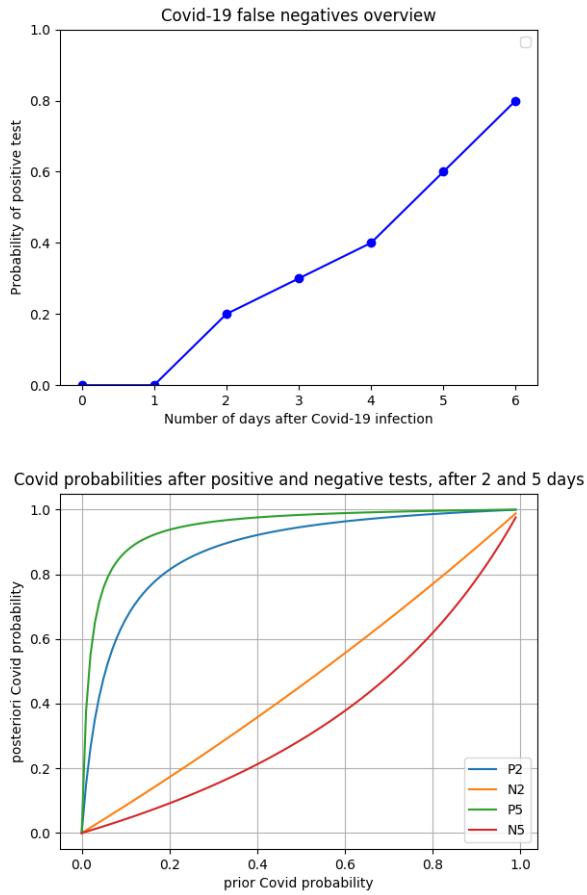


Figure 6.2 Covid-19 false negatives and posteriors after tests. In the lower plot P2 means: positive test after 2 days, that is, in state $(r|c\rangle + (1-r)|c^{\perp}\rangle) \otimes \varphi_2$, where $r \in [0,1]$ is the prior Covid probability. Similarly for N2, P5, N5.

way. The first equation captures the false positives (specificity), and the second one the false negatives (sensitivity), as in the plot at the top of Figure 6.2.

$$t(c^{\perp}, i) = \tfrac{1}{100}|p\rangle + \tfrac{99}{100}|n\rangle \qquad\qquad t(c, i) = \begin{cases} 1|n\rangle & \text{if } i = 0 \text{ or } i = 1 \\ \frac{2}{10}|p\rangle + \frac{8}{10}|n\rangle & \text{if } i = 2 \\ \frac{3}{10}|p\rangle + \frac{7}{10}|n\rangle & \text{if } i = 3 \\ \frac{4}{10}|p\rangle + \frac{6}{10}|n\rangle & \text{if } i = 4 \\ \frac{6}{10}|p\rangle + \frac{4}{10}|n\rangle & \text{if } i = 5 \\ \frac{8}{10}|p\rangle + \frac{2}{10}|n\rangle & \text{if } i = 6. \end{cases}$$

In practice it is often difficult to determine the precise date of infection. We shall consider two

cases, where the infection happened (around) two and five days ago, via the two distributions:

$$\sigma_2 \;=\; \tfrac{1}{4}|1\rangle + \tfrac{1}{2}|2\rangle + \tfrac{1}{4}|3\rangle \qquad\qquad \sigma_5 \;=\; \tfrac{1}{4}|4\rangle + \tfrac{1}{2}|5\rangle + \tfrac{1}{4}|6\rangle.$$

Let's consider the case where we have no (prior) information about the likelihood that the person that is going to be tested has the disease. Therefore we use as prior $\omega = \mathit{unif}_C = \frac{1}{2}|c\rangle + \frac{1}{2}|c^\perp\rangle$.

Suppose in this situation we have a positive test, say after two days. What do we then learn about the disease probability? Our evidence is the postive test predicate $\mathbf{1}_p$ on T, which can be transformed to $t^*(\mathbf{1}_p)$ on $C \times D$. We can use this predicate to update the joint state $\omega \otimes \sigma_2$. The distribution that we are after is the first marginal of this updated state, as in:

$$\Big((\omega \otimes \sigma_2)\big|_{t^*(\mathbf{1}_p)}\Big)[1,0] \;\in\; \mathcal{D}(C).$$

We shall go through the computation step-by-step. First,

$$\begin{aligned} t^*(\mathbf{1}_p) \;&=\; \sum_{x\in C,\, y\in D} t(x,y)(p)\cdot \mathbf{1}_{(x,y)} \\ &=\; \tfrac{2}{10}\cdot\mathbf{1}_{(c,2)} + \tfrac{3}{10}\cdot\mathbf{1}_{(c,3)} + \tfrac{4}{10}\cdot\mathbf{1}_{(c,4)} + \tfrac{6}{10}\cdot\mathbf{1}_{(c,5)} + \tfrac{8}{10}\cdot\mathbf{1}_{(c,6)} + \tfrac{1}{100}\cdot\mathbf{1}_{(c^\perp,0)} + \tfrac{1}{100}\cdot\mathbf{1}_{(c^\perp,1)} \\ &\qquad + \tfrac{1}{100}\cdot\mathbf{1}_{(c^\perp,2)} + \tfrac{1}{100}\cdot\mathbf{1}_{(c^\perp,3)} + \tfrac{1}{100}\cdot\mathbf{1}_{(c^\perp,4)} + \tfrac{1}{100}\cdot\mathbf{1}_{(c^\perp,5)} + \tfrac{1}{100}\cdot\mathbf{1}_{(c^\perp,6)}. \end{aligned}$$

Then:

$$\begin{aligned} \omega\otimes\sigma_2 \models t^*(\mathbf{1}_p) \;&=\; \sum_{x\in C,\, y\in D} \omega(x)\cdot\sigma_2(y)\cdot t^*(\mathbf{1}_p)(x,y) \\ &=\; \tfrac{1}{2}\cdot\tfrac{1}{2}\cdot\tfrac{2}{10} + \tfrac{1}{2}\cdot\tfrac{1}{4}\cdot\tfrac{3}{10} + \tfrac{1}{2}\cdot\tfrac{1}{4}\cdot\tfrac{1}{100} + \tfrac{1}{2}\cdot\tfrac{1}{2}\cdot\tfrac{1}{100} + \tfrac{1}{2}\cdot\tfrac{1}{4}\cdot\tfrac{1}{100} \\ &=\; \tfrac{40+30+1+2+1}{800} \;=\; \tfrac{37}{400}. \end{aligned}$$

But then:

$$\begin{aligned} (\omega\otimes\sigma_2)\big|_{t^*(\mathbf{1}_p)} \;&=\; \frac{1/2\cdot 1/2\cdot 2/10}{37/400}\big|c,2\rangle + \frac{1/2\cdot 1/4\cdot 3/10}{37/400}\big|c,3\rangle + \frac{1/2\cdot 1/4\cdot 1/100}{37/400}\big|c^\perp,1\rangle \\ &\qquad + \frac{1/2\cdot 1/2\cdot 1/100}{37/400}\big|c^\perp,2\rangle + \frac{1/2\cdot 1/4\cdot 1/100}{37/400}\big|c^\perp,3\rangle \\ &=\; \tfrac{20}{37}\big|c,2\rangle + \tfrac{15}{37}\big|c,3\rangle + \tfrac{1}{74}\big|c^\perp,1\rangle + \tfrac{1}{37}\big|c^\perp,2\rangle + \tfrac{1}{74}\big|c^\perp,3\rangle. \end{aligned}$$

Finally:

$$\Big((\omega\otimes\sigma_2)\big|_{t^*(\mathbf{1}_p)}\Big)[1,0] \;=\; \tfrac{35}{37}|c\rangle + \tfrac{2}{37}|c^\perp\rangle \;\approx\; 0.946|c\rangle + 0.054|c^\perp\rangle.$$

Hence a positive test changes the a priori likelihood of 50% to about 95%. In a similar way one can compute the effect of a negative test:

$$\Big((\omega\otimes\sigma_2)\big|_{t^*(\mathbf{1}_n)}\Big)[1,0] \;=\; \tfrac{165}{363}|c\rangle + \tfrac{198}{363}|c^\perp\rangle \;\approx\; 0.455|c\rangle + 0.545|c^\perp\rangle.$$

We see that a negative test, 2 days after infection, reduces the prior disease probability of 50% only slightly, namely to 45%. Thus, a negative PCR test is not very informative, but a positive test is.

Doing the test around 5 days after infection gives more certainty, especially in the case of a negative test:

$$\begin{aligned} \Big((\omega\otimes\sigma_5)\big|_{t^*(\mathbf{1}_p)}\Big)[1,0] \;&=\; \tfrac{60}{61}|c\rangle + \tfrac{1}{61}|c^\perp\rangle \;\approx\; 0.984|c\rangle + 0.016|c^\perp\rangle \\ \Big((\omega\otimes\sigma_5)\big|_{t^*(\mathbf{1}_n)}\Big)[1,0] \;&=\; \tfrac{40}{139}|c\rangle + \tfrac{99}{139}|c^\perp\rangle \;\approx\; 0.288|c\rangle + 0.712|c^\perp\rangle. \end{aligned}$$

The lower plot in Figure 6.2 gives a more elaborate description, for different prior disease probabilities (prevalences) $r \in [0, 1]$ in a distribution $\omega = r|c\rangle + (1 - r)|c^{\perp}\rangle$ as used above. We see that a postive test outcome quickly gives certainty about having the disease. But a negative test outcome gives only a little bit of information with respect to the prior — which in this plot can be represented as the diagonal. For this reason, if you get a negative PCR-test, often a second test is done a few days later.

In the next two examples we look at estimating the number of fish in a pond by counting marked fishes, first in multinomial (actually binomial) mode and then in hypergeometric mode.

Example 6.2.11 Capture and recapture is a methodology used in ecology to estimate the size of a population. So imagine we are looking at a pond and we wish to learn the number of fish that it contains. We catch twenty of them, mark them, and throw them back. Subsequently we catch another twenty, and find out that five of them are marked. What do we learn about the number of fish? This is generally called a Lincoln-Peterson mark-and-recapture problem.

The number of fish in the pond must be at least 20. Let's assume the maximal number is 300. We will be considering units of 10 fish. Hence the underlying sample space F together with the uniform 'prior' state $unif_F$ is:

$$F = \{20, 30, 40, \ldots, 300\} \qquad \text{with} \qquad unif_F = \sum_{x \in F} \tfrac{1}{29}|x\rangle.$$

We now assume that $K = 20$ of the fish in the pond are marked. We can then compute for each value $20, 30, 40, \ldots$ in the fish space F the probability of finding 5 marked fish when 20 of them are caught. In order not to complicate the calculations too much, we catch these 20 fish one by one, check if they are marked, and then throw them back. This means that the probability of catching a marked fish remains the same, and is described by a binomial distribution, see Example 2.1.2 (2). Its parameters are $K = 20$ with probability $\frac{20}{i}$ of catching a marked fish, where $i \in F$ is the assumed total number of fish. This is incorporated in the following 'catch' channel $c\colon F \rightarrowtail \{0, 1, \ldots, K\}$.

$$c(i) := bn[K]\left(\tfrac{K}{i}\right) \overset{(2.1)}{=} \sum_{0 \le k \le K} \binom{K}{k} \cdot \left(\tfrac{K}{i}\right)^k \cdot \left(\tfrac{i-K}{i}\right)^{K-k} |k\rangle.$$

Once this is set up, we construct a posterior state by updating the prior with the information that five marked fish have been found. The latter is expressed as point predicate $\mathbf{1}_5 \in Pred(\{0, \ldots, K\})$ on the codomain of the channel c. We can now do backward inference, as in Definition 6.2.1 (2), and obtain the updated uniform distribution:

$$unif_F\big|_{c^*(\mathbf{1}_5)} = \sum_{i \in F} \frac{(20/i)^5 \cdot (i-20/i)^{15}}{\sum_j (20/j)^5 \cdot (j-20/j)^{15}} |i\rangle.$$

The bar chart of this posterior is in Figure 6.3; it indicates the likelihoods of the various numbers of fish in the pond. One can also compute the expected value (mean) of this posterior; it's 116.5 fish. In case we had caught 10 marked fish out of 20, the expected number would be 47.5.

Note that taking a uniform prior corresponds to the idea that we have no idea about the number of fish in the pond. But possibly we already had a good estimate from previous years. Then we could have used such an estimate as prior distribution, and updated it with this year's evidence.

Later on, in Example **??**, we shall see a 'continuous' version of this example, where the space

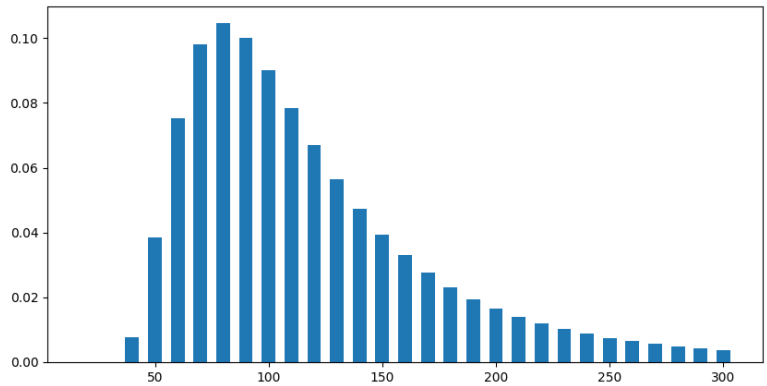


Figure 6.3 The posterior fish number distribution after catching 5 marked fish, in multinomial (throw back) mode, see Example 6.2.11.

of fish numbers is not a finite set $\{20, 30, \ldots, 300\}$ but an interval $[20, 300] \subseteq \mathbb{R}$. The bar chart in Figure 6.3 then becomes a smooth line, see Figure **??**.

Example 6.2.12 We take another look at the previous example. There we used the multinomial distribution (in binomial form) for the probability of catching five marked fish, per pond size. This multinomial mode is appropriate for drawing with replacement, which corresponds to returning each fish that we catch to the pond. This is probably not what happens in practice. So let's try to redescribe the capture-recapture model in hypergeometric mode (like in (Ross, 2018, §4.8.3, Ex. 8h)).

Let's write $M = \{m, m^\perp\}$ for the space with elements m for marked and $m^\perp$ for unmarked. Our recapture catch (draw) of $K = 20$ fish, with 5 of them marked, is thus a multiset $\kappa = 5|m\rangle + 15|m^\perp\rangle$. The urn from which we draw is the pond, in which the total number of fish is unknown, but we do know that 20 of them are marked. The urn / pond is thus a multiset $20|m\rangle + (i-20)|m^\perp\rangle$ with $i \in F$.

We now use a channel:

$$\{20, 30, \ldots, 300\} = F \xrightarrow{d} \mathcal{N}[K](M) \cong \{0, 1, \ldots, K\},$$

given by:

$$d(i) := hg[K]\Big(20|m\rangle + (i-20)|m^\perp\rangle\Big).$$

The updated pond distribution is then:

$$unif_F\big|_{d^*(\mathbf{1}_\kappa)} = \sum_{i \in F} \frac{\binom{i-20}{15}/\binom{i}{20}}{\sum_j \binom{j-20}{15}/\binom{j}{20}} \,|i\rangle.$$

Its bar-chart is in Figure 6.4. It differs minimally from the multinomial one in Figure 6.3. In the hypergeometric case the expected value is 113 fish, against 116.5 in the multinomial case. When the recapture involves 10 marked fish, the expected values are 45.9, against 47.5. As we have

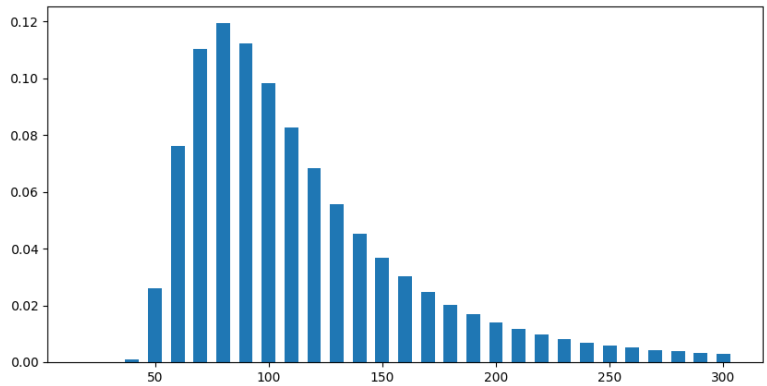


Figure 6.4 The posterior fish number distribution after catching 5 marked fish, in hypergeometric mode, see Example 6.2.12.

already seen in Proposition 3.6.5, the hypergeometric distribution on small draws from a large urn look very much like a multinomial distribution.

We include another example of a similar kind. It does not involve fish but tanks.

Example 6.2.13 During the second world war British statisticians developed a method to estimate the German tank production, of a particular model, from the serial numbers of destroyed tanks found at the battlefield. This has become a challenge that attracted quite a bit of interest. We refer to a Wikipedia page[1] on this topic for background information and for various approaches. The page contains an example with specific numbers, which we copy. Here we develop a channel-based solution, using updating in several ways.

Suppose tanks with serial numbers 19, 40, 42 and 60 have been found. It is assumed that the tanks are produced with increasing numbers, starting from zero, and that they are deployed by the (German) army as soon as they are available. The question is: what is the likely total number of produced tanks?

The first thing that we can say is that at least 60 tanks must have been produced. Thus we use as lower end-point:

$$MiP = 60 \qquad \text{where } MiP \text{ stands for: minimal production.}$$

We don't know the maximum, but since the known serial numbers are relatively low, we expect that the maximal production MaP is not terribly high. We choose:

$$MaP = 250.$$

Our aim is thus to infer a production distribution on the production interval $P = [MiP, MaP] = [60, 250] \subseteq \mathbb{N}$. We start from a uniform prior $unif_P \in \mathcal{D}(P)$.

The observation space O, containing the serial numbers of tanks that can be found is then the

[1] https://en.wikipedia.org/wiki/German_tank_problem

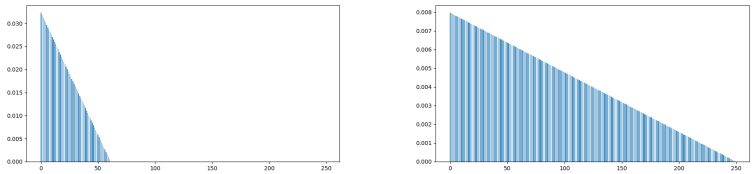


Figure 6.5 The plots of observation distributions for minimum production on the left, and for maximal production on the right, see Example 6.2.13.

interval $O = [0, MaP]$. We thus assume that 0 is the serial number of the tank that was produced first.

Let $n \in P$ be the total number produced tanks. How likely is it to find a serial number $i \leq n$ at the battlefield? It makes sense that this likelihood is higher for lower i. Indeed, a tank with a low number is produced a longer time ago, and is thus in combat for a longer time, and is thus more likely to have been destroyed — so that its serial number can be registered. We make the simplifying assumption that this likelihood descreases linearly when the number i rises.

In order to capture this assumption we use a predicate q_n on the observation space $O = [0, MaP]$, for each production number $n \in P = [MiP, MaP]$. We define, for $i \in O$,

$$q_n(i) \coloneqq \begin{cases} \frac{n-i}{n} & \text{when } i < n \\ 0 & \text{when } i \geq n. \end{cases}$$

Of course, when n tanks have been produced — starting with number 0 — it is impossible to find a serial number $i \geq n$. This explains the last clause. The first clause expresses the linear decline as i rises.

We can now define a tank channel $t \colon P \multimap O$, from production to observation. We use the above predicates q_n to update the uniform distribution on O, as in:

$$t(n) \coloneqq unif_O\big|_{q_n}.$$

Figure 6.5 describes the observation distributions $t(MiP)$ for minimal production on the left. The distribution $t(MaP)$ for maximal production is on the right. Non-zero probabilities exist only for $i < MiP = 60$ on the left, in decreasing order. When production is maximal, all numbers $i < MaP = 250$ occur in the support of the distribution $t(MaP)$.

We now define a posterior distribution $\omega \in \mathcal{D}(P)$ via backward reasoning, using the discovered serial numbers 19, 40, 42 and 60 as point predicates on P. Thus we take:

$$\omega \coloneqq unif_P\big|_{t^*(\mathbf{1}_{19})}\big|_{t^*(\mathbf{1}_{40})}\big|_{t^*(\mathbf{1}_{42})}\big|_{t^*(\mathbf{1}_{60})} = unif_P\big|_{t^*(\mathbf{1}_{19})\,\&\,t^*(\mathbf{1}_{40})\,\&\,t^*(\mathbf{1}_{42})\,\&\,t^*(\mathbf{1}_{60})}.$$

Figure 6.6 contains the resulting posterior tank production distribution. It reaches its maximum at $n = 90$, for the number of produced tanks. This corresponds closely to the number $n = 89$ derived at Wikipedia for what is there called the Bayesian mean. However, the mean of the distribution in Figure 6.6 is 126.5.

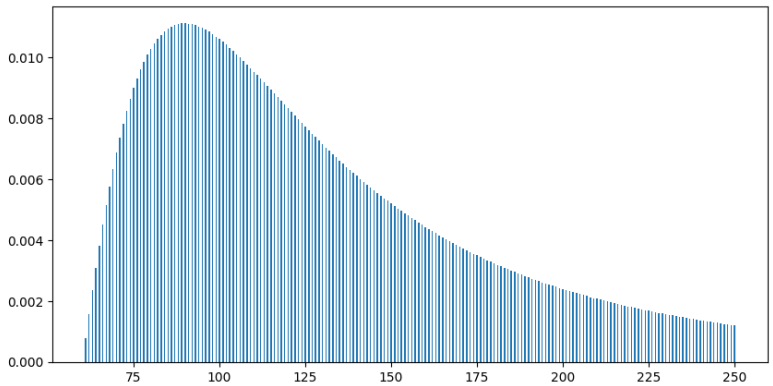


Figure 6.6 The plot of the posterior tank production distribution from Example 6.2.13.

Exercises

6.2.1 We consider some disease with an *a priori* probability (or 'prevalence') of 1%. There is a test for the disease with the following characteristics.

- ('sensitivity') If someone has the disease, then the test is positive with probability of 90%.
- ('specificity') If someone does not have the disease, there is a 95% chance that the test is negative.

1 Take as disease space $D = \{d, d^\perp\}$; describe the prior as a distribution on D;

2 Take as test space $T = \{p, n\}$ and describe the combined sensitivity and specificity as a channel $c\colon D \rightsquigarrow T$;

3 Show that the predicted positive test probability is almost 6%.

4 Assume that a test comes out positive. Use backward reasoning to prove that the probability of having the disease (the posterior, or 'Positive Predictive Value', PPV) is then a bit more than 15% (to be precise: $\frac{18}{117}$). Explain why it is so low — remembering Example 6.2.4.

6.2.2 In the context of the previous exercise we can derive the familiar formulas for Positive Predictive Value (PPV) and Negative Predictive Value (NPV) of medical tests. Let's assume we have a disease prevalence (prior) given by $\omega = p|d\rangle + (1-p)|d^\perp\rangle$ with parameter $p \in [0,1]$ and a channel channel $c\colon \{d, d^\perp\} \rightsquigarrow \{p, n\}$ with sensitivity and specificity parameters $se, sp \in [0,1]$ in:

$$\begin{array}{lrcl} \text{sensitivity:} & c(d) & = & se|p\rangle + (1-se)|n\rangle \\ \text{specificity:} & c(d^\perp) & = & (1-sp)|p\rangle + sp|n\rangle. \end{array}$$

Check that:

$$PPV := \omega|_{c^*(\mathbf{1}_p)}(d) = \frac{p \cdot se}{p \cdot se + (1-p)\cdot(1-sp)}.$$

This is commonly expressed in medical textbooks as:

$$PPV \;=\; \frac{prevalence \cdot sensitivity}{prevalence \cdot sensitivity + (1 - prevalence) \cdot (1 - specificity)}.$$

Check similarly that:

$$NPV \;:=\; \omega|_{c^*(\mathbf{1}_n)}(d^\perp) \;=\; \frac{(1-p)\cdot sp}{p \cdot (1 - se) + (1-p)\cdot sp}.$$

As an aside, the (positive) *likelihood ratio* LR is the fraction:

$$LR \;:=\; \frac{c(d)(p)}{c(d^\perp)(p)} \;=\; \frac{se}{1 - sp}.$$

6.2.3 Give a channel-based analysis and answer to the following question from (Ross, 2007, Chap. I, Exc. 39).

Stores A, B, and C have 50, 75, and 100 employees, and respectively, 50, 60, and 70 percent of these are women. Resignations are equally likely among all employees, regardless of sex. One employee resigns and this is a woman. What is the probability that she works in store C?

6.2.4 The multinomial and hypergeometric charts in Figures 6.3 and 6.4 are very similar, but there are differences, notably when there are few fish in the pond. For instance, when there are only 40 fish in the pond (with 20 of them marked) the charts really differ. Give a conceptual explation for this difference.

6.2.5 The following situation about the relationship between eating hamburgers and having Kreuzfeld-Jacob disease is insprired by (Barber, 2012, §1.2). We have two sets: $E = \{H, H^\perp\}$ about eating Hamburgers (or not), and $D = \{K, K^\perp\}$ about having Kreuzfeld-Jacob disease (or not). The following distributions on these sets are given: half of the people eat hamburgers, and only one in hundred thousand have Kreuzfeld-Jacob disease, which we write as:

$$\omega \;=\; \tfrac{1}{2}|H\rangle + \tfrac{1}{2}|H^\perp\rangle \qquad \text{and} \qquad \sigma \;=\; \tfrac{1}{100{,}000}|K\rangle + \tfrac{99{,}999}{100{,}000}|K^\perp\rangle.$$

1 Suppose that we know that 90% of the people who have Kreuzfeld-Jacob disease eat Hamburgers. Use this additional information to define a channel $c\colon D \twoheadrightarrow E$ with $c_*(\sigma) = \omega$.

2 Compute the probability of getting Kreuzfeld-Jacob for someone eating hamburgers (via backward inference).

6.2.6 Consider in the context of Example 6.2.10 a negative Covid-19 test obtained after 2 days, via the distribution $\sigma_2 = \frac{1}{4}|1\rangle + \frac{1}{2}|2\rangle + \frac{1}{4}|3\rangle$, assuming a uniform disease prior ω. Show that the posterior 'days' distribution is:

$$\Big((\omega \otimes \sigma_2)\big|_{t^*(\mathbf{1}_n)}\Big)[0,1] \;=\; \tfrac{199}{726}|1\rangle + \tfrac{179}{363}|2\rangle + \tfrac{169}{726}|3\rangle \;\approx\; 0.274|1\rangle + 0.493|2\rangle + 0.233|3\rangle.$$

Explain why there a (small) shift 'forward', making the earlier days more likely in this posterior — with respect to the prior σ_2.

6.2.7 Consider the following challenge, copied from Staton (2021).

(i) I have forgotten what day it is.
(ii) There are ten buses per hour in the week and three buses per hour at the weekend.
(iii) I observe four buses in a given hour.
(iv) What is the probability that it is the weekend?

Let $W = \{wd, we\}$ be a set with elements wd for weekday and we for weekend, with prior distribution $\frac{5}{7}|wd\rangle + \frac{2}{7}|we\rangle$. Use the Poisson distribution to define a channel $bus\colon W \to \mathcal{D}_\infty(\mathbb{N})$ and use it to answer the above question via backward inference.

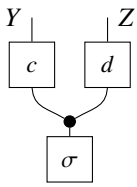
6.3 Analysis of forward and backward inference

The previous section contains many illustrations of probabilistic inference via updating. The current section is more mathematical in nature and looks at the properties of channel-based inference. One recurring topic is the crossover influence through updating of joint states. The abstract results of this section will be illustrated in the next section on Bayesian networks.

Our first result illustrates how forward and backward inference show up naturally in reasoning in a graphical setting: when a joint distribution has a specific form, as indicated below (6.6), then the marginal in one component after an update in the other component can be described via reasoning along the channels involved.

Theorem 6.3.1 *Consider the following situation, with a distribution $\sigma \in \mathcal{D}(X)$ and two channels $c\colon X \to\!\!\!\!\bullet\; Y$ and $d\colon X \to\!\!\!\!\bullet\; Z$. Define a joint distribution $\omega \in \mathcal{D}(Y \times Z)$ as:*

$$\omega \;:=\; \langle c, d\rangle_*(\sigma) \;=\; \tag{6.6}$$



Then, for a factor $q \in Fact(Y)$,

$$(\omega|_{q\otimes\mathbf{1}})[1,0] \;=\; c_*(\omega)|_q \qquad \textit{and} \qquad (\omega|_{q\otimes\mathbf{1}})[0,1] \;=\; d_*(\sigma|_{c^*(q)}). \tag{6.7}$$

The second equation in (6.7) is most interesting. Its right-hand-side involves a forward inference of a backward inference. What happens can be described at an intuitive level using the string diagram in (6.6): the evidence q is first pull backward (down) along channel c, the resulting factor $c^*(q)$ is used to update the distribution σ, and then the result is pushed forward (up) along channel d. This same dynamics occurs in more complicated form in Exercise 6.3.5.

The left-hand-side of the equation (6.7) can also be written in terms of projections $\pi_1\colon Y\times Z \to Y$ and $\pi_2\colon Y \times Z \to Z$, namely as: $(\pi_2)_*(\omega|_{(\pi_1)^*(q)})$. The formulation with the projections will be generalised to an inference query (6.19).

Proof. We first note that:

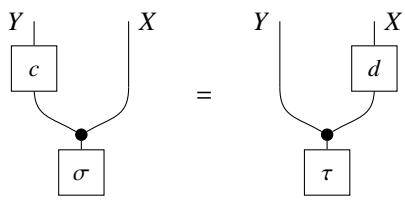
$$\begin{aligned}
\omega \models q \otimes \mathbf{1} \;=\; \langle c, d\rangle_*(\sigma) \models q \otimes \mathbf{1} \;&=\; \sigma \models \langle c, d\rangle^*(q \otimes \mathbf{1}) && \text{by Proposition 4.5.3}\\
&=\; \sigma \models c^*(q) \;\&\; d^*(\mathbf{1}) && \text{by Lemma 4.5.2 (7)}\\
&=\; \sigma \models c^*(q) \;\&\; \mathbf{1}\\
&=\; \sigma \models c^*(q).
\end{aligned}$$

Then, for an element $z \in Z$,

$$
\begin{aligned}
(\omega|_{q\otimes\mathbf{1}})[0,1](z) \;=\; \sum_{y\in Y} \frac{\omega(y,z)\cdot(q\otimes\mathbf{1})(y,z)}{\omega \models q\otimes\mathbf{1}} &= \sum_{y\in Y}\sum_{x\in X} \frac{\sigma(x)\cdot c(x)(y)\cdot d(x)(z)\cdot q(y)}{\sigma \models c^*(q)} \\
&= \sum_{x\in X} \frac{\sigma(x)\cdot \sum_{y\in Y} c(x)(y)\cdot q(y)}{\sigma \models c^*(q)} \cdot d(x)(z) \\
&= \sum_{x\in X} \frac{\sigma(x)\cdot c^*(q)(x)}{\sigma \models c^*(q)} \cdot d(x)(z) \\
&= \sum_{x\in X} \sigma|_{c^*(q)}(x)\cdot d(x)(z) \;=\; d_*(\sigma|_{c^*(q)})(z) \qquad \square
\end{aligned}
$$

At several earlier places in this book we have encountered string-diagrammatic equations of the form (6.8) described below. Such equations between two graphs are of interest for reversal of channels, see Chapter 7.

Theorem 6.3.2 *Let $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$ be distributions with channels $c\colon X \rightsquigarrow Y$ and $d\colon Y \rightsquigarrow X$ for which the following equation between graphs holds.*



(6.8)

As a consequence, via marginalisation, $\sigma = d_*(\tau)$ and $\tau = c_*(\sigma)$.

For a factor $q \in \mathsf{Fact}(Y)$, backward inference along c and forward inference along d coincide:

$$\sigma|_{c^*(q)} \;=\; d_*(\tau|_q). \tag{6.9}$$

In particular, when q is a point predicate $\mathbf{1}_y$ for $y \in Y$ we get:

$$\sigma|_{c^*(\mathbf{1}_y)} \;=\; d(y). \tag{6.10}$$

Thus, the channel d is completely determined by σ and c, under the condition that the validity $\sigma \models c^*(\mathbf{1}_y) = c_*(\sigma) \models \mathbf{1}_y = c_*(\sigma)(y)$ is non-zero, for each $y \in Y$, so that the update $\sigma|_{c^*(\mathbf{1}_y)}$ is well-defined. Equivalently, $c_*(\sigma)$ must have full support.

In the next chapter we shall see that the formula $\sigma|_{c^*(\mathbf{1}_y)}$ defines a reversed 'dagger' channel $c^\dagger_\sigma\colon Y \rightsquigarrow X$ applied to y.

Proof. We show that an arbitrary predicate p gets the same validity in both distributions, see

Remark 4.4.10. We use Proposition 4.5.3 several times.

$$\begin{aligned}
\sigma|_{c^*(q)} \models p &= \frac{\sigma \models c^*(q) \mathbin{\&} p}{\sigma \models c^*(q)} && \text{by Theorem 6.1.4 (1)} \\
&= \frac{\sigma \models \langle c, id\rangle^*(q \otimes p)}{c_*(\sigma) \models q} && \text{by Lemma 4.5.2 (7)} \\
&= \frac{\langle c, id\rangle_*(\sigma) \models q \otimes p}{(\pi_1)_*(\langle c, id\rangle_*(\sigma)) \models q} \\
&= \frac{\langle id, d\rangle_*(\tau) \models q \otimes p}{(\pi_1)_*(\langle id, d\rangle_*(\tau)) \models q} && \text{by assumption} \\
&= \frac{\tau \models q \mathbin{\&} d^*(p)}{\tau \models q} && \text{again by Lemma 4.5.2 (7)} \\
&= \tau|_q \models d^*(p) && \text{by Theorem 6.1.4 (1)} \\
&= d_*(\tau|_q) \models p.
\end{aligned}$$

In the special case when q is a point predicate $\mathbf{1}_y$ we get:

$$\begin{aligned}
\sigma|_{c^*(\mathbf{1}_y)} \overset{(6.9)}{=} d_*(\tau|_{\mathbf{1}_y}) &= d_*(1|y\rangle) && \text{by Lemma 6.1.7 (2)} \\
&= d(y).
\end{aligned}$$

❑

This result can be applied in several situations that we have seen before, involving an equation between graphs, of the form (6.8).

Examples 6.3.3 1 The archetypal equation of the form (6.8) is the one in Theorem 3.3.1, relating accumulation and arrangement (3.15). We can use it in two ways and obtain descriptions of accumulation and arrangement as updates along each other: for $\varphi \in \mathcal{N}[K](X)$ and $\vec{x} \in X^K$,

$$iid[K](\omega)\big|_{acc^*(\mathbf{1}_\varphi)} = arr(\varphi) \qquad\qquad mn[K](\omega)\big|_{arr^*(\mathbf{1}_{\vec{x}})} = \langle acc\rangle(\vec{x}).$$

In these examples the prior distributions $iid[K](\omega)$ and $mn[K](\omega)$ do not play a role in the outcome. This phenomenon will be described in terms of 'sufficient statistic' in Section 7.7.

2 In Exercise 3.2.12 we have seen that the draw-store-delete channel DSD can be obtained via updating, namely as: $\omega \otimes mn[K](\omega)\big|_{mcons_\psi} = DSD(\psi)$, for an arbitrary distribution ω. It is an instance of Equation (6.10), since:

- the predicate $mcons_\psi$ used in Exercise 3.2.12 can also be written as a transformation of a point predicate, namely as $mcons^*(\mathbf{1}_\psi)$;
- there is a graph equations diagram (3.14), that fits the pattern (6.8) in Theorem 6.3.2.

3 In Theorem 3.5.4 we have seen the graph equations diagram (3.27). By applying Equation (6.10) we see how hypergeometric(-store) distributions can be obtained from multinomial ones via updating. For an urn $\upsilon \in \mathcal{N}[L+K](X)$,

$$hgs[K](\upsilon) = mn[K](\omega) \otimes mn[L](\omega)\big|_{sum^*(\mathbf{1}_\upsilon)}.$$

This holds for any distribution ω. By taking the first marginals on both sides we get the hypergeometric distribution as marginal of an update:

$$hg[K](\upsilon) = hgs[K](\upsilon)[1,0] = \Big(mn[K](\omega) \otimes mn[L](\omega)\big|_{sum^*(\mathbf{1}_\upsilon)}\Big)[1,0].$$

A binary / univariate version of this result, for binomial distributions, may be found in (Blitzstein

and Hwang, 2019, Thm. 3.9.2) and in (Johnson et al., 2005, Chap. 3.4). Later, in Example 7.10.7 (1) we shall redescribe this characterisation of the hypergeometric distribution in diagrammatic form and see how the hypergeometric channel $hg[K]$ appears as (marginal of) a dagger channel.

4 In Theorem 3.3.9 we have seen Diagram (3.23) from which we can extra an equation, for an arbitrary distribution $\omega \in \mathcal{D}(X)$ with full support and for each $\lambda \in \mathbb{R}_{>0}$,

$$mn[K](\omega) \;=\; \left(\bigotimes_{x\in X} pois[\lambda \cdot \omega(x)]\right)\Big|_{sum^*(\mathbf{1}_K)} \;=\; Pmn[\lambda](\omega)\Big|_{sum^*(\mathbf{1}_K)}.$$

The second equation uses the characterisation of the Poisson multinomial distribution from Lemma 3.10.3. Formally, we have to add a frequencies isomorphism $Freq\colon \mathbb{N}^X \xrightarrow{\cong} \mathcal{N}(X)$ on the right hand side, so that the types precisely match. This map $Freq$ turns a tuple $t\colon X \to \mathbb{N}$ into a multiset $\sum_x t(x)|x\rangle$.

This way to describe the multinomial distribution as an update of parallel Poisson distributions is well-known, especially in bivariate form, see *e.g.* (Ross, 2018, §6.4). Later on in Example 7.10.7 (2) we shall reformulate the above update equation in graphical form.

Recall from Definition 6.1.1 that we use updating not only for distributions, written as $\omega|_p$, but also for channels c, via elementwise updating: $c|_q(x) := c(x)|_q$. This is used in the next result.

Theorem 6.3.4 *Let $c\colon X \rightsquigarrow Y$ be a channel with a state $\omega \in \mathcal{D}(X)$ on its domain and a factor $q \in Fact(Y)$ on its codomain. Then:*

$$c_*(\omega)\big|_q \;=\; (c|_q)_*(\omega|_{c^*(q)}). \tag{6.11}$$

Proof. For each $y \in Y$,

$$\begin{aligned}
c_*(\omega)|_q(y) \;=\; \frac{c_*(\omega)(y)\cdot q(y)}{c_*(\omega)\models q} &= \sum_{x\in X} \frac{\omega(x)\cdot c(x)(y)\cdot q(y)}{\omega \models c^*(q)} \\
&= \sum_{x\in X} \frac{\omega(x)\cdot c^*(q)(x)}{\omega\models c^*(q)} \cdot \frac{c(x)(y)\cdot q(y)}{c^*(q)(x)} \\
&= \sum_{x\in X} \omega|_{c^*(q)}(x) \cdot \frac{c(x)(y)\cdot q(y)}{c(x)\models q} \\
&= \sum_{x\in X} \omega|_{c^*(q)}(x)\cdot c(x)|_q(y) \\
&= \sum_{x\in X} \omega|_{c^*(q)}(x)\cdot c|_q(x)(y) \;=\; (c|_q)_*(\omega|_{c^*(q)})(y).
\end{aligned}$$

❑

This result has a number of useful consequences.

Corollary 6.3.5 *For appropriately typed channels, states, and factors:*

1 $(d\circ c)|_q = d|_q \circ c|_{d^*(q)}$, *a form of chain rule;*

2 $\langle c, d\rangle_*(\omega)|_{p\otimes q} = \langle c|_p, d|_q\rangle_*(\omega)|_{c^*(p)\&d^*(q)}$;

3 $(e\otimes f)_*(\omega)|_{p\otimes q} = (e|_p \otimes f|_q)_*(\omega|_{e^*(p)\otimes f^*(q)})$.

Proof. 1 Since:

$$\begin{aligned}
(d\circ c)|_q(x) \;=\; (d\circ c)(x)|_q \;=\; d_*(c(x))|_q &\overset{(6.11)}{=} (d|_q)_*(c(x)|_{d^*(q)}) \\
&= (d|_q)_*(c|_{d^*(q)}(x)) \;=\; (d|_q \circ c|_{d^*(q)})(x).
\end{aligned}$$

2 By Exercise 6.1.17 and Lemma 4.5.2 (7):

$$\langle c, d\rangle_*(\omega)|_{p\otimes q} \overset{(6.11)}{=} (\langle c, d\rangle|_{p\otimes q_*}(\omega|_{\langle c,d\rangle^*(p\otimes q)}) \;=\; \langle c|_p, d|_q\rangle_*(\omega|_{c^*(p)\&d^*(q)}).$$

3 Similarly:

$$\begin{aligned}(e\otimes f)_*(\omega)|_{p\otimes q} &\overset{(6.11)}{=} ((e\otimes f)|_{p\otimes q})_*(\omega|_{(e\otimes f)^*(p\otimes q)})\\ &= (e|_p \otimes f|_q)_*(\omega|_{e^*(p)\otimes f^*(q)}).\end{aligned}$$

❑

Theorem 6.1.6 gives valuable leads about how to increase validities, not only by updating both states, but also by updating channels — see Definition 6.1.1 (3).

Proposition 6.3.6 *Let $c\colon X \rightsquigarrow Y$ be a channel with $\omega \in \mathcal{D}(X)$ and $q \in \mathrm{Fact}(Y)$.*

1 There are inequalities between validities:

$$\begin{tikzcd}
 & c_*(\omega) \models q \arrow[dl, phantom, "\geq"] \arrow[dr, phantom, "\leq"] & \\
c_*(\omega|_{c^*(q)}) \models q \arrow[dr, phantom, "\leq"] & & (c|_q)_*(\omega) \models q \arrow[dl, phantom, "\geq"] \\
 & (c|_q)_*(\omega|_{c^*(q)}) \models q &
\end{tikzcd}$$

2 For an additional channel $d\colon X \rightsquigarrow Z$ and a factor $r \in \mathrm{Fact}(Z)$,

$$\begin{tikzcd}
 & \langle c, d\rangle_*(\omega) \models q\otimes r \arrow[dl, phantom, "\geq"] \arrow[dr, phantom, "\leq"] & \\
\langle c, d\rangle_*(\omega|_{c^*(q)\&d^*(r)}) \models q\otimes r \arrow[dr, phantom, "\leq"] & & \langle c|_q, d|_r\rangle_*(\omega) \models q\otimes r \arrow[dl, phantom, "\geq"] \\
 & \langle c|_q, d|_r\rangle_*(\omega|_{c^*(q)\&d^*(r)}) \models q\otimes r &
\end{tikzcd}$$

Proof. For both points it suffices to prove the upper two inequalities.

1 The upper-left inequality occurred already in Theorem 6.2.2. For proving the upper-right inequality we unpack validity:

$$\begin{aligned}(c|_q)_*(\omega) \models q \;=\; \sum_{y\in Y} (c|_q)_*(\omega)(y)\cdot q(y) &= \sum_{x\in X,\, y\in Y} \omega(x)\cdot c(x)|_q(y)\cdot q(y)\\
&= \sum_{x\in X,\, y\in Y} \omega(x)\cdot \frac{c(x)(y)\cdot q(y)}{c(x)\models q}\cdot q(y)\\
&= \sum_{x\in X} \omega(x)\cdot \frac{\sum_{y\in Y} c(x)(y)\cdot q(y)\cdot q(y)}{c(x)\models q}\\
&= \sum_{x\in X} \omega(x)\cdot \frac{c(x)\models q \,\&\, q}{c(x)\models q}\\
&\geq \sum_{x\in X} \omega(x)\cdot \frac{(c(x)\models q)^2}{c(x)\models q} \qquad \text{by Corollary 5.1.4 (1)}\\
&= \sum_{x\in X} \omega(x)\cdot (c(x)\models q) \;=\; \omega\models c^*(q) \;=\; c_*(\omega)\models q.\end{aligned}$$

2 The upper-left inequality follows from the upper-left inequality in the previous point since $\langle c, d\rangle^*(q \otimes r) = c^*(q) \,\&\, d^*(r)$. This follows from Exercise 4.5.9:

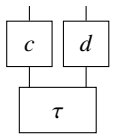
$$\begin{aligned}\langle c, d\rangle^*(q \otimes r) &= ((c \otimes d) \circ \Delta)^*(q \otimes r) = \Delta^*((c \otimes d)^*(q \otimes r)) \\ &= \Delta^*(c^*(q) \otimes d^*(r)) = c^*(q) \,\&\, d^*(r).\end{aligned}$$

Similarly, one obtains the upper-right inequality from point (1) and the equation $\langle c, d\rangle|_{q\otimes r} = \langle c|_q, d|_r\rangle$, which holds by Exercise 6.1.17. ❑

Exercises

6.3.1 Consider a joint distribution $\tau \in \mathcal{D}(X \times Y)$ together with two channel $c\colon X \multimap U$ and $d\colon Y \multimap V$, and define $\omega \in \mathcal{D}(U \times V)$ as:

$$\omega \;:=\; (c \otimes d)_*(\tau) \;=$$



Prove for a factor q on U that:

$$(\omega|_{q\otimes \mathbf{1}})[0, 1] = d_*(\tau|_{c^*(q)\otimes \mathbf{1}}[1, 0]).$$

6.3.2 Consider for a distribution $\omega \in \mathcal{D}(X)$ and a channel $c\colon X \multimap Y$ the associated graph distribution $\langle id, c\rangle_*(\omega) \in \mathcal{D}(X \times Y)$. Show via, or essentially as in, Theorem 6.3.1, that for a factor $q \in Fact(Y)$ one has:

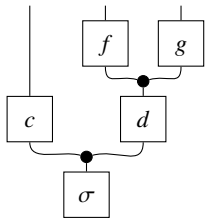
$$\langle id, c\rangle_*(\omega)|_{\mathbf{1}\otimes q}[1, 0] = \omega|_{c^*(q)} \qquad\qquad \langle id, c\rangle_*(\omega)|_{\mathbf{1}\otimes q}[0, 1] = c_*(\omega)|_q.$$

6.3.3 Prove the two equations in Example 6.3.3 (1) directly, via the definition of updating.

6.3.4 Apply (6.10) to diagram (3.20).

6.3.5 Consider a joint distribution ω given by the following string diagram.

$\omega \;=$



Show for two suitably typed factors p, q that:

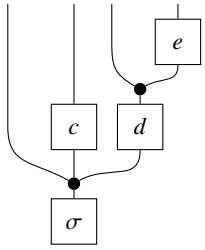
$$\omega\big|_{p\otimes q\otimes \mathbf{1}}[0, 0, 1] = g_*\Big(d_*(\sigma|_{c^*(p)})\big|_{f^*(q)}\Big).$$

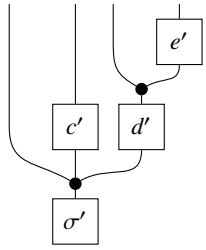
Relate the expression on the right-hand-side of this equation to the shape of the above diagram.

6.3.6 Consider a joint distribution $\omega \in \mathcal{D}(X_1 \times X_2 \times X_3 \times X_4)$ which can be written as:

$$\omega = \Big(id \otimes id \otimes \langle id, e\rangle\Big)_*\Big(\langle id, c, d\rangle_*(\sigma)\Big),$$

as described on the left below. Let $p_i \in Fact(X_i)$ be given. Show that $\omega\big|_{p_1 \otimes p_2 \otimes p_3 \otimes p_4}$ is then of the form as on the right:

$\omega =$ 

$\omega\big|_{p_1 \otimes p_2 \otimes p_3 \otimes p_4} =$ 

Use Theorem 6.3.4 and Corollary 6.3.5 to show specifically that:

$$\sigma' = \sigma\big|_{p_1 \,\&\, c^*(p_2) \,\&\, d^*(p_3 \,\&\, e^*(p_4))} \qquad c' = c\big|_{p_2} \qquad d' = d\big|_{p_3 \,\&\, e^*(p_4)} \qquad e' = e\big|_{p_4}.$$

6.3.7 Let X be a finite set and $\upsilon \in \mathcal{N}(X)$ be an urn with full support. For numbers $K, N \in \mathbb{N}$ and for a multiset $\varphi \in \mathcal{N}[K](X)$, prove that updating a Pólya distribution via backward inference along a hypergeometric channel, as on the left below, can be expressed via a shift:

$$pl[K+N](\upsilon)\big|_{hg[K]^*(\mathbf{1}_\varphi)} = shift\big(\varphi,\, pl[N](\upsilon + \varphi)\big) = \mathcal{D}(- + \varphi)\big(pl[N](\upsilon + \varphi)\big).$$

6.3.8 1 Let $K_1, \ldots, K_N \in \mathbb{N}$ with $\psi \in \mathcal{N}[K](X)$ be given, where $K = \sum_i K_i$. Show that, for any $\omega \in \mathcal{D}(X)$,

$$\begin{aligned}&\big(mn[K_1](\omega) \otimes \cdots \otimes mn[K_N](\omega)\big)\big|_{sum^*(\mathbf{1}_\psi)} \\ &= \sum_{\varphi_1 \leq_{K_1} \psi, \ldots, \varphi_N \leq_{K_N} \psi,\, \sum_i \varphi_i = \psi} \frac{\binom{\psi}{\varphi_1, \ldots, \varphi_N}}{\binom{K}{K_1, \ldots, K_N}} \big|\varphi_1, \ldots, \varphi_N\big\rangle.\end{aligned}$$

(Exercise 1.8.4 guarantees that the probabilities in this sum over multisets add up to one.)

2 Describe the diagram of the form (6.8) that gives rise to the equation in the previous item.

6.3.9 Let $c \colon X \twoheadrightarrow Y$ be a channel, with state $\omega \in \mathcal{D}(X \times Z)$.

1 Prove that for a factor $p \in Fact(Z)$,

$$(c \otimes id)_*(\omega)\big|_{\mathbf{1} \otimes p} = (c \otimes id)_*(\omega\big|_{\mathbf{1} \otimes p}).$$

2 Show also that for $q \in Fact(Y \times Z)$,

$$\Big((c \otimes id)_*(\omega)\big|_q\Big)[0, 1] = \Big(\omega\big|_{(c \otimes id)^*(q)}\Big)[0, 1].$$

6.3.10 Let $c \colon X \twoheadrightarrow Y$ be a channel with a distribution $\omega \in \mathcal{D}(X)$ and with two factors $p \in Fact(X)$, $q \in Fact(Y)$. Check that:

$$(c\big|_q)_*(\omega\big|_{p \& c^*(q)}) = c_*(\omega\big|_p)\big|_q.$$

Check that Theorem 6.3.4 is a special case, for $p = \mathbf{1}$.

6.3.11 Let $p \in Fact(X)$ and $q \in Fact(Y)$ be two factors on spaces X, Y.

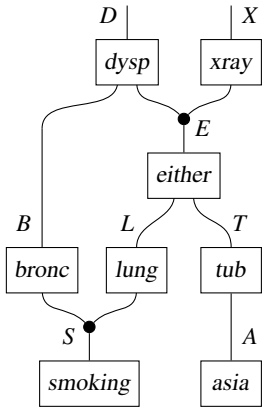


Figure 6.7 The string diagram of the Asia Bayesian network, with node / channel abbreviations: bronc = bronchitis, dysp = dyspnoea, lung = lung cancer, tub = tuberculosis. The wires all have 2-element (yes/no) sets of the form $A = \{a, a^{\perp}\}$. These sets are used in the string diagram as types, annotating the wires.

1 Show that for two channels $c\colon Z \rightsquigarrow X$ and $d\colon Z \rightsquigarrow Y$ with a distribution $\sigma \in \mathcal{D}(Z)$ on their (common) domain, one has:

$$\begin{aligned} \langle c, d\rangle_*(\sigma)|_{p\otimes q}[1,0] &= c_*(\sigma|_{d^*(q)})|_p \\ \langle c, d\rangle_*(\sigma)|_{p\otimes q}[0,1] &= d_*(\sigma|_{c^*(p)})|_q \end{aligned}$$

2 For channels $e\colon U \rightsquigarrow X$ and $d\colon V \rightsquigarrow Y$ with a joint state $\omega \in \mathcal{D}(U \times V)$ one has:

$$\begin{aligned} (e \otimes f)_*(\omega)|_{p\otimes q}[1,0] &= e_*(\omega|_{\mathbf{1}\otimes f^*(q)}[1,0])|_p \\ (e \otimes f)_*(\omega)|_{p\otimes q}[0,1] &= f_*(\omega|_{e^*(p)\otimes \mathbf{1}}[0,1])|_q. \end{aligned}$$

6.3.12 Use Proposition 6.3.6 (1) to prove, for a channel $c\colon X \rightsquigarrow Y$ and a factor q on Y, the following (pointwise) inequality between factors.

$$c^*(q) \;\leq\; (c|_q)^*(q).$$

6.4 Inference in Bayesian networks

In previous sections we have seen several examples of channel-based inference, in forward and backward form. This section shows how to apply these inference methods to Bayesian networks, via an example that is often used in the literature: the 'Asia' Bayesian network, originally from Lauritzen and Spiegelhalter (1988). It captures the situation of patients with a certain probability of smoking and of an earlier visit to Asia; this influences certain lung diseases and the outcome of an xray test.

The Bayesian network example considered here is described in several steps: Figure 6.7 contains the network in the form of a string diagram, where boxes represent channels, and where copiers (black dots) and types of wires are written explicitly. Figure 6.8 gives the conditional probability tables associated with the nodes of this network, in traditional style.

$P(\text{smoke})$
0.5

smoke	$P(\text{bronc})$
s	0.6
$s^\perp$	0.3

smoke	$P(\text{lung})$
s	0.1
$s^\perp$	0.01

$P(\text{asia})$
0.01

asia	$P(\text{tub})$
a	0.05
$a^\perp$	0.01

either	$P(\text{xray})$
e	0.98
$e^\perp$	0.05

bronc	either	$P(\text{dysp})$
b	e	0.9
b	$e^\perp$	0.7
$b^\perp$	e	0.8
$b^\perp$	$e^\perp$	0.1

lung	tub	$P(\text{either})$
ℓ	t	1
ℓ	$t^\perp$	1
$\ell^\perp$	t	1
$\ell^\perp$	$t^\perp$	0

Figure 6.8 The conditional probability tables of the Asia Bayesian network, copied from (Lauritzen and Spiegelhalter, 1988, Table 1).

$$
\begin{aligned}
\mathit{smoke} &= 0.5|s\rangle + 0.5|s^\perp\rangle \\
\mathit{lung}(s) &= 0.1|\ell\rangle + 0.5|\ell^\perp\rangle \\
\mathit{lung}(s^\perp) &= 0.01|\ell\rangle + 0.99|\ell^\perp\rangle \\
\mathit{bronc}(s) &= 0.6|b\rangle + 0.4|b^\perp\rangle \\
\mathit{bronc}(s^\perp) &= 0.3|b\rangle + 0.7|b^\perp\rangle \\
\mathit{dysp}(b, e) &= 0.9|d\rangle + 0.1|d^\perp\rangle \\
\mathit{dysp}(b, e^\perp) &= 0.7|d\rangle + 0.3|d^\perp\rangle \\
\mathit{dysp}(b^\perp, e) &= 0.8|d\rangle + 0.2|d^\perp\rangle \\
\mathit{dysp}(b^\perp, e^\perp) &= 0.1|d\rangle + 0.9|d^\perp\rangle
\end{aligned}
\qquad
\begin{aligned}
\mathit{asia} &= 0.01|a\rangle + 0.99|a^\perp\rangle \\
\mathit{tub}(a) &= 0.05|t\rangle + 0.95|t^\perp\rangle \\
\mathit{tub}(a^\perp) &= 0.01|t\rangle + 0.99|t^\perp\rangle \\
\mathit{xray}(e) &= 0.98|x\rangle + 0.02|x^\perp\rangle \\
\mathit{xray}(e^\perp) &= 0.05|x\rangle + 0.95|x^\perp\rangle \\
\mathit{either}(\ell, t) &= 1|e\rangle \\
\mathit{either}(\ell, t^\perp) &= 1|e\rangle \\
\mathit{either}(\ell^\perp, t) &= 1|e\rangle \\
\mathit{either}(\ell^\perp, t^\perp) &= 1|e^\perp\rangle
\end{aligned}
$$

Figure 6.9 The conditional probability tables from Figure 6.8, reformulated as distributions and channels.

Figure 6.9 reformulates these probability tables as distributions and channels, so that channel-based reasoning techniques can be used (as introduced in Jacobs and Zanasi (2016) for Bayesian networks). Specifically, this Figure 6.9 introduces distributions $\mathit{smoke} \in \mathcal{D}(S)$ and $\mathit{asia} \in \mathcal{D}(A)$ and channels $\mathit{lung}\colon S \multimapinv L$, $\mathit{tub}\colon A \multimapinv T$, $\mathit{bronc}\colon S \multimapinv B$, $\mathit{xray}\colon E \multimapinv X$, $\mathit{dysp}\colon B \times E \multimapinv D$, $\mathit{either}\colon L \times T \multimapinv E$.

The aim of this section is illustrate channel-based inference in this Asia Bayesian network. It is not so much the actual outcomes that we are interested in, but more the systematic methodology that is used to obtain these outcomes. This methodology involves sequential and parallel composition of channels (written as $\odot$ and $\otimes$), transformation of predicates and states along channels, and updating of distributions.

Probability of lung cancer, given no bronchitis

Let's start with the question: what is the probability that someone has lung cancer, given that this person does not have bronchitis. The latter information is the evidence. It takes the form of a point predicate $\mathbf{1}_{b^\perp} = (\mathbf{1}_b)^\perp \colon B \to [0,1]$ on the set $B = \{b, b^\perp\}$ used for presence and absence of bronchitis.

In order to obtain this updated probability of lung cancer we 'follow the graph', as in Theorem 6.3.1. In Figure 6.7 we see that we can transform (pull back) the evidence along the bronchitis channel $\mathit{bronc} \colon S \rightsquigarrow B$, and obtain a predicate $\mathit{bronc}^*(\mathbf{1}_{b^\perp})$ on S. The latter can be used to update the smoking distribution on S. Subsequently, we can push the updated distribution forward along the lung channel $\mathit{lung} \colon S \rightsquigarrow L$ via state transformation. Thus we follow the 'V' shape in the relevant part of the graph, that we studied on its own in Theorem 6.3.1.

Combining this down-update-up steps gives the required outcome:

$$\mathit{lung}_*(\mathit{smoke}\big|_{\mathit{bronc}^*(\mathbf{1}_{b^\perp})}) \;=\; 0.0427|\ell\rangle + 0.9573|\ell^\perp\rangle. \tag{6.12}$$

We see that this calculation combines forward and backward inference, see Definition 6.2.1.

Probability of smoking, given a positive xray

In Figures 6.8 and 6.9 we see a prior smoking probability of 50%. We like to know what this probability becomes if we have evidence of a positive xray. The latter is given by the point predicate $\mathbf{1}_x \in \mathit{Pred}(X)$ for $X = \{x, x^\perp\}$.

There is a long path (down) from xray to smoking, see Figure 6.7, that we need to use for (backward) predicate transformation. Along the way there is a slight complication, namely that the node 'either' has two parent nodes, so that pulling back along the *either* channel yields a predicate on the product set $L \times T$. The only sensible thing to do is to continue predicate transformation downwards, but now with the parallel product channel $\mathit{lung} \otimes \mathit{tub} \colon S \times A \rightsquigarrow L \times T$. The resulting predicate on $S \times A$ can be used to update the product state $\mathit{smoke} \otimes \mathit{asia}$. Then we can take the first marginal to obtain the desired outcome. Thus we compute:

$$\begin{aligned}
&\Big((\mathit{smoke} \otimes \mathit{asia})\big|_{(\mathit{lung}\otimes\mathit{tub})^*\mathit{either}^*(\mathit{xray}^*(\mathbf{1}_x))}\Big)[1,0] \\
&= \Big((\mathit{smoke} \otimes \mathit{asia})\big|_{(\mathit{xray} \circ \mathit{either} \circ (\mathit{lung}\otimes\mathit{tub}))^*(\mathbf{1}_x)}\Big)[1,0] \;=\; 0.6878|s\rangle + 0.3122|s^\perp\rangle.
\end{aligned} \tag{6.13}$$

Thus, a positive xray makes it more likely — w.r.t. the uniform prior — that the patient smokes — as is to be expected. This is obtained by backward inference.

Probability of lung cancer, given both dyspnoea and tuberculosis

Our next inference challenge involves two evidence predicates, namely $\mathbf{1}_d$ on D for dyspnoea and $\mathbf{1}_t$ on T for tuberculosis. We would like to know the updated lung cancer probability.

The situation looks complicated, because of the 'closed loop' in Figure 6.7. But we can proceed in a straightforward manner and combine evidence via conjunction & at a suitable meeting point. We now clearly separate the forward and backward stages of the inference process. We first move the prior states forward to a point that includes the set L — the one that we need to marginalise on to get our conclusion. We abbreviate this state on $B \times L \times T$ as:

$$\begin{aligned}
\sigma &:= (\mathit{bronc} \otimes \mathit{lung} \otimes \mathit{id})_*((\Delta \otimes \mathit{tub})_*(\mathit{smoke} \otimes \mathit{asia})) \\
&= (\langle \mathit{bronc}, \mathit{lung}\rangle \otimes \mathit{tub})_*(\mathit{smoke} \otimes \mathit{asia}) \\
&= \langle \mathit{bronc}, \mathit{lung}\rangle_*(\mathit{smoke}) \otimes \mathit{tub}_*(\mathit{asia}).
\end{aligned}$$

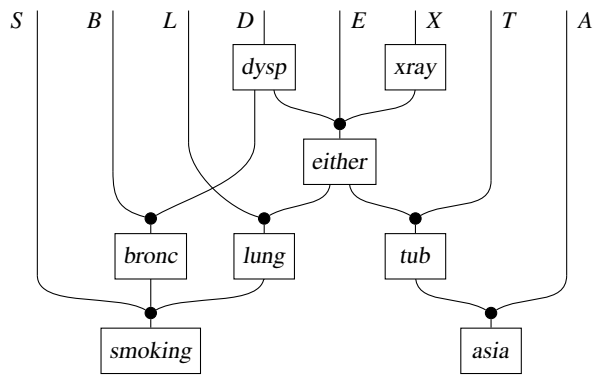


Figure 6.10 The 'joint' version of the Asia network from Figure 6.7, where each wire is copied to the top. This gives a joint distribution α on the product space $S \times B \times L \times D \times E \times X \times T \times A$.

Recall that we write Δ for the copy channel, in this expression of type $S \multimapinv S \times S$.

Going in the backward direction we can form a predicate, called p below, on the set $B \times L \times T$, by predicate transformation *and* conjunction:

$$p \coloneqq (\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1}_t) \mathbin{\&} (id \otimes either)^*(dys^*(\mathbf{1}_d)).$$

The result that we are after is now obtained via updating and marginalisation:

$$\sigma|_p[0,1,0] \;=\; 0.0558|\ell\rangle + 0.9442|\ell^\perp\rangle. \tag{6.14}$$

There is an alternative way to describe the same outcome, using that certain channels can be 'shifted'. In particular, in the definition of the above state σ, the channel *bronc* is used for state transformation. It can also be used in a different role, namely for predicate transformation. We then use a slightly different state, now on $S \times L \times T$,

$$\begin{aligned}\tau &\coloneqq (id \otimes lung \otimes id)_*((\Delta \otimes tub)_*(smoke \otimes asia))\\ &= (\langle id, lung\rangle \otimes tub)_*(smoke \otimes asia)\\ &= \langle id, lung\rangle_*(smoke) \otimes tub_*(asia).\end{aligned}$$

The *bronc* channel is now used for predicate transformation in the predicate:

$$q \coloneqq (\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1}_t) \mathbin{\&} (bronc \otimes either)^*(dys^*(\mathbf{1}_d))$$

The same updated lung cancer distribution is now obtained as:

$$\tau|_q[0,1,0] \;=\; 0.0558|\ell\rangle + 0.9442|\ell^\perp\rangle. \tag{6.15}$$

The reason why the outcomes (6.14) and (6.15) are the same is the topic of Exercise 6.4.3.

We conclude that inference in Bayesian networks can be done compositionally via a combination of forward and backward inference, basically by following the network structure, when represented as a string diagram. The same update outcomes as computed here can be found in ProbLog and WebPPL representations of this network in Examples 6.5.5 and 6.6.4, but without the above computation mechanisms.

In the end we like to remark that one can also use the joint distribution associated with the network for probabilistic reasoning. This is conceptually relevant, but not very efficient, because these joint distributions quickly become very large.

Let us write $\alpha \in \mathcal{D}(S \times B \times L \times D \times E \times X \times T \times A)$ for the joint distribution given by the string diagram in Figure 6.10. It arises by making sure that each distribution and channel in the network has an outgoing wire. We shall not describe this distribution α explicitly, since its underlying space has $2^8 = 256$ elements.

Recall that earlier we asked for the lung cancer probability, given no bronchitis. In the product space $S \times B \times L \times D \times E \times X \times T \times A$ underlying the distribution α, the evidence $\mathbf{1}_{b^\perp}$ is a predicate on the set B, which is the second product component. We are interested in the resulting L-marginal, which is at the third position. Thus we first weaken the evidence $\mathbf{1}_{b^\perp}$ to a predicate $(\pi_2)^*(\mathbf{1}_{b^\perp})$ on the entire product, via the second projection $\pi_2 \colon S \times B \times L \times D \times E \times X \times T \times A \to B$. Similarly, we use the third projection π_3 for marginalisation, to get as outcome:

$$(\pi_3)_*(\alpha|_{(\pi_2)^*(\mathbf{1}_{b^\perp})}) = (\pi_3)_*(\alpha|_{\mathbf{1}\otimes\mathbf{1}_{b^\perp}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}}) = 0.0427|\ell\rangle + 0.9573|\ell^\perp\rangle. \tag{6.16}$$

The fact that we get the same outcome as in the earlier combination (6.12) of backward and forward inference follows essentially from Theorem 6.3.1.

Similarly, the smoking distribution (6.13) of given a positive xray can be obtained via the joint distribution α as:

$$(\pi_1)_*(\alpha|_{(\pi_6)^*(\mathbf{1}_x)}) = (\pi_1)_*(\alpha|_{\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_x\otimes\mathbf{1}\otimes\mathbf{1}}) = 0.6878|s\rangle + 0.3122|s^\perp\rangle. \tag{6.17}$$

The lung cancer distributions (6.14) and (6.15) given dyspnoea and tuberculosis is:

$$(\pi_3)_*(\alpha|_{(\pi_4)^*(\mathbf{1}_d)\,\&\,(\pi_7)^*(\mathbf{1}_t)}) = (\pi_3)_*(\alpha|_{\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}_d\otimes\mathbf{1}\otimes\mathbf{1}_t\otimes\mathbf{1}}) = 0.0558|\ell\rangle + 0.9442|\ell^\perp\rangle. \tag{6.18}$$

These inferences via the joint distribution are often computationally infeasible. However, the inferences may work well via backward and forward reasoning along the channels in the Bayesian network. Making such reasoning possible is one of the main reasons for introducing these Bayesian networks representations in the first place.

The above formulas illustrate a general pattern, involving weakening and marginalisation via projections. Accordingly, we may define an *inference query* on a joint state $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n)$ with evidence $p \in Fact(X_i)$ as input and an output distribution on X_j:

$$(\pi_j)_*(\omega|_{(\pi_i)^*(p)}) \in \mathcal{D}(X_j). \tag{6.19}$$

It combines forward and backward inference along projection channels. When the joint state ω is 'factored' in terms of channels — as illustrated in the string diagram in Figure 6.10 — such inference query (6.19) can be computed more efficiently via inference along the channels — as illustrated in this section. How one factors a joint state into a string diagram of channels is described in Section 7.8.

Exercises

6.4.1 Consider the wetness Bayesian network from Subsection 2.5.2. Write down the channel-based inference formulas for the following inference questions and check the outcomes that are given below.

1 The updated sprinkler distribution, given evidence of a slippery road, is $\frac{63}{260}|b\rangle + \frac{197}{260}|b\rangle$.

2 The updated wet grass distribution, given evidence of a slippery road, is $\frac{4349}{5200}|d\rangle + \frac{851}{5200}|d^{\perp}\rangle$.

6.4.2 We continue with the wetness network.

1 Add copiers and wires to the string diagram in Figure 2.5 to get a joint distribution on the product set $A \times B \times D \times C \times E$ — just like the diagram in Figure 6.10 is a 'joint' version of the original diagram in Figure 6.7.

2 Check that this diagram corresponds to the following (sequential and parallel) composite of channels, giving as joint distribution $\omega \in \mathcal{D}(A \times B \times D \times C \times E)$.

$$\omega = (id \otimes id \otimes wg \otimes id \otimes sr) \circ (id \otimes \Delta_2 \otimes \Delta_3) \circ (id \otimes sp \otimes sr) \circ \Delta_3 \circ wi.$$

3 Explicitly, this joint distribution can be written as:

$$\begin{aligned}\omega = {} & \tfrac{399}{6250}|a,b,d,c,e\rangle + \tfrac{171}{6250}|a,b,d,c,e^{\perp}\rangle + \tfrac{27}{1250}|a,b,d,c^{\perp},e^{\perp}\rangle + \tfrac{21}{6250}|a,b,d^{\perp},c,e\rangle \\ & + \tfrac{9}{6250}|a,b,d^{\perp},c,e^{\perp}\rangle + \tfrac{3}{1250}|a,b,d^{\perp},c^{\perp},e^{\perp}\rangle + \tfrac{672}{3125}|a,b^{\perp},d,c,e\rangle \\ & + \tfrac{288}{3125}|a,b^{\perp},d,c,e^{\perp}\rangle + \tfrac{168}{3125}|a,b^{\perp},d^{\perp},c,e\rangle + \tfrac{72}{3125}|a,b^{\perp},d^{\perp},c,e^{\perp}\rangle \\ & + \tfrac{12}{125}|a,b^{\perp},d^{\perp},c^{\perp},e^{\perp}\rangle + \tfrac{399}{20000}|a^{\perp},b,d,c,e\rangle + \tfrac{171}{20000}|a^{\perp},b,d,c,e^{\perp}\rangle \\ & + \tfrac{243}{1000}|a^{\perp},b,d,c^{\perp},e^{\perp}\rangle + \tfrac{21}{20000}|a^{\perp},b,d^{\perp},c,e\rangle + \tfrac{9}{20000}|a^{\perp},b,d^{\perp},c,e^{\perp}\rangle \\ & + \tfrac{27}{1000}|a^{\perp},b,d^{\perp},c^{\perp},e^{\perp}\rangle + \tfrac{7}{1250}|a^{\perp},b^{\perp},d,c,e\rangle + \tfrac{3}{1250}|a^{\perp},b^{\perp},d,c,e^{\perp}\rangle \\ & + \tfrac{7}{5000}|a^{\perp},b^{\perp},d^{\perp},c,e\rangle + \tfrac{3}{5000}|a^{\perp},b^{\perp},d^{\perp},c,e^{\perp}\rangle + \tfrac{9}{100}|a^{\perp},b^{\perp},d^{\perp},c^{\perp},e^{\perp}\rangle\end{aligned}$$

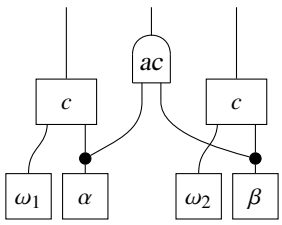
Express the two inferences in the previous exercise as suitable marginalisations of this joint distribution ω, updated with suitably weakened predicates, as in (6.16), (6.17), (6.18).

If you feel mathematically sufficiently fit, you may even recalculate the resulting distributions in this way, as given in the previous exericse.

6.4.3 Check that the equality of the outcomes in (6.14) and in (6.15) can be explained via Exercises 6.3.9 (2) and 4.5.11.

6.4.4 The following illustration comes from a legal context Fenton et al. (2016) and addresses what is sometimes called Cohen's paradox: if one has two pieces of evidence A and B, each with probabibility 0.7, above the (acceptance) threshold of 0.5, their conjunction $A \,\&\, B$ has probability $0.7 \cdot 0.7 = 0.49$ and is thus below the threshold (see also Dawid (1987)). The situation is redescribed in Fenton et al. (2016) in terms of a Bayesian network, with two witnesses ω_1, ω_2, two events α, β, and a credibility channel c, and finally a conjunction channal ac, as in Exercise 2.5.9.

$\gamma \quad =$



(6.20)

The witness distributions $\omega_1, \omega_2 \in \mathcal{D}(R)$ have as sample space $R = \{r, r^{\perp}\}$, where r stands for reliable. Both witnesses are 70% reliable, as expressed by $\omega_1 = \omega_2 = \frac{7}{10}|r\rangle + \frac{3}{10}|r^{\perp}\rangle$.

The event distributions $\alpha, \beta \in \mathcal{D}(\mathbf{2})$ are both uniform, where $\mathbf{2} = \{0, 1\}$ is used for Booleans, with $1 = \mathit{true}$. The credibility channel $c \colon R \times \mathbf{2} \rightsquigarrow \mathbf{2}$ is defined as:

$$c(x, y) \;=\; \begin{cases} 1|y\rangle & \text{if } x = r \\ \frac{1}{2}|1\rangle + \frac{1}{2}|0\rangle & \text{if } x = r^{\perp}. \end{cases}$$

Thus, when dealing with a reliable person, the Boolean input $y \in \mathbf{2}$ is passed on, and for an unreliable person the uniform distribution on $\mathbf{2}$ is returned.

1 Consider the distribution $\gamma \in \mathcal{D}(\mathbf{2} \times \mathbf{2} \times \mathbf{2})$ determined in (6.20), assume what the witnesses claim is true, and concentrate on the conjunction, as on the left below. Check this outcome.

$$\big(\gamma|_{\mathbf{1}_1 \otimes \mathbf{1} \otimes \mathbf{1}_1}\big)[0, 1, 0] \;=\; \tfrac{289}{400}|1\rangle + \tfrac{111}{400}|1\rangle \;=\; 0.7225|1\rangle + 0.2775|0\rangle.$$

This conjunction outcome is now above the threshold of one half — because of the way that the credibility channel c handles the unreliable case.

2 Show that this same outcome can be obtained as:

$$ac_*\big(\alpha' \otimes \beta'\big) \qquad \text{where} \qquad \begin{cases} \alpha' \;:=\; \big((\omega_1 \otimes \alpha)|_{c^*(\mathbf{1}_1)}\big)[0, 1] \\ \beta' \;:=\; \big((\omega_2 \otimes \beta)|_{c^*(\mathbf{1}_1)}\big)[0, 1]. \end{cases}$$

Yet another way to compute this update appears in Exercise 7.10.6.

6.5 Updating in probabilistic logic programming

This section sketches how updating works in the probabilistic programming language ProbLog Raedt et al. (2007); Raedt and Kimmig (2015); Fierens et al. (2015); Derkinderen et al. (2024) — already introduced in Section 2.8. The exposition below starts with predicates and how they are used for updating in ProbLog. It subsequently shows how backward reasoning (along a channel) can be added, via a reformulation (in ProbLog) of several examples from Section 6.2. Finally, inference / updating in the Bayesian network from Section 6.4 is (re)done in ProbLog.

We start with a review of the updating of a fair dice with the 'evenish' predicate from Example 6.1.2 (1). In ProbLog it looks as follows.

```
1/6 :: pips(1); 1/6 :: pips(2); 1/6 :: pips(3);
1/6 :: pips(4); 1/6 :: pips(5); 1/6 :: pips(6).

2/10 :: evenish :- pips(1).
9/10 :: evenish :- pips(2).
1/10 :: evenish :- pips(3).
9/10 :: evenish :- pips(4).
1/10 :: evenish :- pips(5).
8/10 :: evenish :- pips(6).

query(evenish).
```
(6.21)

The first two lines define the fair dice distribution $\mathit{dice} = \sum_{1 \leq i \leq 6} \frac{1}{6}|i\rangle$. The next six lines introduce the `evenish` predicate. Using point predicate $\mathbf{1}_i$ this predicate is:

$$\mathit{evenish} \;=\; \tfrac{2}{10} \cdot \mathbf{1}_1 + \tfrac{9}{10} \cdot \mathbf{1}_2 + \tfrac{1}{10} \cdot \mathbf{1}_3 + \tfrac{9}{10} \cdot \mathbf{1}_4 + \tfrac{1}{10} \cdot \mathbf{1}_5 + \tfrac{8}{10} \cdot \mathbf{1}_6.$$

Recall from Section 2.8 that one can also use an expression `not evenish`, as orthosupplement $evenish^{\perp}$. Indeed, the above description can be read both as a predicate, or as a binary channel, using the correspondence between predicates $X \to [0,1]$ and binary channels $X \rightsquigarrow \mathbf{2}$, using that $\mathcal{D}(\mathbf{2}) \cong [0,1]$, see Remark 4.5.4.

The query at the end of (6.21) leads to the output 0.5, as validity $dice \models evenish$.

Updating in ProbLog is done via the keyword `evidence`. For instance, consider adding the next two lines to (6.21).

```
evidence(evenish).
query(pips(_)).
```

The order of these two lines is irrelevant. The evidence statement leads to an update of the prior *dice* distribution to the posterior version $dice|_{evenish}$ that we have seen in Example 6.1.2 (1):

$$dice|_{evenish} = \tfrac{1}{15}|1\rangle + \tfrac{3}{10}|2\rangle + \tfrac{1}{30}|3\rangle + \tfrac{3}{10}|4\rangle + \tfrac{1}{30}|5\rangle + \tfrac{4}{15}|6\rangle.$$

In presence of the above `evidence(evenish)` statement the pips query no longer returns the original (prior) dice probabilities, but the probabilities of the updated distribution $dice|_{evenish}$ namely:

```
pips(1): 0.066666667          pips(4): 0.3
pips(2): 0.3                  pips(5): 0.033333333
pips(3): 0.033333333          pips(6): 0.26666667
```

One can ask what the effect is of adding another evidence statement, as in:

```
evidence(evenish).
evidence(evenish).
query(pips(_)).
```

This recurrence has no effect. One might have expected the probabilities of repeated updating $dice|_{evenish}|_{evenish} = dice|_{evenish \,\&\, evenish}$ as outcome, but ProbLog first performs logical simplification where a conjunction & of two predicates is performed in a Boolean manner, see Remark 2.8.4. Repeated updates are possible via different instantiations, see Example 6.5.4 below.

Another issue is that the prior distribution is no longer accessible after the update, for comparison. One can use parametrisation with variables and use two instances of the dice to keep the different versions apart, see Figure 6.11. Its output is:

```
prior(1): 0.16666667          posterior(1): 0.066666667
prior(2): 0.16666667          posterior(2): 0.3
prior(3): 0.16666667          posterior(3): 0.033333333
prior(4): 0.16666667          posterior(4): 0.3
prior(5): 0.16666667          posterior(5): 0.033333333
prior(6): 0.16666667          posterior(6): 0.26666667
```

We now reformulate three of the update illustrations from Section 6.2 in ProbLog.

Example 6.5.1 In Example 6.2.3 we saw two urns with black and white balls, in different

```
1/6 :: pips(V,1); 1/6 :: pips(V,2); 1/6 :: pips(V,3);
1/6 :: pips(V,4); 1/6 :: pips(V,5); 1/6 :: pips(V,6) :- dice(V).

dice(a).
dice(b).

prior(X) :- pips(a, X).
posterior(X) :- pips(b, X).

2/10 :: evenish(V) :- pips(V,1).
9/10 :: evenish(V) :- pips(V,2).
1/10 :: evenish(V) :- pips(V,3).
9/10 :: evenish(V) :- pips(V,4).
1/10 :: evenish(V) :- pips(V,5).
8/10 :: evenish(V) :- pips(V,6).

query(prior(_)).
query(posterior(_)).
evidence(evenish(b)).
```

Figure 6.11 Using separate prior and posterior distributions in ProbLog via different instantiations of a parameterised dice distribution.

proportions, with a fair coin deciding from which urn we draw a ball. The draw turns out to be white, and the posterior coin bias is sought. This can be done in ProbLog as follows.

```
1/2 :: head.

2/9  :: white; 7/9  :: black :- head.
5/11 :: white; 6/11 :: black :- not head.

evidence(white).
query(head).
```

The first urn contains two white and seven black balls. In the above formalisation we have turned these numbers into probabilities, essentially by frequentist learning. The same is done for the second urn, associated with the other (non-head) side of the coin.

Running this program produces the posterior head probability 0.32835821. It corresponds to the fraction $\frac{22}{67}$ that we computed in Example 6.2.3. The predicate transformation that we used there to turn the ball evidence into evidence on the coin remains implicit in ProbLog.

Example 6.5.2 In Example 6.2.6 we used an empirical distribution originally stemming from a table (1.31), with two-dimenensions, containing data about combinations of blood pressure (high or low) and medicine usage (no medicine, medicine 1, or medicine 2). We had evidence about the blood pressure, both sharp and fuzzy, and were interested in the associated medicine usage. In ProbLog the situation is formalised in Figure 6.12.

```
10/100 :: blood_med(low,0);  5/100  :: blood_med(high,0);
35/100 :: blood_med(low,1);  10/100 :: blood_med(high,1);
25/100 :: blood_med(low,2);  15/100 :: blood_med(high,2).

blood_marginal(X) :- blood_med(X,Y).
med_marginal(Y)   :- blood_med(X,Y).

high  :- blood_med(low,_).
low   :- blood_med(high,_).
nomed :- blood_med(_,0).
med1  :- blood_med(_,1).
med2  :- blood_med(_,2).

query(high).
query(nomed).

query(med_marginal(_)).
```

Figure 6.12 ProbLog version of the earlier medicine-blood illustration, see Example 6.5.2.

In the first three lines the table (1.31) is described, as a joint distribution τ. Subsequently the two marginals are defined and the relevant predicates. The first two queries yield the validities for high blood pressure and no-medicine. Using the validity notation $\models$ one needs to apply weakening of predicates or marginalisation of distributions to compute these validities, as in:

$$\begin{aligned} \tau \models \mathbf{1}_H \otimes \mathbf{1} &= \mathcal{D}(\pi_1)(\tau) \models \mathbf{1}_H = \tfrac{7}{10} = 0.7 \\ \tau \models \mathbf{1} \otimes \mathbf{1}_0 &= \mathcal{D}(\pi_2)(\tau) \models \mathbf{1}_0 = \tfrac{3}{20} = 0.15. \end{aligned}$$

This marginalisation / weaking occurs in ProbLog in the form of the underscore _ in the definitions of the predicates `high` *etc*. In this way they are defined immediately as binary predicates.

The last query in the above code fragment yields the medicine marginal. It was called $\omega = \frac{3}{20}|0\rangle + \frac{9}{20}|1\rangle + \frac{2}{5}|2\rangle$ in Example 6.2.6.

We are interested in the updated medicine marginal, in presence of information about the blood pressure. This is done in two ways, with sharp and fuzzy evidence.

- When we have sharp evidence of a high blood pressure we add to the above program the statement `evidence(high)`. The program then prints the posterior medicine distribution, with probabilities:

```
med_marginal(0): 0.14285714
med_marginal(1): 0.5
med_marginal(2): 0.35714286
```

  This corresponds to the updated distribution $\omega|_{b^\circ(\mathbf{1}_H)} = \frac{1}{7}|0\rangle + \frac{1}{2}|1\rangle + \frac{5}{14}|2\rangle$ from Example 6.2.6.
- In Example 6.2.6 we also used an alternative, fuzzy predicate as blood pressure evidence, namely $q = \frac{19}{20} \cdot \mathbf{1}_H + \frac{1}{20} \cdot \mathbf{1}_L$. It describes 'highish' blood pressure. We can formalise it in

ProbLog as follows, and then use it as evidence in:

```
19/20 :: highish :- high.
1/20  :: highish :- low.

evidence(highish).
```

ProbLog now gives slightly different posterior medicine probabilities:

```
med_marginal(0): 0.14338235
med_marginal(1): 0.49632353
med_marginal(2): 0.36029412
```

They correspond to the posterior $\omega|_{b^*(q)} = \frac{39}{272}|0\rangle + \frac{135}{272}|1\rangle + \frac{49}{136}|2\rangle$ that we computed earlier in Example 6.2.6.

This example involves crossover effects, like in Example 6.1.2 (2), where updating with evidence in one component of a joint distribution affects the other component. ProbLog handles this crossover appropriately, without making it explicit.

Example 6.5.3 We also briefly review the three prisoner puzzle from Example 6.2.8 in ProbLog. Without further ado, it can be written as follows.

```
1/3 :: gets_pardoned(a); 1/3 :: gets_pardoned(b); 1/3 :: gets_pardoned(c).

1/2 :: gets_executed(b); 1/2 :: gets_executed(c)  :- gets_pardoned(a).
1   :: gets_executed(c)                           :- gets_pardoned(b).
1   :: gets_executed(b)                           :- gets_pardoned(c).

evidence(gets_executed(b)).
query(gets_pardoned(_)).
```

The second block of three lines contains the answer of the warden to prisoner `a` about who gets executed. Recall that the warden should not lie and also should not reveal any information to `a` about `a`'s own fate. Running the programming nicely gives as posterior probabilities:

```
gets_pardoned(a): 0.33333333
gets_pardoned(b): 0
gets_pardoned(c): 0.66666667
```

So indeed, prisoner `a` does not get wiser from the warden's answer: the posterior probability of getting pardoned is (for him) the same as the prior probability.

Example 6.5.4 We recall bag of cherry and lime candies from Example 6.2.9, with multiple successive draws appearing to be lime. Doing this in ProbLog requires a bit of care, since, as we have seen in the beginning of this section, simply repeating an evidence statement has no effect.

This can be circumvented via parametrisation, as below

```
1/10 :: bag(1); 1/5 :: bag(2); 2/5 :: bag(3); 1/5 :: bag(4); 1/10 :: bag(5).

1   :: cherry(X); 0   :: lime(X) :- bag(1).
3/4 :: cherry(X); 1/4 :: lime(X) :- bag(2).
1/2 :: cherry(X); 1/2 :: lime(X) :- bag(3).
1/4 :: cherry(X); 3/4 :: lime(X) :- bag(4).
0   :: cherry(X); 1   :: lime(X) :- bag(5).

evidence(lime(0)).
evidence(lime(1)).
evidence(lime(2)).
```

We see that the channel is parameterised, via a variable x, but the bag distribution is not. In this way one can add repeated evidence with different instantiations to perform multiple updates. The outcome after the above three lime evidence statements give the posterior bag distribution:

```
bag(1): 0
bag(2): 0.013157895
bag(3): 0.21052632
bag(4): 0.35526316
bag(5): 0.42105263
```

This corresponds to the posterior $\omega|_{c^*(\mathbf{1}_L)}|_{c^*(\mathbf{1}_L)}|_{c^*(\mathbf{1}_L)} = \omega|_{c^*(\mathbf{1}_L)\,\&\,c^*(\mathbf{1}_L)\,\&\,c^*(\mathbf{1}_L)}$ that we computed in Example 6.2.9 after three updates.

The nine successive updates that we saw earlier in Figure 6.1 can be obtained separately by running the above program with sufficiently many evidence statements. One can do this more efficiently in ProbLog, for instance via a single parametrised statement:

```
evidence(lime(N)) :- between(0,9,N).
```

It gives the result of ten succesive updates.

Example 6.5.5 We also look at an example of inference in Bayesian networks, in ProbLog. We reconsider the Asia example from Section 6.4. Its ProbLog description is in Figure 6.13. This network involves only binary channels, making the formalisation relatively straightforward — and very similar to the ProbLog formalisation of the wetness Bayesian network in Figure 2.8. The only noticeable thing is that 'either' is no longer a channel, as in Section 6.4, but is handled by the sum mechanism of ProbLog.

In Section 6.4 we elaborated three separate updating questions in terms of updating with forward and backward transformation along the network's channels (as denotations of conditional probability tables). Here we briefly review these three cases, as separate evidence-query matters in ProbLog. Unsurprisingly, the outcomes coincide with what we computed in Section 6.4.

- The first question concerns the probability of lung cancer, given no bronchitis. Below we give

```
0.5  :: smoking.
0.01 :: asia_visit.

0.1  :: lung_cancer :- smoking.
0.01 :: lung_cancer :- not smoking.

0.05 :: tuberculosis :- asia_visit.
0.01 :: tuberculosis :- not asia_visit.

0.6 :: bronchitis :- smoking.
0.3 :: bronchitis :- not smoking.

lung_cancer_or_tuberculosis :- lung_cancer.
lung_cancer_or_tuberculosis :- tuberculosis.

0.98 :: xray :- lung_cancer_or_tuberculosis.
0.05 :: xray :- not lung_cancer_or_tuberculosis.

0.9 :: dyspnoea :- bronchitis,     lung_cancer_or_tuberculosis.
0.7 :: dyspnoea :- bronchitis,     not lung_cancer_or_tuberculosis.
0.8 :: dyspnoea :- not bronchitis, lung_cancer_or_tuberculosis.
0.1 :: dyspnoea :- not bronchitis, not lung_cancer_or_tuberculosis.
```

Figure 6.13 Formalisation of the Asia Bayesian network from Section 6.4 in ProbLog, see Example 6.5.5.

the ProbLog formalisation on the left, with the resulting output on the right.

```
evidence(not bronchitis).
query(lung_cancer).
```

```
lung_cancer: 0.042727273
```

- The second question is about the probability of smoking, given a positive xray. This is obtained as:

```
evidence(xray).
query(smoking).
```

```
smoking: 0.68775385
```

- The third case is about the probability of lung cancer, given both dyspnoea and tuberculosis. In this case we need two separate evidence statements:

```
evidence(dyspnoea).
evidence(tuberculosis).
query(lung_cancer).
```

```
lung_cancer: 0.055798817
```

Having seen these ProbLog equivalents of earlier illustrations we may conclude that ProbLog

```
var validity = function(dist, pred) {
    var infer = Infer(function() {
        var x = sample(dist);
        var y = pred(x);
        return y; });
   return Math.exp(infer.score(true)) }

var update = function(dist, pred) { Infer(function() {
       var x = sample(dist);
       condition( pred(x) == true );
       return x; }) }
```

Figure 6.14 WebPPL functions for validity of a predicate in a distribution and for the Bayesian update of a distribution with a predicate.

closely matches the formalism of distributions, fuzzy predicates, transformations, and updates — without being explicit about these concepts. Moreover, the underlying computational mechanics is left implicit for users of ProbLog. This makes sense, for users who just wish to model probabilistic examples and compute certain outcomes. We dare to think that familiarity with channel-based semantics and updating contributes to one's ProbLog modeling and skills.

6.6 Updating in probabilistic programming

This section looks very much like the previous one exept that it does not concentrate on the language ProbLog, but on WebPPL, introduced in Section 2.9. The latter language is not a form of logic programming, but a form of functional programming, embedded in the more general (non-functional) language Javascript, see Goodman and Stuhlmüller (2014); Goodman and Tenenbaum (2016). This section focuses on updating / conditioning in WebPPL and reviews basically the same examples as for ProbLog. This should facilitate a comparison.

The first question we should ask ourselves is: what is a (fuzzy) predicate in WebPPL? There is no pre-defined class of predicates, or more generally, of observables. We could systematically introduce such a class with associated operations, but for simplicity we choose to use predicates in an *ad hoc* manner, as functions that map inputs to outputs of from `flip(r)`, where `r` is a number in the unit interval $[0, 1]$, representing the probability per input element. We recall that `flip` in WebPPL is not a distribution but a sample from a Bernoulli distribution. In this way, fuzzy predicates on finite domains can be defined by case distinctions, see for instance `evenish_predicate` in (6.22).

Once we have decided on the nature of predicates, the validity and update operations can then be defined in WebPPL as in Figure 6.14. They both use inference. In the case of validity the inference operation produces a distribution on `true` and `false`, of which we select the `true` outcome. Recall from Section 2.9 that the `score` gives the log-probability, so that we still have to apply an e-power. This approach differs from the earlier description of validity via sampling in (4.24), due to the fact that in Figure 6.14 `p(x)` is not a probability in $[0, 1]$, but a Boolean

(`true` or `false`) sampled from a Bernoulli distribution. The `update` function in Figure 6.14 does resemble the later description in (6.28).

Like in the previous section we illustrate this updating with the fair dice and the 'evenish' predicate from Example 6.1.2 (1). In WebPPL it looks as follows.

```
var dice = uniform_distribution([1,2,3,4,5,6])

var evenish_predicate = function(i) {
    return (i == 1) ? flip(2/10) :
           (i == 2) ? flip(9/10) :
           (i == 3) ? flip(1/10) :
           (i == 4) ? flip(9/10) :
           (i == 5) ? flip(1/10) : flip(8/10); }

console.log( validity(dice, evenish_predicate) )
console.log( as_kets(update(dice, evenish_predicate)) )
```
(6.22)

The last line uses the `as_kets` function defined in (2.71), in the beginning of Section 2.9; it turns a discrete WebPPL distrubtion in its ket-representation. The two write commands to the console produce:

```
0.49999999999999994
0.067|1> + 0.3|2> + 0.033|3> + 0.3|4> + 0.033|5> + 0.267|6>
```

We review in WebPPL most of the update illustrations from Section 6.2 that we reformulated in ProbLog in the previous section.

Example 6.6.1 Recall the two urns with black and white balls from Examples 6.2.3 and 6.5.1, with a fair coin deciding from which urn a ball is drawn — that turns out to be white. In WebPPL this is done as follows.

```
var two_urns = Infer (function() {
    var head = flip();
    var white = head ? flip(2/9) : flip(5/11);
    condition( white == true );
    return head })

console.log( as_kets( two_urns ) )
```

Where in ProbLog in Example 6.5.1 the conditioning (via `evidence`) and inference (via `query`) are separated, here in WebPPL they are integrated in one function. The last `console` line yields the same posterior head probability 0.328 that we saw before. The predicate transformation and backward inference that we used in Example 6.2.3 is hard to recognise in the above WebPPL formalisation.

Example 6.6.2 In Examples 6.2.6 and 6.5.2 we used a joint distribution about blood pressure and medicine usage, which we updated in one variable, both with a sharp and a fuzzy predicate, where we were interested in the outcome in the other variable.

First, it is not immediately clear how to define a joint distribution in WebPPL. We use the

```
var blood_med_support = listproduct(['high', 'low'], [0,1,2])
var blood_med = Categorical(
    {ps: [10/100, 35/100, 25/100, 5/100, 10/100, 15/100],
    vs: blood_med_support})

var sharp_update = Infer( function() {
    var bm = sample(blood_med);
    condition( bm[0] == 'high');
    return bm[1]; })

var fuzzy_update = Infer( function() {
    var bm = sample(blood_med);
    var evidence = ( bm[0] == 'high' ) ? flip(19/20) : flip(1/20);
    condition( evidence == true );
    return bm[1]; })

console.log( as_kets( sharp_update ) )
console.log( as_kets( fuzzy_update ) )
```

Figure 6.15 Updating and marginalisation in WebPPL for the medicine-blood illustration, see Example 6.6.2.

`listproduct` function from Figure 2.9 to first form a product support and then assign probabilities appropriately, see the first lines of Figure 6.15. The subsequent two function definitions implement the updates with sharp and fuzzy evidence. The latter form involves an additional predicate, defined in terms of a case distinction with different `flip`'s. The subsequent output commands produce the updated, posterior medicine distributions:

```
0.143|0> + 0.5|1> + 0.357|2>
0.136|0> + 0.486|1> + 0.378|2>
```

These (essentially) coincide with earlier results. We see that the WebPPL approach involves a single exterior `Infer`, with the relevant code inside — as discussed in Remark 2.9.1. This same pattern occurs in the other illustrations in this section.

Example 6.6.3 The WebPPL version of the three prisoner puzzle from Examples 6.2.8 and 6.5.3

```
var asia = function() {
  var smoking =    flip(0.5);
  var asia_visit = flip(0.01);
  var lung_cancer =  smoking    ? flip(0.1)  : flip(0.01);
  var tuberculosis = asia_visit ? flip(0.05) : flip(0.01);
  var bronchitis =   smoking    ? flip(0.6)  : flip(0.3);
  var lung_cancer_or_tuberculosis = lung_cancer | tuberculosis;
  var xray = lung_cancer_or_tuberculosis ? flip(0.98) : flip(0.05);
  var dyspnoea =
    bronchitis  & lung_cancer_or_tuberculosis  ? flip(0.9) :
    bronchitis  & !lung_cancer_or_tuberculosis ? flip(0.7) :
    !bronchitis & lung_cancer_or_tuberculosis  ? flip(0.8) : flip(0.1);
  return { Smoking : smoking, Asia_visit : asia_visit,
           Lung_cancer : lung_cancer, Tuberculosis : tuberculosis,
           Bronchitis : bronchitis, Xray : xray, Dyspnoea : dyspnoea } }
```

Figure 6.16 Implementation of the Asia Bayesian network from Section 6.4 in WebPPL, see Example 6.6.4.

is given below. This time we use a channel explicitly.

```
var prisoners = ['a', 'b', 'c']
var gets_pardoned = Categorical({ ps: [1/3, 1/3, 1/3], vs: prisoners })
var gets_executed_channel = function(prisoner) {
    var probs = (prisoner == 'a') ? [0, 1/2, 1/2] :
                (prisoner == 'b') ? [0, 0, 1] : [0, 1, 0];
    return Categorical({ ps: probs, vs: prisoners }) }

var prisoner_posterior = Infer (function() {
    var p = sample(gets_pardoned);
    var e = sample(gets_executed_channel(p));
    condition( e == 'b' );
    return p })

console.log( as_kets( prisoner_posterior ) )
```

The last line prints `0.333|a> + 0.667|c>`.

Example 6.6.4 We do inference in Bayesian networks, in WebPPL via the Asia example from Section 6.4, see Figure 6.16. It variable are binary, so they can be formalised via `flip`'s instead of via categorical distributions. The asia function in Figure 6.16 returns a record object, like in Figure 2.11. This allows us to consider several earlier update questions in a uniform manner,

without repeating the whole block of asia code.

```
console.log( as_kets( Infer(function() { var a = asia();
                        condition(!a.Bronchitis);
                        return a.Lung_cancer }) ) )
console.log( as_kets( Infer(function() { var a = asia();
                        condition(a.Xray);
                        return a.Smoking }) ) )
console.log( as_kets( Infer(function() { var a = asia();
                        condition(a.Dyspnoea & a.Tuberculosis);
                        return a.Lung_cancer }) ) )
```

They produce:

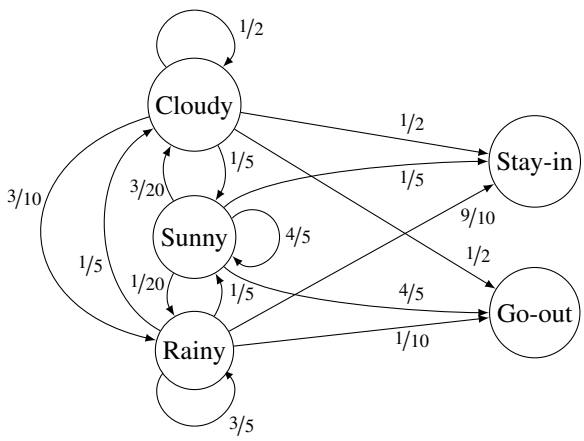
```
0.957|false> + 0.043|true>
0.312|false> + 0.688|true>
0.944|false> + 0.056|true>
```

At the end of this section we can draw similar conclusions about WebPPL, as at the end of the previous section about ProbLog: the approach matches the formalism of distributions, fuzzy predicates, transformations, and updates — without being explicit about these concepts. Since WebPPL is given essentially by a Javascript library it is relatively easy to include channels and predicates into the language, with the associated constructions that we use here.

6.7 Hidden Markov models

Hidden Markov models are much simpler than Bayesian networks: they involve only two channels, one for transitions and one for observations (abstractly called 'emissions'). The channel for transitions can be iterated multiple times, giving observations at each stage. Hidden Markov models are used in many situations. We introduce them here, with special attention for validity and updating. The Markov models are instances of the more general concept of a probabilistic automaton, studied in Chapter 10.

Let us start with a graphical description of an example of what is called a hidden Markov model.



(6.23)

This model has three 'hidden' elements, namely Cloudy, Sunny, and Rainy, representing the weather condition on a particular day. There are 'temporal' transitions with associated probabilities between these elements, as indicated by the labeled arrows. For instance, if it is cloudy today, then there is a 50% chance that it will be cloudy again tomorrow. There are also two 'visible' elements on the right: Stay-in, and Stay-out, describing two possible actions (abstractly: 'emissions') of a person, depending on the weather condition. There are transitions with probabilities from the hidden elements, on the left in (6.23), to the visible elements, on the right. The idea is that with every time step a transition is made between hidden elements, resulting in a visible outcome. Such steps may be repeated for a finite number of times — or even forever. The interaction between what is hidden and what can be observed is a key element of hidden Markov models. For instance, one may ask: given a certain initial position, what is the probability of seeing a consecutive sequence of the four visible elements: Stay-in, Stay-in, Go-out, Stay-in?

Hidden Markov models are simple statistical models that have many applications in temporal pattern recognition, in speech, handwriting or gestures Rabiner (1989); Jurafsky and Martin (2018); Manning and Schütze (2003), but also in robotics Thrun et al. (2005) and in biological sequences Eddy (2004); Yoon (2009) or wheather forecasts Khiatani and Ghose (2017). Markov models are used in situations where one has successor probabilities. For instance, in Figure 2.1, we have seen the distribution of letter frequencies in English. One can form a Markov model that assigns to each letter the distribution of successor letters (in English). This can be useful for generating, recognising or correcting words. Of course, one can also have a Markov model acting on words and providing distributions of successor words.

This section will briefly look into hidden Markov models, using the notation and terminology of channels and string diagrams, with special emphasis on validity and updating. Indeed, a hidden Markov model can be defined easily in terms of channels and forward and backward transformation of states and observables. In addition, conditioning of states by observables can be used to formulate and answer elementary questions about hidden Markov models. Learning for Markov models will be described separately in Sections 8.8 and 8.9. Markov models are examples of probabilistic automata. Such automata will be studied separately, in Chapter 10.

Definition 6.7.1 A *Markov model* (or a *Markov chain*) is given by a set X of 'internal positions', typically finite, with a 'transition' channel $t\colon X \rightsquigarrow X$ and an initial state / distribution $\sigma \in \mathcal{D}(X)$.

A *hidden Markov model*, often abbreviated as HMM, is a Markov model, as just described, with an additional 'emission' channel $e\colon X \rightsquigarrow Y$, where Y is a set of 'outputs'.

In the above illustration (6.23) we have as sets of positions and outputs:

$$X = \{\text{Cloudy}, \text{Sunny}, \text{Rainy}\} \qquad Y = \{\text{Stay-in}, \text{Go-out}\},$$

with transition channel $t\colon X \rightsquigarrow X$,

$$\begin{aligned} t(\text{Cloudy}) &= \tfrac{1}{2}\big|\text{Cloudy}\rangle + \tfrac{1}{5}\big|\text{Sunny}\rangle + \tfrac{3}{10}\big|\text{Rainy}\rangle \\ t(\text{Sunny}) &= \tfrac{3}{20}\big|\text{Cloudy}\rangle + \tfrac{4}{5}\big|\text{Sunny}\rangle + \tfrac{1}{20}\big|\text{Rainy}\rangle \\ t(\text{Rainy}) &= \tfrac{1}{5}\big|\text{Cloudy}\rangle + \tfrac{1}{5}\big|\text{Sunny}\rangle + \tfrac{3}{5}\big|\text{Rainy}\rangle, \end{aligned}$$

and emission channel $e\colon X \rightsquigarrow Y$,

$$\begin{aligned} e(\text{Cloudy}) &= \tfrac{1}{2}\big|\text{Stay-in}\big\rangle + \tfrac{1}{2}\big|\text{Go-out}\big\rangle \\ e(\text{Sunny}) &= \tfrac{1}{5}\big|\text{Stay-in}\big\rangle + \tfrac{4}{5}\big|\text{Go-out}\big\rangle \\ e(\text{Rainy}) &= \tfrac{9}{10}\big|\text{Stay-in}\big\rangle + \tfrac{1}{10}\big|\text{Go-out}\big\rangle. \end{aligned}$$

An initial state is missing in the picture (6.23).

In the literature on Markov models, the elements of the set X are often called states. This clashes with the terminology in this book, since we use 'state' as synonym for 'distribution'. So, here we call $\sigma \in \mathcal{D}(X)$ an (initial) state / distribution, and we call elements of X (internal) positions. At the same time we may call X the sample space. The transition channel $t\colon X \rightsquigarrow X$ is an endo-channel on X, that is, a channel from the space X to itself. As a function, it is of the form $t\colon X \to \mathcal{D}(X)$; it is an instance of a *coalgebra*, that is, a map of the form $A \to F(A)$ for a functor F, see Section 10.4 for more information. A HMM is an instance of a coalgebra, of the form $X \longrightarrow \mathcal{D}(X) \times \mathcal{D}(Y)$.

In a Markov chain / model one can iteratitively compute successor distributions. For an initial distribution σ and transition channel t one can form successor states via state transformation:

$$\begin{array}{l} \sigma \\ t_*(\sigma) \\ t_*(t_*(\sigma)) \;=\; (t \circ t)_*(\sigma) \;=\; (t^2)_*(\sigma) \qquad \text{where} \qquad t^n := \begin{cases} \mathit{unit} & \text{if } n = 0 \\ t \circ t^{n-1} & \text{if } n > 0. \end{cases} \\ \quad\vdots \\ (t^n)_*(\sigma) \end{array}$$

In these transitions the state at stage $n+1$ depends only on the state at stage n: in order to predict a future step, all we need is the immediate predecessor state. This makes HMMs relatively easy dynamical models. Multi-stage dependencies can be handled as well, by enlarging the sample space, see Exercise 6.7.6 below.

One interesting problem in the area of Markov chains is to find a 'stationary' state σ_∞ with $t_*(\sigma_\infty) = \sigma_\infty$, see Exercise 6.7.2 for an illustration, and also Exercise 2.4.22 for a sufficient condition. As an aside, an observable q with $c^*(q) = q$ is called a harmonic function.

Here we are more interested in hidden Markov models $\mathbf{1} \overset{\sigma}{\rightsquigarrow} X \overset{t}{\rightsquigarrow} X \overset{e}{\rightsquigarrow} Y$. The elements of the set Y are observable — and hence sometimes called *signals* — whereas the elements of X are hidden. Thus, many questions related to hidden Markov models concentrate on what one can learn about X via Y, in a finite number of steps. Hidden Markov models are examples of models with *latent* variables.

We briefly discuss some basic issues related to HMMs in separate subsections. A recurring theme is the relationship between 'parallel' and 'sequential' formulations.

6.7.1 Validity in hidden Markov models

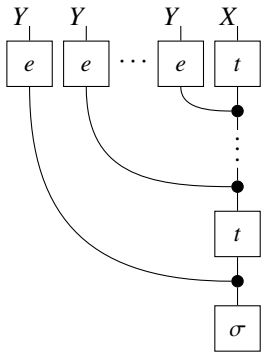
The first question that we like to address is: given a sequence of observables, what is their probability (validity) in a HMM? Standardly in the literature, one only looks at the probability of a sequence of point observations (elements), but here we use a more general approach. After all, one may not be certain about observing a specific point at a particular position, or some point observations may be missing; in the latter case one may wish to replace them by a constant (uniform) observation.

We proceed by defining validity of a sequence of observables in a joint state first; subsequently we look at (standard) algorithms for computing these validities efficiently. We thus start by defining a relevant joint state.

We fix a HMM $\mathbf{1} \overset{\sigma}{\rightsquigarrow} X \overset{t}{\rightsquigarrow} X \overset{e}{\rightsquigarrow} Y$. For each $n \in \mathbb{N}$ a channel $\langle e, t\rangle_n \colon X \rightsquigarrow Y^n \times X$ is defined in the following manner:

$$\begin{aligned} \langle e, t\rangle_0 &:= \Big(X \xrightarrow{id} X \cong \mathbf{1} \times X \cong Y^0 \times X\Big) \\ \langle e, t\rangle_{n+1} &:= \Big(X \xrightarrow{\langle e,t\rangle_n} Y^n \times X \xrightarrow{id^n \otimes \langle e,t\rangle} Y^n \times (Y \times X) \cong Y^{n+1} \times X\Big). \end{aligned} \tag{6.24}$$

We recall that the tuple $\langle e, t\rangle$ of channels is $(e \otimes t) \circ \Delta$, see Definition 2.4.4 (3). With these tuples we can form a joint state $(\langle e, t\rangle_n)_*(\sigma) \in \mathcal{D}(Y^n \times X)$. As a string diagram it looks as follows.



(6.25)

We consider the combined likelihood of a sequence of observables on the set Y in a hidden Markov model. In the literature these observables are typically point predicates $\mathbf{1}_y \colon Y \to [0, 1]$, for $y \in Y$, but, as mentioned, here we allow more general observables $Y \to \mathbb{R}$.

Definition 6.7.2 Let $\mathcal{H} = (\mathbf{1} \overset{\sigma}{\rightsquigarrow} X \overset{t}{\rightsquigarrow} X \overset{e}{\rightsquigarrow} Y)$ be a hidden Markov model and let $\vec{p} = p_1, \dots, p_n$ be a list of observables on Y. The validity $\mathcal{H} \models \vec{p}$ of this sequence $\vec{p}$ in the model $\mathcal{H}$ is defined via the tuples (6.24) as:

$$\begin{aligned} \mathcal{H} \models \vec{p} &:= (\langle e, t\rangle_n)_*(\sigma)[1, \dots, 1, 0] \models p_1 \otimes \cdots \otimes p_n \\ &\overset{(4.11)}{=} (\langle e, t\rangle_n)_*(\sigma) \models p_1 \otimes \cdots \otimes p_n \otimes \mathbf{1} \\ &\overset{(4.15)}{=} \sigma \models (\langle e, t\rangle_n)^*(p_1 \otimes \cdots \otimes p_n \otimes \mathbf{1}). \end{aligned} \tag{6.26}$$

The marginalisation mask $[1, \dots, 1, 0]$ contains n times the number 1. It ensures that the X outcome in (6.25) is discarded.

We describe an alternative way to formulate this validity without using the (big) joint state on $Y^n \times X$. It forms the essence of the classical 'forward' and 'backward' algorithms for validity in HMMs, see *e.g.* Bishop (2006); Ross (2007) or (Jurafsky and Martin, 2018, App. A). An alternative algorithm is described in Exercise 6.7.4.

Proposition 6.7.3 *The HMM-validity* (6.26) *can be computed as:*

$$\begin{aligned} \mathcal{H} \models \vec{p} \;=\; \sigma \models\; & e^*(p_1) \;\&\; \\ & t^*\big(\, e^*(p_2) \;\&\; \\ & \quad t^*\big(\, e^*(p_3) \;\&\; \cdots \\ & \qquad t^*\big(\, e^*(p_n) \;\&\; \cdots\big)\big)\big). \end{aligned} \tag{6.27}$$

This validity can be calculated recursively in forward manner as:

$$\sigma \models \alpha(\vec{p}) \qquad \textit{where} \qquad \begin{cases} \alpha([q]) \;=\; e^*(q) \\ \alpha([q_1] +\!\!+ \vec{q}) \;=\; e^*(q_1) \mathbin{\&} t^*(\alpha(\vec{q})). \end{cases}$$

Alternatively, this validity can be calculated recursively in backward manner as:

$$\sigma \models \beta(\vec{p}, \mathbf{1}) \qquad \textit{where} \qquad \begin{cases} \beta([q]) \;=\; q \\ \beta(\vec{q} +\!\!+ [q_n, q_{n+1}]) \;=\; \beta(\vec{q} +\!\!+ [e^*(q_n) \mathbin{\&} t^*(q_{n+1})]). \end{cases}$$

Proof. We first prove, by induction on $n \geq 1$ that for observables p_i on Y and q on X one has:

$$(\langle e, t\rangle_n)^*(p_1 \otimes \cdots \otimes p_n \otimes q) \;=\; e^*(p_1) \mathbin{\&} t^*(e^*(p_2) \mathbin{\&} t^*(\cdots t^*(e^*(p_n) \mathbin{\&} t^*(q)) \cdots)) \qquad (*)$$

The base case $n = 1$ is easy by Lemma 4.5.2 (7):

$$(\langle e, t\rangle_1)^*(p_1 \otimes q) \;=\; e^*(p_1) \mathbin{\&} t^*(q).$$

For the induction step we reason as follows.

$$\begin{aligned} &(\langle e, t\rangle_{n+1})^*(p_1 \otimes \cdots \otimes p_n \otimes p_{n+1} \otimes q) \\ &\;=\; \langle e, t\rangle_n{}^*((id^n \otimes \langle t, e\rangle)^*(p_1 \otimes \cdots \otimes p_n \otimes p_{n+1} \otimes q)) \\ &\;=\; (\langle e, t\rangle_n)^*(p_1 \otimes \cdots \otimes p_n \otimes \langle e, t\rangle^*(p_{n+1} \otimes q)) \\ &\;=\; (\langle e, t\rangle_n)^*(p_1 \otimes \cdots \otimes p_n \otimes (e^*(p_{n+1}) \mathbin{\&} t^*(q))) \qquad \text{as just shown} \\ &\overset{\text{(IH)}}{=}\; e^*(p_1) \mathbin{\&} t^*(e^*(p_2) \mathbin{\&} t^*(\cdots t^*(e^*(p_n) \mathbin{\&} t^*(e^*(p_{n+1}) \mathbin{\&} t^*(q))) \cdots)). \end{aligned}$$

We can now prove Equation (6.27):

$$\begin{aligned} \mathcal{H} \models \vec{p} &\overset{(6.26)}{=} \sigma \models (\langle e, t\rangle_n)^*((p_1 \otimes \cdots \otimes p_n \otimes \mathbf{1})) \\ &\overset{(*)}{=}\; e^*(p_1) \mathbin{\&} t^*(e^*(p_2) \mathbin{\&} t^*(\cdots t^*(e^*(p_n) \mathbin{\&} t^*(\mathbf{1}) \cdots))) \\ &=\; e^*(p_1) \mathbin{\&} t^*(e^*(p_2) \mathbin{\&} t^*(\cdots t^*(e^*(p_n) \cdots))). \end{aligned}$$

❑

6.7.2 Filtering and prediction

Given a sequence $\vec{p}$ of factors, one can compute their validity $\mathcal{H} \models \vec{p}$ in a HMM $\mathcal{H}$, as described above. One can also use these factors to 'guide' the evolution of the HMM. At each state i the factor p_i is used to update the current state, via backward inference. The new state is then moved forward via the transition function. This process is called *filtering*, after the Kalman filter from the 1960s that is used for instance in trajectory optimisation in navigation and in rocket control (*e.g.* for the Apollo program). The system can evolve autonomously via its transition function, but observations at regular intervals can update (correct) the current state.

Definition 6.7.4 Let $\mathcal{H} = (\mathbf{1} \overset{\sigma}{\rightsquigarrow} X \overset{t}{\rightsquigarrow} X \overset{e}{\rightsquigarrow} Y)$ be a hidden Markov model and let $\vec{p} = p_1, \ldots, p_n$ be a list of factors on Y.

1 The *filtered* sequence of distributions $\sigma_1, \ldots, \sigma_{n+1} \in \mathcal{D}(X)$ is obtained via the following the observe-update-proceed principle:

$$\sigma_1 \;:=\; \sigma \qquad \text{and} \qquad \sigma_{i+1} \;:=\; t_*\big(\sigma_i|_{e^*(p_i)}\big).$$

2 The (visible) *prediction* after $\vec{p}$ is the distribution:

$$e_*(\sigma_{n+1}) \;\in\; \mathcal{D}(Y).$$

In the terminology of Definition 6.2.1, the definition of the state σ_{i+1} involves both forward and backward inference. Below we show that the final state σ_{n+1} in the filtered sequence can also be obtained via crossover inference on a joint state, obtained via the tuple channels (6.24). This fact gives a theoretical justification, but is of little practical relevance — since joint states quickly become too big to handle.

Proposition 6.7.5 *In the context of Definition 6.7.4,*

$$\sigma_{n+1} \;=\; (\langle e,t\rangle_n)_*(\sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}}[0,\ldots,0,1]$$

The marginalisation mask $[0,\ldots,0,1]$ *has n zero's.*

Proof. By induction on $n \geq 1$. The base case with $\langle e,t\rangle_1 = \langle e,t\rangle$ is handled as follows.

$$\begin{aligned}
\langle e,t\rangle_*(\sigma)\big|_{p_1\otimes\mathbf{1}}[0,1] &= (\pi_2)_*\Big(\langle e|_{p_1},t\rangle_*\big((\sigma|_{\langle e,t\rangle^*(p_1\otimes\mathbf{1})})\big)\Big) && \text{by Corollary 6.3.5 (2)}\\
&= t_*(\sigma|_{e^*(p_1)}) && \text{by Lemma 4.5.2 (7)}\\
&= \sigma_2.
\end{aligned}$$

The induction step requires a bit more work:

$$\begin{aligned}
&(\langle e,t\rangle_{n+1})_*(\sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes p_{n+1}\otimes\mathbf{1}}[0,\ldots,0,0,1]\\
&= (\pi_{n+2})_*\Big((id^n\otimes\langle e,t\rangle)_*(\langle e,t\rangle_n)_*(\sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes p_{n+1}\otimes\mathbf{1}}\Big)\\
&\overset{(6.11)}{=} (\pi_{n+2})_*\Big(\big((id^n\otimes\langle e,t\rangle)\big|_{p_1\otimes\cdots\otimes p_n\otimes p_{n+1}\otimes\mathbf{1}}\big)_*\Big(\big(\langle e,t\rangle_n)_*(\sigma)\big|_{(id^n\otimes\langle e,t\rangle)^*(p_1\otimes\cdots\otimes p_n\otimes p_{n+1}\otimes\mathbf{1})}\Big)\Big)\\
&= (\pi_{n+2})_*\Big(\big(id^n\otimes\langle e|_{p_{n+1}},t\rangle\big)_*\Big(\big(\langle e,t\rangle_n)_*(\sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes(e^*(p_{n+1})\,\&\,t^*(\mathbf{1}))}\Big)\Big)\\
&= (t\circ\pi_{n+1})_*\Big(\big(\langle e,t\rangle_n)_*(\sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes e^*(p_{n+1})}\Big)\\
&= t_*\Big(\big(\langle e,t\rangle_n)_*(\sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}}\big|_{\mathbf{1}\otimes\cdots\otimes\mathbf{1}\otimes e^*(p_{n+1})}[0,\ldots,0,1]\Big)\\
&\qquad\text{by Lemma 6.1.7 (3) and Lemma 4.4.9 (1)}\\
&= t_*\Big(\big(\langle e,t\rangle_n)_*(\sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}}[0,\ldots,0,1]\big|_{e^*(p_{n+1})}\Big)\\
&\qquad\text{by Lemma 6.1.7 (6)}\\
&\overset{(IH)}{=} t_*\big(\sigma_{n+1}\big|_{e^*(p_{n+1})}\big) \;=\; \sigma_{n+2}.
\end{aligned}$$

❑

The next consequence of the previous proposition may be understood as Bayes' rule for HMMs — or more accurately, the product rule for HMMs, see Proposition 6.1.4.

Corollary 6.7.6 *Still in the context of Definition 6.7.4, let q be a predicate on X. Its validity in the final state in the sequence* $\sigma_1,\ldots,\sigma_{n+1}$*, filtered by factors* $p_1,\ldots,p_n$*, is given by:*

$$\sigma_{n+1}\models q \;=\; \frac{(\langle e,t\rangle_n)_*(\sigma)\models p_1\otimes\cdots\otimes p_n\otimes q}{\mathcal{H}\models\vec{p}}.$$

Proof. This follows from Bayes' rule, in Proposition 6.1.4 (1):

$$\begin{aligned}
\sigma_{n+1}\models q &= (\langle e,t\rangle_n)_*(\sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}}[0,\ldots,0,1]\models q\\
&= (\langle e,t\rangle_n)_*(\sigma)\big|_{p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}}\models\mathbf{1}\otimes\cdots\otimes\mathbf{1}\otimes q && \text{by (4.11)}\\
&= \frac{(\langle e,t\rangle_n)_*(\sigma)\models(p_1\otimes\cdots\otimes p_n\otimes\mathbf{1})\;\&\;(\mathbf{1}\otimes\cdots\otimes\mathbf{1}\otimes q)}{(\langle e,t\rangle_n)_*(\sigma)\models p_1\otimes\cdots\otimes p_n\otimes\mathbf{1}}\\
&= \frac{(\langle e,t\rangle_n)_*(\sigma)\models p_1\otimes\cdots\otimes p_n\otimes q}{\mathcal{H}\models\vec{p}}.
\end{aligned}$$

The last equation uses Lemma 4.4.9 (1) and Definition 6.7.2. ❑

6.7.3 Metropolis-Hastings

A question with practical relevance is the following. Suppose we have a non-empty multiset $\varphi \in \mathcal{M}(X)$, so with real-valued multiplicities, how can we sample from its normalisation $\mathit{Flrn}(\varphi) = \frac{1}{\|\varphi\|} \cdot \varphi \in \mathcal{D}(X)$, without computing the size $\|\varphi\|$? When φ is a distribution with large support, calculating this sum $\|\varphi\| = \sum_x \varphi(x)$ may be undoable.

The Metropolis-Hastings algorithm from the early 1950s provides an answer. It shows how to turn φ into a Markov chain $c\colon X \rightsquigarrow X$ for which $\mathit{Flrn}(\varphi)$ is a stationary distribution / state: $c_*(\mathit{Flrn}(\varphi)) = \mathit{Flrn}(\varphi)$, that is, a fixed point of state transformation along the channel c. Now we can form a sequence of elements from X by starting with an arbitrary point $x_0 \in X$, and sample successively $x_{n+1} \leftarrow c(x_n)$. One obtains a sequence $x_0, \dots, x_N$ whose accumulation $\mathit{acc}(x_0, \dots, x_N) \in \mathcal{D}(X)$ approaches $\mathit{Flrn}(\varphi)$, as N increases, see *e.g.* Chib and Greenberg (1995); Tierney (1994) for more information.

We sketch how to obain the channel $c\colon X \rightsquigarrow X$, from a multiset $\varphi \in \mathcal{M}(X)$. We assume that φ has full support, *i.e.* that $X = \mathit{supp}(\varphi)$ and thus that the set X is finite. We describe a simplified version; the construction can be generalised, by factoring in an arbitrary 'candidate' channel $r\colon X \rightsquigarrow X$.

One first defines a function $\alpha\colon X \to \mathcal{M}(X)$ as:

$$\alpha(x) \coloneqq \sum_{y \in X} \min\left(1, \frac{\varphi(y)}{\varphi(x)}\right) \big|y\rangle.$$

The channel $c\colon X \rightsquigarrow X$ is then defined as:

$$c(x)(y) \coloneqq \begin{cases} \dfrac{\alpha(x)(y)}{|X|} & \text{if } x \neq y \\ 1 - \displaystyle\sum \Big\{ \frac{\alpha(x)(z)}{|X|} \;\Big|\; z \in X,\ z \neq x \Big\} & \text{if } x = y. \end{cases}$$

The second clause ensures that each $c(x)$ is a distribution,

We need to show is that $c_*(\mathit{Flrn}(\varphi)) = \mathit{Flrn}(\varphi)$. We give an exemplaric proof for $X = \{u, v, w\}$ with the multiset $\varphi \in \mathcal{M}(X)$ satisfying $\varphi(u) \leq \varphi(w) \leq \varphi(v)$. Then:

$$\begin{aligned} \alpha(u) &= 1\big|u\rangle + 1\big|v\rangle + 1\big|w\rangle \\ \alpha(v) &= \frac{\varphi(u)}{\varphi(v)}\big|u\rangle + 1\big|v\rangle + \frac{\varphi(w)}{\varphi(v)}\big|w\rangle \\ \alpha(w) &= \frac{\varphi(u)}{\varphi(w)}\big|u\rangle + 1\big|v\rangle + 1\big|w\rangle. \end{aligned}$$

And:

$$\begin{aligned} c(u) &= \frac{1}{3}\big|u\rangle + \frac{1}{3}\big|v\rangle + \frac{1}{3}\big|w\rangle \\ c(v) &= \frac{\varphi(u)}{3\varphi(v)}\big|u\rangle + \left(1 - \frac{\varphi(u)}{3\varphi(v)} - \frac{\varphi(w)}{3\varphi(v)}\right)\big|v\rangle + \frac{\varphi(w)}{3\varphi(v)}\big|w\rangle \\ c(w) &= \frac{\varphi(u)}{3\varphi(w)}\big|u\rangle + \frac{1}{3}\big|v\rangle + \left(1 - \frac{\varphi(u)}{3\varphi(w)} - \frac{1}{3}\right)\big|w\rangle. \end{aligned}$$

We check that $Flrn(\varphi)$ is a stationary distributin.

$$\begin{aligned} c_*(Flrn(\varphi))(u) &= \frac{1}{\|\varphi\|} \cdot \sum_{x \in X} \varphi(x) \cdot c(x)(u) = \frac{1}{\|\varphi\|} \cdot \left(\varphi(u) \cdot \frac{1}{3} + \varphi(v) \cdot \frac{\varphi(u)}{3\varphi(v)} + \varphi(w) \cdot \frac{\varphi(u)}{3\varphi(w)} \right) \\ &= \frac{1}{\|\varphi\|} \cdot \varphi(u) = Flrn(\varphi)(u). \end{aligned}$$

Similarly,

$$\begin{aligned} c_*(Flrn(\varphi))(v) &= \frac{1}{\|\varphi\|} \cdot \left(\varphi(u) \cdot \frac{1}{3} + \varphi(v) \cdot \left(1 - \frac{\varphi(u)}{3\varphi(v)} - \frac{\varphi(w)}{3\varphi(v)} \right) + \varphi(w) \cdot \frac{1}{3} \right) \\ &= \frac{1}{\|\varphi\|} \cdot \varphi(v) = Flrn(\varphi)(v) \\ c_*(Flrn(\varphi))(w) &= \frac{1}{\|\varphi\|} \cdot \left(\varphi(u) \cdot \frac{1}{3} + \varphi(v) \cdot \frac{\varphi(w)}{3\varphi(v)} + \varphi(w) \cdot \left(1 - \frac{\varphi(u)}{3\varphi(w)} - \frac{1}{3} \right) \right) \\ &= \frac{1}{\|\varphi\|} \cdot \varphi(w) = Flrn(\varphi)(w). \end{aligned}$$

The proof may be generalised to arbitrary multisets, but the many case distinctions easily obfuscate what is going on.

Exercises

6.7.1 Consider the HMM example 6.23 with initial state $\sigma = 1|\text{Cloudy}\rangle$.

1. Compute successive states $(t^n)_*(\sigma)$ for $n = 0, 1, 2, 3$.
2. Compute successive observations $e_*((t^n)_*(\sigma))$ for $n = 0, 1, 2, 3$.
3. Check that the validity of the sequence of (point-predicate) observations Go-out, Stay-in, Stay-in is $\frac{837}{5000} = 0.1674$.
4. Show that filtering, as in Definition 6.7.4, with these same three (point) observations yields as final outcome:

$$\begin{aligned} &\tfrac{1867}{6696}|\text{Cloudy}\rangle + \tfrac{347}{1395}|\text{Sunny}\rangle + \tfrac{15817}{33480}|\text{Rainy}\rangle \\ &\approx 0.279|\text{Cloudy}\rangle + 0.249|\text{Sunny}\rangle + 0.472|\text{Rainy}\rangle. \end{aligned}$$

6.7.2 Consider the transition channel t associated with the HMM example 6.23. Check that in order to find a stationary distribution $\sigma_\infty = x|\text{Cloudy}\rangle + y|\text{Sunny}\rangle + z|\text{Rainy}\rangle$ one has to solve the equations:

$$x = \tfrac{1}{2}x + \tfrac{3}{20}y + \tfrac{1}{5}z \qquad y = \tfrac{1}{5}x + \tfrac{4}{5}y + \tfrac{1}{5}z \qquad z = \tfrac{3}{10}x + \tfrac{1}{20}y + \tfrac{3}{5}z$$

Deduce that $\sigma_\infty = \frac{1}{4}|\text{Cloudy}\rangle + \frac{1}{2}|\text{Sunny}\rangle + \frac{1}{4}|\text{Rainy}\rangle$ and double-check that $t_*(\sigma_\infty) = \sigma_\infty$.

6.7.3 (The set-up of this exercise is copied from machine learning lecture notes of Doina Precup.) Consider a 5-state hallway of the form:

1	2	3	4	5

Thus we use a space $X = \{1, 2, 3, 4, 5\}$ of positions, together with a space $Y = \{2, 3\}$

of outputs, for the number of surrounding walls. The transition and emission channels $t\colon X \rightsquigarrow X$ and $e\colon X \rightsquigarrow Y$ for a robot in this hallway are given by:

$$\begin{aligned} t(1) &= \tfrac{3}{4}|1\rangle + \tfrac{1}{4}|2\rangle & e(1) &= 1|3\rangle \\ t(2) &= \tfrac{1}{4}|1\rangle + \tfrac{1}{2}|2\rangle + \tfrac{1}{4}|3\rangle & e(2) &= 1|2\rangle \\ t(3) &= \tfrac{1}{4}|2\rangle + \tfrac{1}{2}|3\rangle + \tfrac{1}{4}|4\rangle & e(3) &= 1|2\rangle \\ t(4) &= \tfrac{1}{4}|3\rangle + \tfrac{1}{2}|3\rangle + \tfrac{1}{4}|5\rangle & e(4) &= 1|2\rangle \\ t(5) &= \tfrac{1}{4}|4\rangle + \tfrac{3}{4}|5\rangle & e(5) &= 1|3\rangle. \end{aligned}$$

We use $\sigma = 1|3\rangle$ as start state, and we have a sequence of observations $\alpha = [2, 2, 3, 2, 3, 3]$, formally as a sequence of point predicates $[\mathbf{1}_2, \mathbf{1}_2, \mathbf{1}_3, \mathbf{1}_2, \mathbf{1}_3, \mathbf{1}_3]$.

1 Check that $(\sigma, t, e) \models \alpha = \frac{3}{512}$.

2 Next we filter with the sequence α. Show that it leads succesively to the following states σ_i as in Definition 6.7.4

$$\begin{aligned} \sigma_1 \;:=\; \sigma &= 1|3\rangle \\ \sigma_2 &= \tfrac{1}{4}|2\rangle + \tfrac{1}{2}|3\rangle + \tfrac{1}{4}|4\rangle \\ \sigma_3 &= \tfrac{1}{16}|1\rangle + \tfrac{1}{4}|2\rangle + \tfrac{3}{8}|3\rangle + \tfrac{1}{4}|4\rangle + \tfrac{1}{16}|5\rangle \\ \sigma_4 &= \tfrac{3}{8}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|4\rangle + \tfrac{3}{8}|5\rangle \\ \sigma_5 &= \tfrac{1}{8}|1\rangle + \tfrac{1}{4}|2\rangle + \tfrac{1}{4}|3\rangle + \tfrac{1}{4}|4\rangle + \tfrac{1}{8}|5\rangle \\ \sigma_6 &= \tfrac{3}{8}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|4\rangle + \tfrac{3}{8}|5\rangle \\ \sigma_7 &= \tfrac{3}{8}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|4\rangle + \tfrac{3}{8}|5\rangle. \end{aligned}$$

6.7.4 Apply Bayes' rule to the validity formulation (6.27) in order to prove the correctness of the following HMM validity algorithm.

$$\begin{aligned} (\sigma, t, e) \models [] \;&:=\; 1 \\ (\sigma, t, e) \models [p_1] +\!\!+ \vec{p} \;&:=\; \big(\sigma \models e^*(p_1)\big) \cdot \Big(\big(t_*(\sigma|_{e^*(p_1)}), t, e\big) \models \vec{p}\Big). \end{aligned}$$

(Notice the connection with filtering from Definition 6.7.4.)

6.7.5 A *random walk* is a Markov model $d\colon \mathbb{Z} \rightsquigarrow \mathbb{Z}$ given by $d(n) = r|n-1\rangle + (1-r)|n+1\rangle$ for some $r \in [0, 1]$. This captures the idea that a step-to-the-left or a step-to-the-right are the only possible transitions. (The letter 'd' hints at modeling a drunkard.)

1 Start from initial state $\sigma = 1|0\rangle \in \mathcal{D}(\mathbb{Z})$ and describe a couple of subsequent states $d_*(\sigma)$, $(d^2)_*(\sigma)$, $(d^3)_*(\sigma)$, ... Which pattern emerges?

2 Prove that for $K \in \mathbb{N}$,

$$\begin{aligned} (d^K)_*(\sigma) &= \sum_{0 \le k \le K} bn[K](1-r)(k)\,\big|2k - K\rangle \\ &= \sum_{0 \le k \le K} \binom{K}{k} \cdot (1-r)^k \cdot r^{K-k}\,\big|2k - K\rangle \end{aligned}$$

6.7.6 A Markov chain $X \rightsquigarrow X$ has a 'one-stage history' only, in the sense that the state at stage $n+1$ depends only on the state at stage n. The following situation from (Ross, 2007, Chap. III, Ex. 4.4) involves a two-stage history.

Suppose that whether or not it rains today depends on wheather conditions through the last two days. Specifically, suppose that if it has rained for the past two days, then it will rain tomorrow with probability 0.7; if it rained today but not yesterday, then it will rain tomorrow with probability 0.5; if

it rained yesterday but not today, then it will rain tomorrow with probability 0.4; if it has not rained in the past to days, then it will rain tomorrow with probability 0.2.

1 Write $R = \{r, r^{\perp}\}$ for the state space of rain and no-rain outcomes, and capture the above probabilities via a channel $c\colon R \times R \rightsquigarrow R$.

2 Turn this channel c into a Markov chain $\langle \pi_2, c\rangle\colon R{\times}R \rightsquigarrow R{\times}R$, where the second component of $R{\times}R$ describes whether or not it rains on the current day, and the first component on the previous day. Describe $\langle \pi_2, c\rangle$ both as a function and as a string diagram.

3 Generalise this approach to a history of length $N > 1$: turn a channel $X^N \rightsquigarrow X$ into a Markov model $X^N \rightsquigarrow X^N$, where the relevant history is incorporated into the sample space.

6.7.7 Use the approach of the prevous exercise to turn a hidden Markov model into a Markov model.

6.7.8 Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two HMMs. Define their parallel product $\mathcal{H}_1 \otimes \mathcal{H}_2$ using the tensor operation $\otimes$ on states and channels.

6.7.9 Consider the multiset $\varphi = 5|1\rangle + 4|2\rangle + 3|3\rangle + 2|4\rangle + 1|5\rangle \in \mathcal{M}(X)$ for $X = \{1, 2, 3, 4, 5\}$. Use the Metropolis-Hastings method from Subsection 6.7.3 to obtain a channel $c\colon X \rightsquigarrow X$ with $c_*(Flrn(\varphi)) = Flrn(\varphi)$.

6.8 Updating draw distributions

In this section we apply updating to distributions that are obtained by drawing from an urn, as described in Chapter 3. We first show that updating commutes with multinomial channels. The proof is not particularly difficult, but the result is noteworthy because it shows how well basic probabilistic notions are integrated.

Recall from Definition 4.6.6 that $\overline{p}^{\bullet}\colon \mathcal{N}(X) \to \mathbb{R}$ is the multiplicative extension of an observable $p\colon X \to \mathbb{R}$, see, defined by $\overline{p}^{\bullet}(\varphi) = \prod_{x \in X} p(x)^{\varphi(x)}$. Proposition 4.6.7 (2) we already saw that the validity $mn[K](\omega) \models \overline{p}^{\bullet}$ in a multinomial distribution equals the K-power $(\omega \models p)^K$ of the validity in the original distribution. We now extend this result to updating.

Theorem 6.8.1 *Let $\omega \in \mathcal{D}(X)$ be given. For a factor (or predicate) $p\colon X \to \mathbb{R}_{\geq 0}$ and a number $K \in \mathbb{N}$, one has the following results about updating with the multiplicative extension $\overline{p}^{\bullet}\colon \mathcal{N}(X) \to \mathbb{R}_{\geq 0}$.*

1 By updating a multinomial distribution with the multiplicative extension $\overline{p}^{\bullet}$ one still has a multinomial distribution, with an updated urn distribution:

$$mn[K](\omega)\big|_{\overline{p}^{\bullet}} = mn[K](\omega|_p).$$

2 Similarly, for the parallel multinomial law pml *from Section 3.7,*

$$pml(\textstyle\sum_i n_i|\omega_i\rangle)\big|_{\overline{p}^{\bullet}} = pml(\textstyle\sum_i n_i|\,\omega_i|_p\rangle).$$

3 Updating a Poisson point process does not only affect the distribution ω but also the rate λ:

$$Pmn[\lambda](\omega)\big|_{\overline{p}^{\bullet}} = Pmn[\lambda \cdot (\omega \models p)](\omega|_p).$$

Proof. 1 Via the validity formula of Proposition 4.6.7 (2), for $\varphi \in \mathcal{N}[K](X)$,

$$\begin{aligned}
mn[K](\omega)\big|_{\overline{p}^{\bullet}}(\varphi) &= \frac{mn[K](\omega)(\varphi)\cdot \overline{p}^{\bullet}(\varphi)}{mn[K](\omega)\models \overline{p}^{\bullet}} = (\varphi)\cdot \frac{\big(\prod_x \omega(x)^{\varphi(x)}\big)\cdot\big(\prod_x p(x)^{\varphi(x)}\big)}{(\omega\models p)^K} \\
&= (\varphi)\cdot \prod_x \left(\frac{\omega(x)\cdot p(x)}{\omega\models p}\right)^{\varphi(x)} \qquad \text{since } \|\varphi\| = K \\
&= (\varphi)\cdot \prod_x \omega|_p(x)^{\varphi(x)} \\
&= mn[K](\omega|_p)(\varphi).
\end{aligned}$$

2 Similarly, via Proposition 4.6.10:

$$\begin{aligned}
pml(\textstyle\sum_i n_i|\omega_i\rangle)\big|_{\overline{p}^{\bullet}} &= \sum_{\varphi\in\mathcal{N}[K](X)} \frac{pml(\sum_i n_i|\omega_i\rangle)(\varphi)\cdot \overline{p}^{\bullet}(\varphi)}{pml(\sum_i n_i|\omega_i\rangle)\models \overline{p}^{\bullet}}\,|\varphi\rangle \quad \text{where } K = \textstyle\sum_i n_i \\
&\overset{(3.36)}{=} \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \frac{\prod_i\, mn[n_i](\omega_i)(\varphi_i)\cdot \overline{p}^{\bullet}(\varphi_i)}{\prod_i (\omega_i\models p)^{n_i}}\,\big|\textstyle\sum_i \varphi_i\rangle \\
&= \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \prod_i (\varphi_i)\cdot \frac{\prod_x (\omega(x)\cdot p(x))^{\varphi_i(x)}}{(\omega_i\models p)^{n_i}}\,\big|\textstyle\sum_i \varphi_i\rangle \\
&= \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \prod_i (\varphi_i)\cdot \prod_x \left(\frac{\omega(x)\cdot p(x)}{\omega_i\models p}\right)^{\varphi_i(x)}\,\big|\textstyle\sum_i \varphi_i\rangle \\
&= \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \prod_i (\varphi_i)\cdot \prod_x \omega_i|_p(x)^{\varphi_i(x)}\,\big|\textstyle\sum_i \varphi_i\rangle \\
&= \sum_{i,\,\varphi_i\in\mathcal{N}[n_i](X)} \prod_i mn[n_i](\omega_i|_p)(\varphi_i)\,\big|\textstyle\sum_i \varphi_i\rangle \;=\; pml(\textstyle\sum_i n_i|\,\omega_i|_p\rangle).
\end{aligned}$$

3 Using Proposition 4.6.7 (2) and (4) together with the item (1), we get for an arbitrary multiset $\varphi \in \mathcal{N}(X)$, say with of size $K = \|\varphi\|$,

$$\begin{aligned}
Pmn[\lambda](\omega)\big|_{\overline{p}^{\bullet}}(\varphi) &= \frac{Pmn[\lambda](\omega)(\varphi)\cdot \overline{p}^{\bullet}(\varphi)}{Pmn[\lambda](\omega)\models \overline{p}^{\bullet}} \\
&\overset{(3.53)}{=} \frac{e^{-\lambda}\cdot\lambda^K\cdot mn[K](\varphi)\cdot \overline{p}^{\bullet}(\varphi)}{K!\cdot e^{-\lambda\cdot(\omega\models p^{\perp})}} \\
&= \frac{e^{-\lambda\cdot(1-(\omega\models p^{\perp}))}\cdot\lambda^K\cdot(\omega\models p)^K}{K!}\cdot\frac{mn[K](\varphi)\cdot \overline{p}^{\bullet}(\varphi)}{mn[K](\omega)\models \overline{p}^{\bullet}} \\
&= \frac{e^{-\lambda\cdot(\omega\models p)}\cdot(\lambda\cdot(\omega\models p))^K}{K!}\cdot mn[K](\omega|_p)(\varphi) \\
&\overset{(3.53)}{=} Pmn[\lambda\cdot(\omega\models p)](\omega|_p)(\varphi).
\end{aligned}$$

❑

Example 6.8.2 We continue Example 3.10.2, with a disease distribution $\omega = \frac{1}{6}|a\rangle + \frac{1}{8}|b\rangle + \frac{3}{8}|c\rangle + \frac{1}{3}|d\rangle$ with $\lambda = 5$ patients arriving per hour. Let's now assume that people with diseases a, b are treated separately. When we wish to consider the arrival probabilities for these diseases only, we update with the (sharp) predicate $p\colon \{a,b,c,d\} \to [0,1]$ given by $p(a) = p(b) = 1$ and $p(c) = p(d) = 0$. The validity $\omega \models p$ is $\omega(a) + \omega(b) = \frac{1}{8} + \frac{1}{6} = \frac{7}{24}$, and the updated distribution ρ is:

$$\rho \coloneqq \omega|_p = \frac{1/8}{7/24}|a\rangle + \frac{1/6}{7/24}|b\rangle = \tfrac{3}{7}|a\rangle + \tfrac{4}{7}|b\rangle.$$

The adapted rate for these two diseases is $\lambda \cdot (\omega \models p) = 5 \cdot \frac{7}{24} = \frac{35}{24}$. Thus the arrival probabilities, for multisets of these two diseases a, b only, are:

$$\begin{aligned}
Pmn[5](\omega)\big|_{\overline{p}^{\bullet}} &= Pmn[5 \cdot \tfrac{7}{24}](\omega|_p) = Pmn[\tfrac{35}{24}](\rho) \\
&= pois[\tfrac{35}{24}](0) \cdot mn[0](\rho) + pois[\tfrac{35}{24}](1) \cdot mn[1](\rho) + pois[\tfrac{35}{24}](2) \cdot mn[2](\rho) + \cdots \\
&= e^{-35/24} \cdot \Big(1\big|\mathbf{0}\big\rangle + \tfrac{5}{8}\big|1|a\rangle\big\rangle + \tfrac{5}{6}\big|1|b\rangle\big\rangle + \tfrac{25}{128}\big|2|a\rangle\big\rangle + \tfrac{25}{48}\big|1|a\rangle + 1|b\rangle\big\rangle + \tfrac{25}{72}\big|2|b\rangle\big\rangle + \cdots\Big) \\
&= 0.233\big|\mathbf{0}\big\rangle + 0.145\big|1|a\rangle\big\rangle + 0.194\big|1|b\rangle\big\rangle \\
&\qquad + 0.0454\big|2|a\rangle\big\rangle + 0.121\big|1|a\rangle + 1|b\rangle\big\rangle + 0.0808\big|2|b\rangle\big\rangle + \cdots
\end{aligned}$$

We now restrict to multiplicative extensions $\overline{\mathbf{1}_E}^{\bullet}$ of sharp predicates, which are sharp themselves, see Exercise 4.6.4. The action $\varphi \bullet p$ of a factor p on a multiset φ from Exercise 4.4.14 can be used to get an analogue of the first item in the next proposition for hypergeometric and Pólya distributions. However, this result is more restricted and works only for a *sharp* predicate p. The sharp predicate restricts both the urn and the draws from it to balls of certain colours only.

Theorem 6.8.3 *Let $\upsilon \in \mathcal{N}(X)$ be an urn and $E \subseteq X$ an event / subset, forming a sharp indicator predicate $\mathbf{1}_E \colon X \to [0,1]$.*

1 If $\big\|\upsilon \bullet \mathbf{1}_E\big\| \geq K$,

$$hg[K](\upsilon)\big|_{\overline{\mathbf{1}_E}^{\bullet}} = hg[K]\big(\upsilon \bullet \mathbf{1}_E\big).$$

2 Similarly, if $\upsilon \bullet \mathbf{1}_E$ is non-empty,

$$pl[K](\upsilon)\big|_{\overline{\mathbf{1}_E}^{\bullet}} = pl[K](\upsilon \bullet \mathbf{1}_E).$$

Proof. 1 We first recall from Exercise 4.6.4 that $\overline{\mathbf{1}_E}^{\bullet}(\varphi) = 1$ iff $supp(\varphi) \subseteq E$, for $\varphi \in \mathcal{N}(X)$. Next, let's write $L = \|\upsilon\|$ for the size of the urn and $L_E = \|\upsilon \bullet \mathbf{1}_E\|$ for the size of the restricted urn — where it is assumed that $L_E \geq K$. Then:

$$\begin{aligned}
hg[K](\upsilon) \models \overline{\mathbf{1}_E}^{\bullet} = \sum_{\varphi \leq_K \upsilon} hg[K](\upsilon)(\varphi) \cdot \overline{\mathbf{1}_E}^{\bullet}(\varphi) &= \sum_{\varphi \leq_K \upsilon,\, supp(\varphi) \subseteq E} hg[K](\upsilon)(\varphi) \\
&= \sum_{\varphi \leq_K \upsilon \bullet \mathbf{1}_E} hg[K](\upsilon)(\varphi) \\
&= \frac{\sum_{\varphi \leq_K \upsilon \bullet \mathbf{1}_E} \binom{\upsilon \bullet \mathbf{1}_E}{\varphi}}{\binom{L}{K}} = \frac{\binom{L_E}{K}}{\binom{L}{K}} \qquad \text{by Lemma 1.8.2.}
\end{aligned}$$

As a result:

$$\begin{aligned}
hg[K](\upsilon)\big|_{\overline{\mathbf{1}_E}^{\bullet}}(\varphi) = \sum_{\varphi \leq_K \upsilon} \frac{hg[K](\upsilon)(\varphi) \cdot \overline{\mathbf{1}_E}^{\bullet}(\varphi)}{hg[K](\upsilon) \models \overline{\mathbf{1}_E}^{\bullet}} \big|\varphi\rangle &= \sum_{\varphi \leq_K \upsilon \bullet \mathbf{1}_E} \frac{\prod_x \binom{\upsilon(x)}{\varphi(x)}}{\binom{L}{K}} \cdot \frac{\binom{L}{K}}{\binom{L_E}{K}} \big|\varphi\rangle \\
&= \sum_{\varphi \leq_K \upsilon \bullet \mathbf{1}_E} \frac{\binom{\upsilon \bullet \mathbf{1}_E}{\varphi}}{\binom{L_E}{K}} \big|\varphi\rangle = hg[K](\upsilon \bullet \mathbf{1}_E).
\end{aligned}$$

2 Similarly. ❑

We include an example involving updates of a hypergeometric distribution, but it does not exploit the equations in the last theorem.

Example 6.8.4 Suppose a bank robbery took place, involving three robbers. Afterwards, the police arrests five suspects, for whom we use a set of labels $X = \{a, b, c, d, f\}$. Their sex is included via the set $S = \{M, F\}$, so the arrested group can be described as multiset:

$$\psi = 1\big|a, F\rangle + 1\big|b, M\rangle + 1\big|c, M\rangle + 1\big|d, F\rangle + 1\big|e, M\rangle \in \mathcal{N}(X \times S).$$

We take as prior distribution for the gang of robbers the hypergeometric distribution of size 3. It forms a uniform distribution over all options:

$$\begin{aligned}
\omega_0 \coloneqq hg[3](\psi) = {} & \tfrac{1}{10}\Big|\,1|a, F\rangle + 1|b, M\rangle + 1|c, M\rangle\Big\rangle + \tfrac{1}{10}\Big|\,1|a, F\rangle + 1|b, M\rangle + 1|d, F\rangle\Big\rangle \\
& + \tfrac{1}{10}\Big|\,1|a, F\rangle + 1|c, M\rangle + 1|d, F\rangle\Big\rangle + \tfrac{1}{10}\Big|\,1|b, M\rangle + 1|c, M\rangle + 1|d, F\rangle\Big\rangle \\
& + \tfrac{1}{10}\Big|\,1|a, F\rangle + 1|b, M\rangle + 1|e, M\rangle\Big\rangle + \tfrac{1}{10}\Big|\,1|a, F\rangle + 1|c, M\rangle + 1|e, M\rangle\Big\rangle \\
& + \tfrac{1}{10}\Big|\,1|b, M\rangle + 1|c, M\rangle + 1|e, M\rangle\Big\rangle + \tfrac{1}{10}\Big|\,1|a, F\rangle + 1|d, F\rangle + 1|e, M\rangle\Big\rangle \\
& + \tfrac{1}{10}\Big|\,1|b, M\rangle + 1|d, F\rangle + 1|e, M\rangle\Big\rangle + \tfrac{1}{10}\Big|\,1|c, M\rangle + 1|d, F\rangle + 1|e, M\rangle\Big\rangle.
\end{aligned}$$

There is a witness with relevant information. The witness first of all tells to be sure that two of the robbers are male (and one female). We incorproate this in the sharp predicate $p\colon \mathcal{N}[3](X \times S) \to \{0, 1\}$ given by:

$$p(\varphi) = 1 \iff \varphi[0, 1](M) = 2 \iff \sum_{x \in X} \varphi(x, M) = 2.$$

This brings us to the first update, of the prior hypergeometric distribution ω_0, giving us another uniform distribution, indeed with only gangs with two males:

$$\begin{aligned}
\omega_1 \coloneqq \omega_0\big|_p = {} & \tfrac{1}{6}\Big|\,1|a, F\rangle + 1|b, M\rangle + 1|c, M\rangle\Big\rangle + \tfrac{1}{6}\Big|\,1|b, M\rangle + 1|c, M\rangle + 1|d, F\rangle\Big\rangle \\
& + \tfrac{1}{6}\Big|\,1|a, F\rangle + 1|b, M\rangle + 1|e, M\rangle\Big\rangle + \tfrac{1}{6}\Big|\,1|a, F\rangle + 1|c, M\rangle + 1|e, M\rangle\Big\rangle \\
& + \tfrac{1}{6}\Big|\,1|b, M\rangle + 1|d, F\rangle + 1|e, M\rangle\Big\rangle + \tfrac{1}{6}\Big|\,1|c, M\rangle + 1|d, F\rangle + 1|e, M\rangle\Big\rangle.
\end{aligned}$$

The police then forms a lineup of the five suspects and asks the witness to tell, via a number between 0 and 10, how certain the witness is about the presence of each individual separately. The number 0 stands for: certainly not present, whereas 10 means: certainly present. This yields the following table.

a	b	c	d	e
8	5	1	0	9

The table easily translates into a factor $q\colon X \to \mathbb{R}_{\geq 0}$, namely $q = 8 \cdot \mathbf{1}_a + 5 \cdot \mathbf{1}_b + 1 \cdot \mathbf{1}_c + 9 \cdot \mathbf{1}_e$. Notice that the (female) subspect d is excluded, by the witness. We weaken q to a factor $q \otimes \mathbf{1}\colon X \times S \to \mathbb{R}_{\geq 0}$ and then we multicplicatively extend it to a factor $\overline{q \otimes \mathbf{1}}^{\bullet}\colon \mathcal{N}[3](X \times S) \to [0, 1]$ on multisets, as in Definition 4.6.6. Now we perform another update:

$$\begin{aligned}
\omega_2 \coloneqq \omega_1\big|_{\overline{q \otimes \mathbf{1}}^{\bullet}} = {} & \tfrac{5}{59}\Big|\,1|a, F\rangle + 1|b, M\rangle + 1|c, M\rangle\Big\rangle + \tfrac{45}{59}\Big|\,1|a, F\rangle + 1|b, M\rangle + 1|e, M\rangle\Big\rangle \\
& + \tfrac{9}{59}\Big|\,1|a, F\rangle + 1|c, M\rangle + 1|e, M\rangle\Big\rangle \\
\approx {} & 0.09\Big|\,1|a, F\rangle + 1|b, M\rangle + 1|c, M\rangle\Big\rangle + 0.76\Big|\,1|a, F\rangle + 1|b, M\rangle + 1|e, M\rangle\Big\rangle \\
& + 0.15\Big|\,1|a, F\rangle + 1|c, M\rangle + 1|e, M\rangle\Big\rangle.
\end{aligned}$$

This distribution ω_2 gives a pretty good picture of which gang of three most likely robbed the bank.

6.8.1 Sampling from updated distributions

The following program fragments perform an update of a distribution $\omega \in \mathcal{D}(X)$, on the left with a sharp predicate to $\omega|_{\mathbf{1}_E}$, for a subset $E \subseteq X$, and on the right to $\omega|_p$, for a fuzzy predicate $p\colon X \to [0,1]$.

```
x ← ω
if x in E:
   return x
```

```
x ← ω
y ← flip(p(x))
if y == 1:
   return x
```

(6.28)

The mechanism on the left is *rejection* sampling, where all the samples from ω outside the subset E are rejected. The approach on the right is called *importance* sampling. The value $p(x) \in [0,1]$ is used as input r for $\mathit{flip}(r) = r|1\rangle + (1-r)|0\rangle$ and thus determines if the orginal sample x is returned.

We briefly look into the correctness of these programs. In the sharp case, on the left, we can describe the indicator map $\mathbf{1}_E$ as a function $X \to \mathbf{2}$, where $\mathbf{2} = \{0,1\}$. We use the tuple of functions $\langle id, \mathbf{1}_E\rangle\colon X \to X \times \mathbf{2}$ to transform the multinomial distribution $mn[K](\omega) \in \mathcal{D}(\mathcal{N}[K](X))$ to a distribution in $\mathcal{D}(\mathcal{N}[K](X \times \mathbf{2}))$. The second projection $\pi_2\colon X \times \mathbf{2} \to \mathbf{2}$ can be used as a sharp predicate. Its multiplicative extension $\overline{\pi_2}^{\bullet}\colon \mathcal{N}[K](X \times \mathbf{2}) \to \mathbf{2}$ can then be used for updating. It singles out those cases where E holds. Finally, we marginalise to $\mathcal{D}(\mathcal{N}[K](X))$. Thus, as semantics of the rejection sampling on the left in (6.28), we use:

$$
\begin{aligned}
&\mathcal{D}\mathcal{N}(\pi_1)\Big(\mathcal{D}\mathcal{N}(\langle id, \mathbf{1}_E\rangle)(mn[K](\omega))\big|_{\overline{\pi_2}^{\bullet}}\Big) \\
&= \mathcal{D}\mathcal{N}(\pi_1)\Big(mn[K](\mathcal{D}(\langle id, \mathbf{1}_E\rangle)(\omega))\big|_{\overline{\pi_2}^{\bullet}}\Big) && \text{by naturality of multinomial} \\
&= \mathcal{D}\mathcal{N}(\pi_1)\Big(mn[K](\mathcal{D}(\langle id, \mathbf{1}_E\rangle)(\omega)\big|_{\pi_2})\Big) && \text{by Theorem 6.8.1 (1)} \\
&= \mathcal{D}\mathcal{N}(\pi_1)\Big(mn[K](\mathcal{D}(\langle id, \mathbf{1}_E\rangle)(\omega|_{\pi_2 \circ \langle id, \mathbf{1}_E\rangle}))\Big) && \text{by Lemma 6.1.7 (7)} \\
&= mn[K]\Big(\mathcal{D}(\pi_1)(\mathcal{D}(\langle id, \mathbf{1}_E\rangle)(\omega|_{\mathbf{1}_E}))\Big) && \text{again by naturality} \\
&= mn[K](\omega|_{\mathbf{1}_E}).
\end{aligned}
$$

Hence, the program produces samples from the updated distribution $\omega|_{\mathbf{1}_E}$, as intended.

We turn to the update on the right in (6.28), with a fuzzy predicate $p\colon X \to [0,1]$. We identify this function p with the composite $\mathit{flip} \circ p$, giving a channel $p\colon X \to\!\!\!\!\!\!\circ\; \mathbf{2}$, see Exercise 4.5.7, as used above. We can form a tuple $\langle id, p\rangle\colon X \to\!\!\!\!\!\!\circ\; X \times \mathbf{2}$, but now it is a channel. Hence we have to use the extension $\mathcal{N}[K]$ of multisets to the category of channels, from Corollary 3.8.7. We thus

take as semantics of the above program on the right in (6.28):

$$
\begin{aligned}
&\mathcal{DN}(\pi_1)\Big(\mathcal{N}[K](\langle id\,,p\rangle)_*(mn[K](\omega))\big|_{\overline{\pi_2}\bullet}\Big) \\
&= \mathcal{DN}(\pi_1)\Big(mn[K](\langle id\,,p\rangle_*(\omega))\big|_{\overline{\pi_2}\bullet}\Big) && \text{by channel naturality (3.44)} \\
&= \mathcal{DN}(\pi_1)\Big(mn[K](\langle id\,,p\rangle_*(\omega)|_{\pi_2})\Big) && \text{by Theorem 6.8.1 (1)} \\
&\overset{(*)}{=} \mathcal{DN}(\pi_1)\Big(mn[K](\textstyle\sum_x \omega|_p(x)\big|x,1\rangle)\Big) \\
&= mn[K]\Big(\mathcal{D}(\pi_1)(\textstyle\sum_x \omega|_p(x)\big|x,1\rangle)\Big) && \text{by naturality of multinomial} \\
&= mn[K](\omega|_p).
\end{aligned}
$$

For the marked equation $\overset{(*)}{=}$ we first note that the channel $\langle id\,,p\rangle\colon X \twoheadrightarrow X\times\mathbf{2}$ is of the form:

$$\langle id\,,p\rangle(x) \;=\; \sum_{x\in X} p(x)\,\big|x,1\rangle + (1-p(x))\,\big|x,0\rangle.$$

Hence:

$$
\begin{aligned}
\langle id\,,p\rangle_*(\omega)\models\pi_2 \;&=\; \sum_{x\in X} \omega(x)\cdot p(x)\cdot\pi_2(x,1) + \omega(x)\cdot(1-p(x)\cdot\pi_2(x,0) \\
&=\; \sum_{x\in X} \omega(x)\cdot p(x) \;=\; \omega\models p.
\end{aligned}
$$

And thus:

$$\langle id\,,p\rangle_*(\omega)|_{\pi_2} \;=\; \sum_{x\in X} \frac{\omega(x)\cdot p(x)}{\omega\models p}\,\big|x,1\rangle \;=\; \sum_{x\in X} \omega|_p(x)\,\big|x,1\rangle.$$

An alternative to sample from $\omega|_p$ is the Metropolis-Hastings method, see Subsection 6.7.3.

Exercises

6.8.1 Consider the following situation and questions, describing a typical update situation[2], but with a special twist.

A friend of mine has three children aged 4 and 5 with one twin.

(a) What is the probability that there are three girls, assuming that the probability of a girl is $\frac{1}{2}$?
(b) I ring this friend's doorbell and I hear a girl's voice say that she will open the door soon. If I can assume that this is one of the three children, what is the probability that my friend has three daughters?
(c) A 4-year-old girl opens the door. Still assuming this is one of the children, what is now the probability that my friend has three daughters?

Write $A = \{4,5\}$ for the ages and $C = \{b,g\}$ for the children. We form a prior distribution ω in three steps.

- First take the uniform distribution $\upsilon := \frac{1}{4}\big|b,4\rangle + \frac{1}{4}\big|b,5\rangle + \frac{1}{4}\big|g,4\rangle + \frac{1}{4}\big|g,5\rangle$ on $C\times A$.
- Next take the multinomial distribution $\mu := mn[3](\upsilon) \in \mathcal{D}(\mathcal{N}[3](C\times A))$.
- Finally we form the (sharp) twin predicate $twin\colon \mathcal{N}[3](C\times A) \to \{0,1\}$ given by $twin(\varphi) = 1$ iff $\big|supp(\varphi[0,1])\big| = 2$. Thus the $twin$ predicate holds when both ages 4 and 5 occur in the multiset φ of size 3, which means that one of the ages must occur twice in φ. We then take as prior $\omega := \mu|_{twin}$.

[2] See the special page wikipedia.org/wiki/Boy_or_girl_paradox.

1 Show (or accept) that:

$$\begin{aligned}\omega \;=\; & \tfrac{1}{16}\Big|2|b,4\rangle + 1|b,5\rangle\Big\rangle + \tfrac{1}{16}\Big|1|b,4\rangle + 2|b,5\rangle\Big\rangle + \tfrac{1}{8}\Big|1|b,4\rangle + 1|b,5\rangle + 1|g,4\rangle\Big\rangle \\ & + \tfrac{1}{16}\Big|2|b,5\rangle + 1|g,4\rangle\Big\rangle + \tfrac{1}{16}\Big|1|b,5\rangle + 2|g,4\rangle\Big\rangle + \tfrac{1}{8}\Big|1|b,4\rangle + 1|b,5\rangle + 1|g,5\rangle\Big\rangle \\ & + \tfrac{1}{16}\Big|2|b,4\rangle + 1|g,5\rangle\Big\rangle + \tfrac{1}{8}\Big|1|b,4\rangle + 1|g,4\rangle + 1|g,5\rangle\Big\rangle + \tfrac{1}{16}\Big|2|g,4\rangle + 1|g,5\rangle\Big\rangle \\ & + \tfrac{1}{8}\Big|1|b,5\rangle + 1|g,4\rangle + 1|g,5\rangle\Big\rangle + \tfrac{1}{16}\Big|1|b,4\rangle + 2|g,5\rangle\Big\rangle + \tfrac{1}{16}\Big|1|g,4\rangle + 2|g,5\rangle\Big\rangle.\end{aligned}$$

2 Compute the marginals of the multisets inside the (big) kets, that is, compute the children and ages distributions:

$$\mathcal{D}\big(\mathcal{N}(\pi_1)\big)(\omega) \in \mathcal{D}\big(\mathcal{N}[3](C)\big) \qquad\qquad \mathcal{D}\big(\mathcal{N}(\pi_2)\big)(\omega) \in \mathcal{D}\big(\mathcal{N}[3](A)\big).$$

The first distribution is the multinomial distribution, with draw size 3, for the uniform distribution on C. The second distribution is uniform.

3 Define the girls-number function $gn\colon \mathcal{N}[3](C \times A) \to \mathbf{4} = \{0,1,2,3\}$ as $gn(\varphi) = \varphi[1,0](g)$. Use this function gn as deterministic channel and prove that:

$$gn_*(\omega) \;=\; \tfrac{1}{8}|0\rangle + \tfrac{1}{4}|1\rangle + \tfrac{1}{4}|2\rangle + \tfrac{1}{8}|3\rangle.$$

Conclude that the answer to the above question (a) is $\frac{1}{8}$.

4 Use the (sharp) greater-than-one predicate $\mathbf{1}_{\geq 1} := (\mathbf{1}_0)^{\perp}\colon \mathbf{4} \to \{0,1\}$ to prove:

$$gn_*\Big(\omega|_{gn^*(\mathbf{1}_{\geq 1})}\Big) \;=\; \tfrac{3}{7}|1\rangle + \tfrac{3}{7}|2\rangle + \tfrac{1}{7}|3\rangle.$$

This means that the answer to question (b) is $\frac{1}{7}$.

5 Let's write $g4\colon \mathcal{N}[3](C \times A) \to \{0,1\}$ for the (sharp) at-least-one-4-year-girl predicate given by $g4(\varphi) = 1$ iff $\varphi(g,4) \geq 1$. Prove that:

$$gn_*\Big(\omega|_{g4}\Big) \;=\; \tfrac{3}{10}|1\rangle + \tfrac{1}{2}|2\rangle + \tfrac{1}{5}|3\rangle,$$

so that the answer to question (c) is $\frac{1}{5}$.

6.8.2 Restricting a binomial distribution to a smaller size, via updating, does not yield a binomial distribution of that smaller size, that is, in general, for $L < K$,

$$bn[K](r)|_{\mathbf{1}_{\{n \mid n \leq L\}}} \;\neq\; bn[L](r).$$

Take $r = \frac{1}{3}$, $K = 3$, $L = 2$, and check that:

1 $bn[K](r) = \frac{8}{27}|0\rangle + \frac{4}{9}|1\rangle + \frac{2}{9}|2\rangle + \frac{1}{27}|3\rangle$;
2 $bn[K](r)|_{\mathbf{1}_{\{n \mid n \leq L\}}} = \frac{4}{13}|0\rangle + \frac{6}{13}|1\rangle + \frac{3}{13}|2\rangle$;
3 $bn[L](r) = \frac{4}{9}|0\rangle + \frac{4}{9}|1\rangle + \frac{1}{9|2}$.

6.8.3 Prove Theorem 6.8.3 (2) yourself.

6.8.4 The question below is adapted from (Ross, 2018, §6.4). Let $\omega \in \mathcal{D}(X)$ be a distribution with a subset $E \subseteq X$ and with numbers $L \leq K$. Recall, for a multiset φ, that $(\varphi \bullet \mathbf{1}_E)(x) = \varphi(x) \cdot \mathbf{1}_E(x)$. Define sharp predicate $E_L\colon \mathcal{N}[K](X) \to [0,1]$ by:

$$E_L(\varphi) := \begin{cases} 1 & \text{if } \displaystyle\sum_{x \notin E} \varphi(x) = L \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $E_L(\varphi) = 1$ iff $\big\|\varphi \bullet \mathbf{1}_{\neg E}\big\| = L$.

Prove that:

$$\mathcal{D}(- \bullet \mathbf{1}_E)\Big(mn[K](\omega)\big|_{E_L}\Big) \;=\; mn[K-L](\omega|_{\mathbf{1}_E}).$$

6.8.5 Fix a multiset $\psi \in \mathcal{N}[K+1](X)$ of size $K+1$ and define the sharp predicate $mcons_\psi \colon X \times \mathcal{N}[K](X) \to [0,1]$ as:

$$mcons_\psi(x,\varphi) \;:=\; \begin{cases} 1 & \text{if } 1|x\rangle + \varphi = \psi \\ 0 & \text{otherwise.} \end{cases}$$

Show that for any distribution $\omega \in \mathcal{D}(X)$,

$$\omega \otimes mn[K](\omega)\big|_{mcons_\psi} \;=\; DSD(\psi),$$

where the draw-store-delete channel DSD is from Exercise 3.2.12. Earlier we saw that this equation follows from a general result, see Example 6.3.3 (2).

6.8.6 We return to the multiplicative extension $\overline{p}^\bullet \colon \mathcal{L}(X) \to \mathbb{R}_{\geq 0}$ to lists, of a factor $p \colon X \to \mathbb{R}_{\geq 0}$, from Exercise 4.6.8.

1 Show that:

$$iid[K](\omega)\big|_{\overline{p}^\bullet} \;=\; iid[K](\omega|_p).$$

2 Prove next for Poisson-iid that:

$$Piid[\lambda](\omega)\big|_{\overline{p}^\bullet} \;=\; Piid[\lambda \cdot (\omega \models p)](\omega).$$

6.9 Discretisation, and coin bias learning

In this chapter we have introduced updating for distributions $\omega \in \mathcal{D}(X)$. It involves incorporating evidence p into a new distribution $\omega|_p$ making p more true, see Theorem 6.1.6. In this way we learn which element $x \in X$ has the highest probability, and thus is the most likely.

This approach can also be used for what is called parameter learning. Consider a coin $flip(r)$, with a bias $r \in [0,1]$ that is unkown. What we do know are a few coin flips: a certain sequence (or multiset) of heads and tails. What can we then infer about r? By following the above approach we seek a distribution ω on the space of parameters r, and update this distribution with the evidence (of flips). A problem is that the bias r ranges over a continuous space $[0,1]$, whereas the discrete distributions that we have used so far have a finite support — or possibly countable support, for $\mathcal{D}_\infty$. This situation forms a good motivation for continuous probability theory, which we postpone to Chapter **??**.

In the meantime the situation is not hopeless. What we can do is discretise: chop up the continuous interval $[0,1]$ into finitely many points and apply the discrete techniques that we do know (like in Riemann integration). If we let the number of points increase, we may still get reasonably good results. This is indeed the case, as we will illustrate below. In fact, computers doing computation in continuous probability perform such fine-grained discretisation.

We start with discretisation of an interval of real numbers.

Definition 6.9.1 Let $a, b \in \mathbb{R}$ with $a < b$ and $N \in \mathbb{N}$ with $N > 0$ be given.

1 We write $[a,b]_N \subseteq [a,b] \subseteq \mathbb{R}$ for the interval $[a,b]$ reduced to N elements:

$$\begin{aligned}[a,b]_N &:= \left\{ a + (i + \tfrac{1}{2})s \;\middle|\; 0 \leq i < N \right\} \qquad \text{where} \qquad s := \frac{b-a}{N} \\ &= \left\{ a + \tfrac{1}{2}s,\ a + \tfrac{3}{2}s, \ldots, a + \tfrac{2N-1}{2}s \right\} \\ &= \left\{ a + \tfrac{1}{2}s,\ a + \tfrac{3}{2}s, \ldots, b - \tfrac{1}{2}s \right\}.\end{aligned}$$

2 Let $f\colon S \to \mathbb{R}_{\geq 0}$ be a function, defined on a finite subset $S \subseteq \mathbb{R}$. We write $Disc(f,S) \in \mathcal{D}(S)$ for the discrete distribution defined as:

$$Disc(f,S) := \sum_{x \in S} \tfrac{f(x)}{t} \big| x \rangle \qquad \text{where} \qquad t := \sum_{x \in S} f(x).$$

Often we combine the notations from these two items and use discretised states of the form $Disc(f,[a,b]_N)$.

To see an example of item (1), consider the interval $[1,2]$ with $N = 3$. The step size s is then $s = \frac{2-1}{3} = \frac{1}{3}$, so that:

$$[1,2]_3 = \left\{ 1 + \tfrac{1}{2} \cdot \tfrac{1}{3},\ 1 + \tfrac{3}{2} \cdot \tfrac{1}{3},\ 1 + \tfrac{5}{2} \cdot \tfrac{1}{3} \right\} = \left\{ 1 + \tfrac{1}{6},\ 1 + \tfrac{1}{2},\ 1 + \tfrac{5}{6} \right\}.$$

We choose to use internal points only and exclude the end-points in this finite subset since the end-points sometimes give rise to boundary problems, with functions being undefined. When N goes to infinity, the smallest and largest elements in $[a,b]_N$ will approximate the end-points a and b — from above and from below, respectively.

The 'total' number t in item (2) normalises the formal sum and ensures that the multiplicities add up to one. In this way we can define a uniform distribution on $[a,b]_N$ as $unif_{[a,b]_N}$, like before, or alternatively as $Disc(\mathbf{1},[a,b]_N)$, where $\mathbf{1}$ is the constant-one function.

Example 6.9.2 We look at the following classical question: suppose we are given a coin with an unknown bias, we flip it eight times, and observe the following list of heads (H) and tails (T):

$$[T,H,H,H,T,T,H,H].$$

What can we then say about the bias of the coin?

The frequentist approach that we have seen in Section 2.2 would turn the above list into a multiset and then into a distribution, by frequentist learning, see also Diagram (2.24). This gives:

$$[T,H,H,H,T,T,H,H] \longmapsto 5|H\rangle + 3|T\rangle \longmapsto \tfrac{5}{8}|H\rangle + \tfrac{3}{8}|T\rangle.$$

Here we do not use this frequentist approach to learning the bias parameter, but take a Bayesian route. We assume that the bias parameter itself is given by a distribution, describing the likelihoods of various bias values. We assume no prior knowledge and therefore start from the uniform distribution. It will be updated based on successive observations, using the technique of backward inference, see Definition 6.2.1 (2).

The bias b of a coin is a number in the unit interval $[0,1]$, giving rise to a coin distribution $flip(b) = b|H\rangle + (1-b)|T\rangle$. Thus we can see $flip$ as a channel $flip\colon [0,1] \rightsquigarrow \{H,T\}$. At this stage we avoid continuous distributions and discretise the unit interval. We choose $N = 100$ in the chop up, giving as underlying space $[0,1]_N$ with $N = 100$ points, on which we take the uniform

distribution *unif* as prior:

$$\begin{aligned}\upsilon \;:=\; \mathit{unif}_{[0,1]_{100}} \;:=\; \mathit{Disc}(\mathbf{1}, [0,1]_{100}) &= \sum_{x \in [0,1]_N} \tfrac{1}{N}|x\rangle \\ &= \sum_{0 \leq i < N} \tfrac{1}{N}\Big|\tfrac{2i+1}{2N}\Big\rangle \;:=\; \tfrac{1}{N}\Big|\tfrac{1}{2N}\Big\rangle + \tfrac{1}{N}\Big|\tfrac{3}{2N}\Big\rangle + \cdots + \tfrac{1}{N}\Big|\tfrac{2N-1}{2N}\Big\rangle.\end{aligned}$$

We use the *flip* operation as a channel, restricted to the discretised space:

$$[0,1]_N \xrightarrow{\mathit{flip}} \{H, T\} \qquad \text{given by} \qquad \mathit{flip}(b) \;=\; b|H\rangle + (1-b)|T\rangle.$$

There are the two (sharp, point) predicates $\mathbf{1}_H$ and $\mathbf{1}_T$ on the codomain $\{H, T\}$. It is not hard to show, see Exercise 6.9.2 below, that:

$$\mathit{flip}_*(\upsilon) \models \mathbf{1}_H \;=\; \mathit{flip}_*(\upsilon) \models \mathbf{1}_T \;=\; \tfrac{1}{2}.$$

Predicate transformation along *flip* yields two predicates on $[0,1]_N$ given by:

$$\mathit{flip}^*(\mathbf{1}_H)(r) \;=\; r \qquad \text{and} \qquad \mathit{flip}^*(\mathbf{1}_T)(r) \;=\; 1 - r.$$

Given the above sequence of head/tail observations $[T, H, H, H, T, T, H, H]$, we perform successive predicate transformations $\mathit{flip}^*(\mathbf{1}_{(-)})$ and update the prior (uniform) state accordingly. This gives, via Lemma 6.1.7 (3),

$$\begin{aligned}
&\upsilon|_{\mathit{flip}^*(\mathbf{1}_T)} \\
&\upsilon|_{\mathit{flip}^*(\mathbf{1}_T)}|_{\mathit{flip}^*(\mathbf{1}_H)} \;=\; \upsilon|_{\mathit{flip}^*(\mathbf{1}_T)\,\&\,\mathit{flip}^*(\mathbf{1}_H)} \;=\; \upsilon|_{\mathit{flip}^*(\mathbf{1}_H)\,\&\,\mathit{flip}^*(\mathbf{1}_T)} \\
&\upsilon|_{\mathit{flip}^*(\mathbf{1}_T)}|_{\mathit{flip}^*(\mathbf{1}_H)}|_{\mathit{flip}^*(\mathbf{1}_H)} \;=\; \upsilon|_{\mathit{flip}^*(\mathbf{1}_T)\,\&\,\mathit{flip}^*(\mathbf{1}_H)\,\&\,\mathit{flip}^*(\mathbf{1}_H)} \\
&\qquad\qquad\qquad\qquad\qquad\qquad = \upsilon|_{\mathit{flip}^*(\mathbf{1}_H)^2\,\&\,\mathit{flip}^*(\mathbf{1}_T)} \\
&\qquad \vdots \\
&\upsilon|_{\mathit{flip}^*(\mathbf{1}_H)^5\,\&\,\mathit{flip}^*(\mathbf{1}_T)^3}
\end{aligned}$$

As we already know from Lemma 6.1.7 (3), the order of updating does not matter. Indeed, the order of the coin flips in the list $[T, H, H, H, T, T, H, H]$ of observations does not matter. What matters is the associated multiset of observations $5|H\rangle + 3|T\rangle$. The multiplicities 5 and 3 reappear in the predicate $\mathit{flip}^*(\mathbf{1}_H)^5 \;\&\; \mathit{flip}^*(\mathbf{1}_T)^3$ in the last line above.

An overview of the distributions arising from these succesive updates is given in Figure 6.17. Later on, in Definition 8.3.5 (2), we shall identify this as updating according to Pearl (and not Jeffrey). These distributions approximate (continuous) Beta distributions, see Example **??** later on. The probability distribution functions of these Beta distributions form a smoothed out version of the bar charts in Figure 6.17.

After these eight updates, let's write $\rho \;:=\; \mathit{unif}|_{\mathit{flip}^*(\mathbf{1}_H)^5 \& \mathit{flip}^*(\mathbf{1}_T)^3}$ for the resulting distribution. We now ask three questions.

1 Where does ρ reach its highest value, and what is it? The answers are given by the singleton set of maximum values:

$$\operatorname{argmax}(\rho) \;=\; \Big\{\tfrac{5}{8}\Big\} \qquad \text{with} \qquad \rho(\tfrac{5}{8}) \;\approx\; 0.025347.$$

2 What is the predicted coin distribution? The outcome, with truncated multiplicities, is:

$$\mathit{flip}_*(\rho) \;=\; 0.6|H\rangle + 0.4|T\rangle.$$

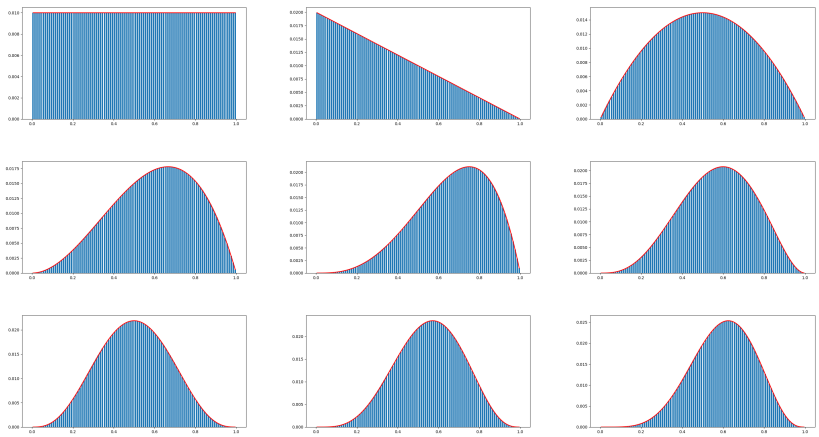


Figure 6.17 Coin bias distributions arising from the prior discrete uniform distribution on the discretised unit interval $[0,1]_N$ for $N = 100$, via backward inferences using successive coin observations $[T, H, H, H, T, T, H, H]$. The red line in these plots is given by $1/100$ times the probability density function of the corresponding continuous *Beta*-distribution. It shows how closely the discrete approximations match the continuous probability density function.

3 What is the expected value of ρ? It is:

$$Mean(\rho) = 0.6 = \rho \models flip^*(\mathbf{1}_H).$$

For mathematical reasons[3] the exact outcome is 0.6. However, we have used approximation via discretisation. The value computed with this discretisation is 0.599999985. We can conclude that chopping the unit interval up with $N = 100$ already gives a fairly good approximation.

It turns out that that all the distributions in Figure 6.17 can be described in a uniform manner, as a certain parameterised class of distributions, which is closed under backward inference, along the *flip* channel. Updating along *flip* corresponds simply to an increase of parameters. This is very convenient since it means that we do not have to perform the — computationally costly — distribution updates, but we can simply do parameter updates. Such situations occur more often and involve a 'conjugate prior' relationship. We shall study it more systematically later on, in Subsection **??**. At this stage we only describe the relevant class of distributions and show how it is closed under backward inference along *flip*.

Thus we define, dependent on the discretisation parameter $N \in \mathbb{N}$, the channel $Beta_N \colon \mathbb{N}_{>0} \times \mathbb{N}_{>0} \rightsquigarrow [0,1]_N$ as normalisation:

$$Beta_N(a,b) := Flrn\left(\sum_{r\in[0,1]_N} r^{a-1}\cdot(1-r)^{b-1}\big|r\rangle\right) = \sum_{r\in[0,1]_N} \frac{r^{a-1}\cdot(1-r)^{b-1}}{\sum_{s\in[0,1]_N} s^{a-1}\cdot(1-s)^{b-1}}\big|r\rangle. \quad (6.29)$$

Then we have the following results.

[3] The distribution ρ is an approximation of the probability density function $Beta(6,4)$, which has mean $\frac{6}{6+4} = \frac{3}{5} = 0.6$, see Exercise **??**.

Proposition 6.9.3 *Let $N \in \mathbb{N}$ be the the discretisation parameter, used in the chopped subspace $[0,1]_N \subseteq [0,1]$, with uniform distribution $\upsilon_N := \mathit{unif}_{[0,1]_N} \in \mathcal{D}([0,1]_N)$, and let $a, b \in \mathbb{N}_{>0}$.*

1 The discretised Beta distribution $\mathrm{Beta}_N(a,b)$ satisfies:

$$\mathit{Beta}_N(a,b) \;=\; \upsilon_N \,\big|_{\mathit{flip}^*(\mathbf{1}_H)^{a-1} \,\&\, \mathit{flip}^*(\mathbf{1}_T)^{b-1}}.$$

2 The class of discretised Beta distributions is closed under backward inference along flip:

$$\begin{aligned} \mathit{Beta}_N(a,b)\big|_{\mathit{flip}^*(\mathbf{1}_H)} &= \mathit{Beta}_N(a+1,b) \\ \mathit{Beta}_N(a,b)\big|_{\mathit{flip}^*(\mathbf{1}_T)} &= \mathit{Beta}_N(a,b+1). \end{aligned}$$

More generally, for $n, m \in \mathbb{N}$ one has:

$$\mathit{Beta}_N(a,b)\big|_{\mathit{flip}^*(\mathbf{1}_H)^n \,\&\, \mathit{flip}^*(\mathbf{1}_T)^m} \;=\; \mathit{Beta}_N(a+n,b+m).$$

This second item shows how backward inference of a prior $\mathit{Beta}_N(a,b)$ along a flip can be expressed via an update of the 'hyperparameters' a, b.

Proof. 1 For $r \in [0,1]_N$ we have:

$$\begin{aligned} \upsilon_N\big|_{\mathit{flip}^*(\mathbf{1}_H)^{a-1} \,\&\, \mathit{flip}^*(\mathbf{1}_T)^{b-1}}(r) &= \frac{\upsilon_N(r) \cdot \mathit{flip}^*(\mathbf{1}_H)^{a-1}(r) \cdot \mathit{flip}^*(\mathbf{1}_T)^{b-1}(r)}{\upsilon_N \models \mathit{flip}^*(\mathbf{1}_H)^{a-1} \;\&\; \mathit{flip}^*(\mathbf{1}_T)^{b-1}} \\ &= \frac{{}^1\!/_N \cdot r^{a-1} \cdot (1-r)^{b-1}}{\sum_{s \in [0,1]_N} {}^1\!/_N \cdot s^{a-1} \cdot (1-s)^{b-1}} \;:=\; \mathit{Beta}_N(a,b). \end{aligned}$$

2 We use this result and Lemma 6.1.7 (3) to prove the general statement:

$$\begin{aligned} \mathit{Beta}_N(a,b)\big|_{\mathit{flip}^*(\mathbf{1}_H)^n \,\&\, \mathit{flip}^*(\mathbf{1}_T)^m} &= \upsilon_N \,\big|_{\mathit{flip}^*(\mathbf{1}_H)^{a-1} \,\&\, \mathit{flip}^*(\mathbf{1}_T)^{b-1}}\big|_{\mathit{flip}^*(\mathbf{1}_H)^n \,\&\, \mathit{flip}^*(\mathbf{1}_T)^m} \\ &= \upsilon_N \,\big|_{\mathit{flip}^*(\mathbf{1}_H)^{a-1} \,\&\, \mathit{flip}^*(\mathbf{1}_T)^{b-1} \,\&\, \mathit{flip}^*(\mathbf{1}_H)^n \,\&\, \mathit{flip}^*(\mathbf{1}_T)^m} \\ &= \upsilon_N \,\big|_{\mathit{flip}^*(\mathbf{1}_H)^{a+n-1} \,\&\, \mathit{flip}^*(\mathbf{1}_T)^{b+m-1}} \;:=\; \mathit{Beta}_N(a+n,b+m). \qquad \square \end{aligned}$$

6.9.1 Discretisation of distributions

In the beginning of this section we have described how to chop up intervals $[a,b]$ of real numbers into a discrete sample space. We have used this in particular for the unit interval $[0,1]$ that we used as space for a coin bias. Since there is an isomorphism $[0,1] \cong \mathcal{D}(\mathbf{2})$, this discretisation of $[0,1]$ might as well be seen as a discretisation of the set of distributions on $\mathbf{2} = \{0,1\}$. Can we do such discretisation more generally, for sets of distributions $\mathcal{D}(X)$? There is an easy way to do so via normalisation of natural multisets.

Recall from Theorem 4.8.11 (1) that for $K > 0$ we collect 'fractional' distributions in a subset:

$$\mathcal{D}[K](X) \;:=\; \Big\{ \mathit{Flrn}(\varphi) \;\Big|\; \varphi \in \mathcal{N}[K](X) \Big\} \;=\; \Big\{ \tfrac{1}{K} \cdot \varphi \;\Big|\; \varphi \in \mathcal{N}[K](X) \Big\}.$$

Recall also — from Proposition 1.8.7 — that if the set X has n elements, then $\mathcal{N}[K](X)$ contains $\left(\!\binom{n}{K}\!\right)$ multisets, so that $\mathcal{D}[K](X)$ contains $\left(\!\binom{n}{K}\!\right)$ distributions. For instance, when $X = \{a,b\}$, then $\mathcal{M}[5](X)$ contains $\left(\!\binom{2}{5}\!\right) = \frac{6!}{5! \cdot 1!} = 6$ the multisets:

$$5|a\rangle \qquad 4|a\rangle + 1|b\rangle \qquad 3|a\rangle + 2|b\rangle \qquad 2|a\rangle + 3|b\rangle \qquad 1|a\rangle + 4|b\rangle \qquad 5|b\rangle.$$

Hence, $\mathcal{D}[5](\{a,b\})$ contains the six fractional distributions:

$$1|a\rangle \qquad \tfrac{4}{5}|a\rangle + \tfrac{1}{5}|b\rangle \qquad \tfrac{3}{5}|a\rangle + \tfrac{2}{5}|b\rangle \qquad \tfrac{2}{5}|a\rangle + \tfrac{3}{5}|b\rangle \qquad \tfrac{1}{5}|a\rangle + \tfrac{4}{5}|b\rangle \qquad 1|b\rangle.$$

By taking large K in $\mathcal{D}[K](X)$ we can obtain fairly good approximations, since the union of these subsets $\mathcal{D}[K](X) \subseteq \mathcal{D}(X)$ is dense, by Theorem 4.8.11 (1). Figure 4.5 gives a visual representation.

Example 6.9.4 Consider an election[4] with three candidates a, b, c, which we collect in a set X of candidates, as $X := \{a,b,c\}$. A candidate wins if (s)he gets more than 50% of the votes. A poll has been conducted among 100 possible voters, giving the following preferences.

candidates	a	b	c
poll numbers	52	28	20

What is the probability that candidate a then wins in the election? Notice that these are poll numbers, before the election, not actual votes. If they were actual votes, candidate a gets more than 50%, and thus wins. However, this is a poll and we like to learn about voter preferences.

A state in $\mathcal{D}(X)$ is used as distribution of preferences over the three candidates in $X = \{a,b,c\}$. In this example we discretise the set of states and work with the finite subset $\mathcal{D}[K](X) \hookrightarrow \mathcal{D}(X)$ of fractional distributions. There is no initial knowledge about voter preferences, so our prior υ_K is the uniform distribution over these preference distributions:

$$\upsilon_K \coloneqq unif_{\mathcal{D}[K](X)} = \sum_{\varphi \in \mathcal{N}[K](X)} \frac{1}{\binom{\binom{3}{K}}{}} \big|Flrn(\varphi)\big\rangle \in \mathcal{D}\big(\mathcal{D}[K](X)\big).$$

We use the following sharp predicates on $\mathcal{D}[K](X)$, called: aw for a wins, bw for b wins, cw for c wins, and nw for no-one wins. They are defined on a fractional distribution $\sigma \in \mathcal{D}[K](X)$ as:

$$\begin{aligned}
aw(\sigma) = 1 &\iff \sigma(a) > \tfrac{1}{2} \\
bw(\sigma) = 1 &\iff \sigma(b) > \tfrac{1}{2} \\
cw(\sigma) = 1 &\iff \sigma(c) > \tfrac{1}{2} \\
nw(\sigma) = 1 &\iff \sigma(a) \leq \tfrac{1}{2} \text{ and } \sigma(b) \leq \tfrac{1}{2} \text{ and } \sigma(c) \leq \tfrac{1}{2}.
\end{aligned}$$

The prior validities:

$$\upsilon_K \models aw \qquad \upsilon_K \models bw \qquad \upsilon_K \models cw \qquad \upsilon_K \models nw$$

all approximate $\frac{1}{4}$, as K goes to infinity.

We use the inclusion $\mathcal{D}[K](X) \hookrightarrow \mathcal{D}(X)$ as a channel $i\colon \mathcal{D}[K](X) \rightarrowtail X$. Like in the above coin scenario, we perform successive backward inferences, using the above poll numbers, to obtain a posterior ρ_K obtained via updates:

$$\rho_K \coloneqq \upsilon_K|_{i^*(\mathbf{1}_a)^{52}}|_{i^*(\mathbf{1}_b)^{28}}|_{i^*(\mathbf{1}_c)^{20}} = \upsilon_K|_{i^*(\mathbf{1}_a)^{52} \,\&\, i^*(\mathbf{1}_b)^{28} \,\&\, i^*(\mathbf{1}_c)^{20}}.$$

The posterior validity $\rho_K \models aw$ expresses the probability that a wins, based on the poll numbers. This validity has the following outcomes, for several values of the discretisation parameter K.

[4] This illustration is copied and adapted from Sam Staton's tutorial at the Mathematical Foundations of Programming Semantics conference, 2021.

	$K = 100$	$K = 500$	$K = 1000$
probability that a wins	0.578	0.609	0.613

Clearly, these validities are approximations. When modeled via continuous probability theory, see Example **??**, the probability that a wins can be calculated more accurately as 0.617. This illustrates that the above discretisations of states work reasonably well.

To give a bit more perspective, the posterior probability that b or c wins with the above poll numbers is close to zero. The probability that no-one wins is substantial, namely almost 0.39.

Exercises

6.9.1 Recall the N-element set $[a, b]_N$ from Definition 6.9.1 (1).

1 Show that its largest element is $b - \frac{1}{2}s$, where $s = \frac{b-a}{N}$.

2 Prove that $\sum_{x \in [a,b]_N} x = \frac{N(a+b)}{2}$, via Proposition 1.2.6 (1).

3 Show that $[0, 1]_N$ is closed under orthosupplement: if $r \in [0, 1]_N$, then also $1-r \in [0, 1]_N$. It it an effect algebra?

6.9.2 Consider coin parameter learning in Example 6.9.2.

1 Show that the predicition in the prior (uniform) distribution $\upsilon_N := \mathit{unif}_{[0,1]_N}$ on $[0, 1]_N$ gives a fair coin, *i.e.*

$$\mathit{flip}_*(\upsilon_N) = \tfrac{1}{2}|1\rangle + \tfrac{1}{2}|0\rangle.$$

This equality is independent of N.

2 Prove that, also independently of N,

$$\upsilon_N \models \mathit{flip}^*(\mathbf{1}_H) = \tfrac{1}{2}.$$

3 Show next that:

$$\upsilon_N|_{\mathit{flip}^*(\mathbf{1}_H)} = \sum_{x \in [0,1]_N} \tfrac{2x}{N}|x\rangle.$$

4 Use Proposition 1.2.6 (2) to prove that:

$$\mathit{flip}_*(\upsilon_N|_{\mathit{flip}^*(\mathbf{1}_H)})(H) = \tfrac{2}{3} - \tfrac{1}{6N^2}.$$

Conclude that $\mathit{flip}_*(\upsilon_N|_{\mathit{flip}^*(\mathbf{1}_H)})$ approaches $\frac{2}{3}|H\rangle + \frac{1}{3}|T\rangle$ as N goes to infinity.

6.9.3 Check that the probabilities $\mathit{flip}_*(\rho)(H)$ and $\mathit{Mean}(\rho)$ are the same, in Example 6.9.2.

6.9.4 The following example involving the reconstruction of a prior from a posterior is adapted from Griffiths and Tenenbaum (2007). Person A claims to have psychokinetic abilities and to be able to infleunce the outcome of a fair coin toss in such a way that in three quarters of the cases heads comes up. This is tried in an experiment, where out of 10 toin tosses 9 turn out to be heads. Person B watching the experiment tells afterwards to now be 60% sure that A indeed has psychokinetic powers. How sure was B before the experiment?

We take a set $X = \{p, p^\perp\}$ capturing whether A has psychokinetic powers (p) or not ($p^\perp$). There is a channel $c\colon X \rightsquigarrow \{H, T\}$ where $c(p) = \mathit{flip}(\frac{3}{4})$ and $c(p^\perp) = \mathit{flip}(\frac{1}{2})$.

1 Define $q \in \mathit{Pred}(X)$ as $q := \mathit{flip}^*(\mathbf{1}_H)^9 \mathbin{\&} \mathit{flip}^*(\mathbf{1}_T)^1$. Show that:

$$q = \tfrac{3^9}{4^{10}} \cdot \mathbf{1}_p + \tfrac{1}{2^{10}} \cdot \mathbf{1}_{p^\perp} = \tfrac{19683}{1048576} \cdot \mathbf{1}_p + \tfrac{1}{2024} \cdot \mathbf{1}_{p^\perp}.$$

2 Let $\omega(r) := r|p\rangle + (1-r)|p^\perp\rangle$, where r is a variable. Show that solving $\omega(r)|_q = \frac{3}{5}|p\rangle + \frac{2}{5}|p^\perp\rangle$ yields as solution:

$$r = \frac{3 \cdot 2^{10}}{2 \cdot 3^9 + 3 \cdot 2^{10}} \approx 0.0724.$$

Conclude that B's prior believe in psychokinesis is just over 7%.

3 Form the channel $mn[10](c) = mn[10] \circ c \colon X \rightsquigarrow \mathcal{N}[10](\{H,T\})$, with multiset $\varphi = 9|H\rangle + 1|T\rangle$. Check that updating an arbitrary distribution $\rho \in \mathcal{D}(\{H,T\})$ with the above predicate q is the same as updating ρ with $mn[10](c)^*(\mathbf{1}_\varphi)$.

(This fact will be appear in general form in Proposition 8.4.1 (2).)

6.9.5 Check that $Beta_N(1,1)$ in (6.29) is the uniform distribution on $[0,1]_N$.

6.9.6 Consider the discretised beta channel $Beta_N \colon \mathbb{N} \times \mathbb{N} \to \mathcal{D}(X)$ in the following diagram:

$$\begin{tikzcd}
(\mathbb{N}_{>0} \times \mathbb{N}_{>0}) \times \mathcal{N}(\{H,T\}) \arrow[r, "Beta_N \times id"] \arrow[d, "add"'] & \mathcal{D}([0,1]_N) \times \mathcal{N}(\{H,T\}) \arrow[d, "infer"] \\
\mathbb{N}_{>0} \times \mathbb{N}_{>0} \arrow[r, "Beta_N"] & \mathcal{D}([0,1]_N)
\end{tikzcd}$$

where the maps *add* and *infer* are given by:

$$\begin{aligned}
add(a,\, b,\, n|H\rangle + m|T\rangle) &= (a+n, b+m) \\
infer(\omega,\, n|H\rangle + m|T\rangle) &= \omega|_{flip^*(\mathbf{1}_H)^n \,\&\, flip^*(\mathbf{1}_T)^m}.
\end{aligned}$$

Prove the following points.

1 The map *add* is an action of the monoid $\mathcal{N}(\{H,T\})$ on $\mathbb{N}_{>0} \times \mathbb{N}_{>0}$.

2 The function $infer \colon \mathcal{D}([0,1]_N) \times \mathcal{N}(\{H,T\}) \to \mathcal{D}([0,1]_N)$ is a monoid action — when possible undefinedness is ignored.

3 The above rectangle commutes — making $Beta_N$ a homomorphism of monoid actions. (This description of a conjugate prior relationship as a monoid action comes from Jacobs (2023b).)

6.9.7 We accept, without proof, the following equation. For $a, b \in \mathbb{N}$,

$$\lim_{N\to\infty} \frac{1}{N} \cdot \sum_{r \in [0,1]_N} r^{a-1} \cdot (1-r)^{b-1} = \frac{(a-1)! \cdot (b-1)!}{(a+b-1)!}. \qquad (*)$$

Use equation $(*)$ to prove that the binary Pólya distribution can be obtained from the binomial, via the limit of the discretised *Beta* distribution (6.29):

$$\lim_{N\to\infty} bn[K]_*(Beta_N(a,b))(i) = pl[K]\Big(a|H\rangle + b|T\rangle\Big)\Big(i|H\rangle + (K-i)|T\rangle\Big).$$

Later, in Exercise **??** we shall see a proper continuous version of this result.

6.9.8 Redo the coin bias learning from the beginning of this section in the style of Example 6.9.4, via discretisation $\mathcal{D}[K](\mathbf{2})$ of states in $\mathcal{D}(\mathbf{2}) \cong [0,1]$.

6.10 Frequentist and Bayesian probability

Here, at the end of this chapter, we have encountered some of the basic concepts in (our treatment of) discrete probability theory, namely: distributions (states), predicates, validity, and updating. Before continuing we like to sit back and reflect on what we have seen so far, from a more abstract perspective.

We start with a fundamental result, for which we recall that the set $Pred(X) = [0,1]^X$ of fuzzy predicates on a set X has the structure of an effect module: there are truth and falsity predicates $\mathbf{1}, \mathbf{0}$, orthocomplement $p^\perp$, partial sum $p \ovee q$, and scalar multiplication $r \cdot p$, see Subsection 4.4.3.

The next result is a Riesz-style representation result, representing distributions on X via a double dual $[0,1]^{[0,1]^X}$.

Theorem 6.10.1 *1 Let X be a finite set. There is then a 'representation' isomorphism,*

$$\mathcal{D}(X) \xrightarrow[\cong]{\mathcal{V}} \mathbf{EMod}\big(Pred(X),\ [0,1]\big) \tag{6.30}$$

given by validity: $\mathcal{V}(\omega)(p) := \omega \models p$. It gives a bijective correspondence between distributions in $\mathcal{D}(X)$ and 'predicate evaluations' $\mathsf{Pred}(X) \to [0,1]$ that preserve the effect module structure.

2 This isomorphism (6.30) is natural in X, with respect to channels $c\colon X \rightarrowtail Y$; they give a commuting diagram of the following form, combining state transformation c_* and predicate transformation c^*.

$$\begin{tikzcd}
\mathcal{D}(X) \arrow[r, "\mathcal{V}", "\cong"'] \arrow[d, "c_*"'] & \mathbf{EMod}(Pred(X), [0,1]) \arrow[d, "h \mapsto h \circ c^*"] \\
\mathcal{D}(Y) \arrow[r, "\mathcal{V}", "\cong"'] & \mathbf{EMod}(Pred(Y), [0,1])
\end{tikzcd}$$

This result uses the notation $\mathbf{EMod}(Pred(X),\ [0,1])$ for the 'hom' set of homomorphisms of effect modules $Pred(X) \to [0,1]$. The mapping $X \mapsto \mathbf{EMod}(Pred(X),\ [0,1])$ forms a monad on **Sets**, known as the expectation monad, see Jacobs et al. (2016).

Proof. 1 We have already seen that $\mathcal{V}(\omega)\colon Pred(X) \to [0,1]$ preserves the effect module structure, see Lemma 4.4.6. We exploit that X is a finite set for showing that the function $\mathcal{V}$ is an isomorphism. Let $h\colon Pred(X) \to [0,1]$ be a homomorphism of effect modules. It gives rise to a distribution, via point predicates $\mathbf{1}_x \in Pred(X)$, for $x \in X$.

$$\mathcal{V}^{-1}(\phi) := \sum_{x \in X} \phi(\mathbf{1}_x) \big|x\rangle \in \mathcal{D}(X).$$

The probabilities $\phi(\mathbf{1}_x)$ add up to one since ϕ is a map of effect modules:

$$\sum_{x \in X} \phi(\mathbf{1}_x) = \phi\Big(\bigovee_{x \in X} \mathbf{1}_x\Big) = \phi(\mathbf{1}) = 1.$$

These $\mathcal{V}$ and $\mathcal{V}^{-1}$ are each other's inverses:

$$(\mathcal{V}^{-1} \circ \mathcal{V})(\omega) = \sum_{x \in X} \mathcal{V}(\omega)(\mathbf{1}_x) \big|x\rangle = \sum_{x \in X} (\omega \models \mathbf{1}_x) \big|x\rangle = \sum_{x \in X} \omega(x) \big|x\rangle = \omega$$

$$(\mathcal{V} \circ \mathcal{V}^{-1})(\phi)(p) = \Big(\sum_{x \in X} \phi(\mathbf{1}_x) \big|x\rangle\Big) \models p = \sum_{x \in X} \phi(\mathbf{1}_x) \cdot p(x) = \phi\Big(\bigovee_{x \in X} p(x) \cdot \mathbf{1}_x\Big) = \phi(p).$$

In the last line we use that ϕ is a homomorphism of effect modules and that the predicate p has a normal form $\bigovee_x p(x) \cdot \mathbf{1}_x$, see Lemma 4.4.3 (2).

2 Naturality for the representation isomorphism $\mathcal{V}$ in (6.30) follows from Proposition 4.5.3: for a channel $c\colon X \rightsquigarrow Y$,

$$\begin{aligned}\big((h \mapsto h \circ c^*) \circ \mathcal{V}\big)(\omega)(q) \;=\; \mathcal{V}(\omega)(c^*(q)) &= \omega \models c^*(q) \\ &= c_*(\omega) \models q \\ &= \mathcal{V}(c_*(\omega))(q) \;=\; \big(\mathcal{V} \circ c_*\big)(\omega)(q). \qquad \square\end{aligned}$$

There are two leading interpretations of probability, namely a frequentist interpretation and a Bayesian interpretation. The frequentist approach treats probability distributions as records of probabilities of occurrences, obtained from long-term accumulations, *e.g.* from frequentist learning. One can associate this view with the set $\mathcal{D}(X)$ of distributions on the left-hand-side of the representation isomorphism (6.30). The right-hand-side fits the Bayesian view, which focuses on assigning probabilities to predicates, see de Finetti (2017), in this situation via special functions $Pred(X) \to [0,1]$ that preserve the effect module structure. The representation isomorphism demonstrates that the frequentist and Bayesian interpretations of probability are thightly connected, at least in finite discrete probability. We shall see a similar isomorphism below, in Theorem 6.10.6, for infinite discrete probability. The situation for continuous probability will be described in Theorem **??**.

But first we look at how operations on distributions, like marginalisation, tensor product, and updating work accross the above representation isomorphism (6.30). The next result establishes a close connection between operations on distributions and logical operations, like between marginalisation and weakening. In the case of updating (conditioning) of a distribution, we see that the corresponding logical formulation is the common formulation of conditonal probability as a fraction of probability assignments. Not suprisingly, the connection relies on Bayes' rule.

Proposition 6.10.2 *Fix finite sets X, Y.*

1 *Let $\phi\colon Pred(X \times Y) \to [0,1]$ be a map of effect modules. Define first and second marginals $\phi[1,0]\colon Pred(X) \to [0,1]$ and $\phi[0,1]\colon Pred(Y) \to [0,1]$ of ϕ as:*

$$\phi[1,0](p) \coloneqq \phi(p \otimes \mathbf{1}) \qquad \textit{and} \qquad \phi[0,1](q) \coloneqq \phi(\mathbf{1} \otimes q).$$

Then: $\mathcal{V}^{-1}(\phi[1,0]) = \mathcal{V}^{-1}(\phi)[1,0]$ and $\mathcal{V}^{-1}(\phi[0,1]) = \mathcal{V}^{-1}(\phi)[0,1]$.

2 *Let $\phi\colon Pred(X) \to [0,1]$ and $\nu\colon Pred(Y) \to [0,1]$ be maps of effect modules. We define their tensor product $\phi \otimes \nu\colon Pred(X \times Y) \to [0,1]$ as:*

$$(\phi \otimes \nu)(r) \coloneqq \phi\big(x \mapsto \nu(r(x,-))\big) = \nu\big(y \mapsto \phi(r(-,y))\big).$$

Then: $\mathcal{V}^{-1}(\phi \otimes \nu) = \mathcal{V}^{-1}(\phi) \otimes \mathcal{V}^{-1}(\nu)$.

3 *Let $\phi\colon Pred(X) \to [0,1]$ be a map of effect modules and let $p \in Pred(X)$ be a predicate with $\phi(p) \neq 0$. We define an update $\phi|_p\colon Pred(X) \to [0,1]$ as:*

$$\phi|_p(q) \coloneqq \frac{\phi(p \,\&\, q)}{\phi(p)}.$$

Then: $\mathcal{V}^{-1}(\phi|_p) = \mathcal{V}^{-1}(\phi)\big|_p$.

Proof. 1 We consider the first marginal only. It is easy to see that the map $\phi[1,0]\colon Pred(X) \to [0,1]$, as defined above, preserves the effect module structure. We get an equality of distributions since for $x \in X$,

$$\begin{aligned}
\mathcal{V}^{-1}(\phi[1,0])(x) \;=\; \phi[1,0](\mathbf{1}_x) &= \phi(\mathbf{1}_x \otimes \mathbf{1}) \\
&= \phi\Big(\mathbf{1}_x \otimes \bigvee_{y\in Y} \mathbf{1}_y\Big) \\
&= \sum_{y\in Y} \phi(\mathbf{1}_x \otimes \mathbf{1}_y) \\
&= \sum_{y\in Y} \phi(\mathbf{1}_{(x,y)}) \;=\; \sum_{y\in Y} \mathcal{V}^{-1}(\phi)(x,y) \;=\; \mathcal{V}^{-1}(\phi)[1,0](x).
\end{aligned}$$

2 For elements $a \in X$ and $b \in Y$ we have:

$$\begin{aligned}
\mathcal{V}^{-1}(\phi \otimes \nu)(a,b) \;=\; (\phi \otimes \nu)(\mathbf{1}_{(a,b)}) &= \phi(x \mapsto \nu(\mathbf{1}_{(a,b)}(x,-))) \\
&= \phi(x \mapsto \nu(\mathbf{1}_a(x) \cdot \mathbf{1}_b)) \\
&= \phi(x \mapsto \mathbf{1}_a(x) \cdot \nu(\mathbf{1}_b)) \\
&= \phi(k(\mathbf{1}_b) \cdot \mathbf{1}_a) \\
&= \nu(\mathbf{1}_b) \cdot \phi(\mathbf{1}_a) \\
&= \mathcal{V}^{-1}(\phi)(a) \cdot \mathcal{V}^{-1}(\nu)(b) \;=\; \big(\mathcal{V}^{-1}(\phi) \otimes \mathcal{V}^{-1}(\nu)\big)(a,b).
\end{aligned}$$

The same can be shown for the other formulation of $\phi \otimes k$ in item (2).

3 Again, $\phi|_p\colon Pred(X) \to [0,1]$ is a map of effect modules. Next, for $x \in X$,

$$\begin{aligned}
\mathcal{V}^{-1}(\phi|_p)(x) \;=\; \phi|_p(\mathbf{1}_x) \;=\; \frac{\phi(p \otimes \mathbf{1}_x)}{\phi(p)} &= \frac{\mathcal{V}^{-1}(\phi) \models p \otimes \mathbf{1}_x}{\mathcal{V}^{-1}(\phi) \models p} \\
&= \mathcal{V}^{-1}(\phi)\big|_p \models \mathbf{1}_x \quad \text{by Bayes, see Theorem 6.1.4 (2)} \\
&= \mathcal{V}^{-1}(\phi)\big|_p(x).
\end{aligned}$$

❑

The above results make it possible to do probability theory in a distribution-free style, with only predicates and predicate transformers (effect module maps). We illustrate how this works using an earlier example.

Example 6.10.3 We illustrate the purely Bayesian approach, on the right-hand-side of the isomorphism (6.30), for the Medicine-Blood updates from Example 6.2.6. There, we had a prior medicine distribution $\omega = \frac{3}{20}|0\rangle + \frac{9}{20}|1\rangle + \frac{2}{5}|2\rangle$, which we now replace by the predicate evaluation $\Omega := \mathcal{V}(\omega)\colon Pred(M) \to [0,1]$, where $M = \{0,1,2\}$ is the set of medicines. Thus:

$$\Omega(p) \;:=\; \mathcal{V}(\omega)(p) \;=\; \tfrac{3}{20} \cdot p(0) + \tfrac{9}{20} \cdot p(1) + \tfrac{2}{5} \cdot p(2).$$

It can be updated with the sharp predicate $\mathbf{1}_E \in Pred(M)$, where $E = \{1,2\} \subseteq M$ gives a restriction to medicines $1,2$. Via the update formula of Proposition 6.10.2 (3) we get a new predicate evaluation $\Omega|_{\mathbf{1}_E}\colon Pred(M) \to [0,1]$, namely:

$$\Omega|_{\mathbf{1}_E}(p) \;=\; \frac{\Omega(p \,\&\, \mathbf{1}_E)}{\Omega(\mathbf{1}_E)} \;=\; \frac{{}^{9}/_{20} \cdot p(1) + {}^{2}/_{5} \cdot p(2)}{{}^{9}/_{20} + {}^{2}/_{5}} \;=\; \tfrac{9}{17} \cdot p(1) + \tfrac{8}{17} \cdot p(2).$$

It corresponds to the updated distribution $\omega|_{\mathbf{1}_E} = \frac{9}{17}|1\rangle + \frac{8}{17}|2\rangle$ that we obtained in Example 6.2.6.

Next we recall from Example 6.2.6 the channel $b\colon M \rightsquigarrow B$, where $B = \{H, L\}$ the set for

high / low blood pressure. This channel is formulated in terms of distributions (on B), which we now like to avoid. We recall from Exercise 4.5.12 that such a channel bijectively corresponds to a predicate transformer operation $\beta = b^* \colon Pred(B) \to Pred(M)$, preserving the effect module structure of predicates. This map β puts channels 'on the Bayesian' side, in the world of predicates. Explicitly, for $q \in Pred(B)$, it is defined as:

$$\beta(q)(0) \coloneqq b^*(q)(0) = b(0)(H) \cdot q(H) + b(0)(L) \cdot q(L) = \tfrac{2}{3} \cdot q(H) + \tfrac{1}{3} \cdot q(L).$$

Similarly, $\beta(q)(1) \coloneqq \frac{7}{9} \cdot q(H) + \frac{2}{9} \cdot q(L)$ and $\beta(q)(2) \coloneqq \frac{5}{8} \cdot q(H) + \frac{3}{8} \cdot q(L)$.

We are interested in learning the new predicate evaluation (on M) when we have evidence of a high blood pressure. We then update Ω with $\beta(\mathbf{1}_H)$, giving as new predicate evaluation:

$$\begin{aligned}
&\Omega|_{\beta(\mathbf{1}_H)}(p) \\
&\qquad = \frac{\Omega(p \mathbin{\&} \beta(\mathbf{1}_H))}{\Omega(\beta(\mathbf{1}_H))} \\
&\qquad = \frac{{}^3/_{20} \cdot \beta(\mathbf{1}_H)(0) \cdot p(0) + {}^9/_{20} \cdot \beta(\mathbf{1}_H)(1) \cdot p(1) + {}^2/_5 \cdot \beta(\mathbf{1}_H)(2) \cdot p(2)}{{}^3/_{20} \cdot \beta(\mathbf{1}_H)(0) + {}^9/_{20} \cdot \beta(\mathbf{1}_H)(1) + {}^2/_5 \cdot \beta(\mathbf{1}_H)(2)} \\
&\qquad = \frac{{}^3/_{20} \cdot {}^2/_3 \cdot p(0) + {}^9/_{20} \cdot {}^7/_9 \cdot p(1) + {}^2/_5 \cdot {}^5/_8 \cdot p(2)}{{}^3/_{20} \cdot {}^2/_3 + {}^9/_{20} \cdot {}^7/_9 + {}^2/_5 \cdot {}^5/_8} \\
&\qquad = \tfrac{1}{7} \cdot p(0) + \tfrac{1}{2} \cdot p(1) + \tfrac{5}{14} \cdot p(2).
\end{aligned}$$

This predicate evaluation corresponds, via the isomorphism (6.30), to the updated distribution $\omega|_{b^*(\mathbf{1}_H)} = \frac{1}{7}|0\rangle + \frac{1}{2}|1\rangle + \frac{5}{14}|2\rangle$ computed in Example 6.2.6.

We finish this section with an extension of the correspondence (6.30) between distributions in $\mathcal{D}(X)$ and predicate evaluations $Pred(X) \to [0, 1]$ to distributions in $\mathcal{D}_\infty(X)$, with countable support. This involves effect algebras / modules with countable joins (least upper bounds). We shall use the terms ∞-joins and ∞-continuity for what are commonly called ω-joins and ω-continuity. Since we often use the Greek letter ω for states / distributions, we choose to avoid using it also for the first infinite ordinal ω in these expressions. The use of the sign ∞ for the countable case is in line with the notation $\mathcal{D}_\infty(X)$ for the set of distribution with countable support.

Posets with countable join are basic structures in the semantics of programming languages, see Plotkin (1983); Abramsky and Jung (1994), since they provide least fixed points of continuous functions, see Exercise 6.10.4. Instead of requiring joins of countable chains, one can also use joins of directed sets.

Definition 6.10.4 Consider a poset $(A, \leq)$.

1 A sequence (or chain) of elements $a_n \in A$, for $n \in \mathbb{N}$, is called *ascending* if $a_n \leq a_{n+1}$ for each n.
2 The poset $(A, \leq)$ will be called an ∞-cpo, or, in words, *a countably complete partial order*, if each ascending chain (a_n) in A has a least upperbound (join) $\bigvee_n a_n \in A$.
3 A monotone function $f \colon A \to B$ between two ∞-complete posets A, B is called ∞-*continuous* if it preserves joins of ascending chains, that is, if $f(\bigvee_n a_n) = \bigvee_n f(a_n)$, for each ascending chain (a_n) in A.
4 We shall write ∞-**Cpo** for the category of ∞-cpo's with ∞-continuous (and, implicitly, monotone) functions between them.

These countable joins can be combined with effect algebra / module structure, introduced in Definition 4.4.1, see *e.g.* Gudder (1998); Westerbaan et al. (2020). Later on, in Section **??** we shall see how these effect algebras with joins capture the order-theoretic essentials of so-called σ-algebras of measurable subsets.

Definition 6.10.5 1 An ∞-*effect algebra* is an effect algebra which is ∞-complete as a poset, using the order described in Exercise 4.4.15. From the same exercise we know that the sum $\ovee$ automatically preserves joins in each argument.

We shall write ∞-**EA** for the category with ∞-effect algebras as objects, and with ∞-continuous effect algebra maps as morphisms.

2 Similarly, an ∞-*effect module* is an effect module E which is ∞-complete as poset, and whose scalar multiplication $[0,1] \times E \to E$ is ∞-continuous in each argument.

The category ∞-**EMod** contains such ∞-effect modules, with ∞-continous effect module maps between them.

In the last point we implicitly use that the unit interval $[0,1]$ is an ∞-cpo. This is obvious.

For each set X, the powerset $\mathcal{P}(X)$ is an effect algebra with arbitrary joins, so in particular ∞-joins. The unit interval $[0,1]$ is an ∞-effect module, and more generally, each set of predicates $Pred(X) = [0,1]^X$ is an ∞-effect module, via pointwise joins. Later on, in continuous probability, we shall see measurable spaces whose sets of measurable subsets form examples of ∞-effect algebras. There is a free completion construction that turns an arbitrary effect algebra into an ∞-effect algebra. Categorically this is formulated as: the inclusion functor ∞-**EA** $\to$ **EA** has a left adjoint, see (Lachman, 2024, Thm. 33).

We recall from Exercise 4.4.16 that the sum operation $\ovee$ of an ∞-effect algebra is continuous in both arguments separately. Explicitly: $\bigvee_n x \ovee y_n = x \ovee \bigvee_n y_n$. Further, if there are joins $\bigvee$, there are also meets $\bigwedge$ via orthosupplement.

Theorem 6.10.6 *For each countable set X there is a representation isomorphism,*

$$\mathcal{D}_\infty(X) \xrightarrow[\cong]{\mathcal{V}} \infty\text{-}\mathbf{EMod}\big(Pred(X),\ [0,1]\big) \tag{6.31}$$

also given by validity: $\mathcal{V}(\omega)(p) := \omega \models p$. *This isomorphism is natural in X, like in Theorem 6.10.1* (2).

Proof. Much of this works as in the proof of Theorem 6.10.1, except that we have to prove that $\mathcal{V}(\omega)\colon Pred(X) \to [0,1]$ is ∞-continuous, now that we have $\omega \in \mathcal{D}_\infty(X)$. So let (p_n) be an ascending chain of predicates, with pointwise join $p = \bigvee_n p_n$. Then:

$$\begin{aligned}
\mathcal{V}(\omega)\Big(\bigvee\nolimits_n p_n\Big) = \omega \models \bigvee\nolimits_n p_n &= \sum_{x \in X} \omega(x) \cdot \Big(\bigvee\nolimits_n p_n\Big)(x) \\
&= \sum_{x \in X} \omega(x) \cdot \Big(\bigvee\nolimits_n p_n(x)\Big) \\
&= \sum_{x \in X} \bigvee\nolimits_n \omega(x) \cdot p_n(x) \\
&\overset{(*)}{=} \bigvee\nolimits_n \sum_{x \in X} \omega(x) \cdot p_n(x) = \bigvee\nolimits_n \mathcal{V}(\omega)(p_n).
\end{aligned}$$

The direction $(\geq)$ of the marked equation $\overset{(*)}{=}$ holds by monotonicity. For $(\leq)$ we reason as follows.

Since X is countable, we can write it as $X = \{x_1, x_2, \ldots\}$. For each $N \in \mathbb{N}$ we can use that finite sums preserve ∞-joins in:

$$\sum_{k \leq N} \bigvee_n \omega(x_k) \cdot p_n(x_k) \;=\; \bigvee_n \sum_{k \leq N} \omega(x_k) \cdot p_n(x_k) \;\leq\; \bigvee_n \sum_{x \in X} \omega(x) \cdot p_n(x).$$

Hence:

$$\sum_{x \in X} \bigvee_n \omega(x) \cdot p_n(x) \;=\; \lim_{N \to \infty} \sum_{k \leq N} \bigvee_n \omega(x_k) \cdot p_n(x_k) \;\leq\; \bigvee_n \sum_{x \in X} \omega(x) \cdot p_n(x).$$

The inverse of $\mathcal{V}$ in (6.31) is defined as countable formal sum, using that X is countable: for a homomorphism of ∞-effect modules $\phi \colon Pred(X) \to [0,1]$ we take:

$$\mathcal{V}^{-1}(\phi) \;:=\; \sum_{x \in X} \phi(\mathbf{1}_x) \,\big|x\rangle.$$

In order to see that the probabilities $\phi(\mathbf{1}_x)$ add up to one, we write the set X as $X = \{x_1, x_2, \ldots\}$. We express the sum over these x_i via a join of an ascending chain:

$$\sum_{x \in X} \phi(\mathbf{1}_x) \;=\; \bigvee_n \sum_{k \leq n} \phi(\mathbf{1}_{x_k}) \;=\; \bigvee_n \phi\Big(\bigvee\!\!\!\!\bigcirc_{k \leq n} \mathbf{1}_{x_k}\Big) \;=\; \phi\Big(\bigvee_n \mathbf{1}_{\{x_1, \ldots, x_n\}}\Big) \;=\; \phi(\mathbf{1}) \;=\; 1.$$

The arguments that $\mathcal{V}$ and $\mathcal{V}^{-1}$ are each other's inverses, and that $\mathcal{V}$ is natural, work essentially in the same way as in the proof of Theorem 6.10.1. ❑

Exercises

6.10.1 Consider the representation isomorphism in Theorem 6.10.1.

1 Show that a uniform distribution corresponds to the mapping that sends a predicate to its average value.

2 Show that for finite sets X, Y and maps $\phi \in \mathbf{EMod}(Pred(X), [0,1])$ and $\nu \in \mathbf{EMod}(Pred(Y), [0,1])$ one has:

$$(\phi \otimes \nu)[1, 0] \;=\; \phi.$$

6.10.2 Recall from Theorem 4.4.5 that $Pred(X)$ is the free effect module on the effect algebra $\mathcal{P}(X)$, for a finite set X.

1 Deduce from this fact that there is an isomorphism:

$$\mathbf{EA}\big(\mathcal{P}(X),\ [0,1]\big) \;\cong\; \mathbf{EMod}\big(Pred(X),\ [0,1]\big).$$

Describe this isomorphism in detail.

2 Conclude that an alternative version of the representation isomorphism (6.30) is:

$$\mathcal{D}(X) \xrightarrow[\cong]{} \mathbf{EA}\big(\mathcal{P}(X),\ [0,1]\big)$$

6.10.3 Recall the Poisson distribution $pois[\lambda] = \sum_{k \in \mathbb{N}} e^{-\lambda} \cdot \frac{\lambda^k}{k!} \,|k\rangle \in \mathcal{D}_\infty(\mathbb{N})$ from (2.10). Describe the corresponding homomorphism of ∞-effect modules $Pred(\mathbb{N}) \to [0,1]$ via (6.31).

6.10.4 Let $(A, \leq)$ be an ∞-cpo.

1 Suppose that A has a least element $\mathbf{0} \in A$. For an ∞-continuous function $f\colon A \to A$ define:

$$\mathit{fix}(f) := \bigvee_{n\in\mathbb{N}} f^n(\mathbf{0}) \in A.$$

Check that this is well-defined, *i.e.* that the chain $(f^n(0))$ of iterated applications of f is ascending. Prove that $\mathit{fix}(f) \in A$ is the least fixed point of f, *i.e.* that: $f(\mathit{fix}(f)) = \mathit{fix}(f)$ and for any $x \in A$, if $f(x) = x$, then $\mathit{fix}(f) \leq x$.

2 Call a subset $U \subseteq A$ *Scott open* if U is an upset (*i.e.* $y \geq x \in U$ implies $y \in U$) and satisfies, for any ascending chain (a_n),

$$\bigvee_n a_n \in U \implies a_n \in U, \text{ for some } n \in \mathbb{N}.$$

Check that the Scott open subsets form a topology on A.

3 Let $f\colon A \to B$ be a monotone function, where B is also an ∞-cpo. Show that:

$$f \text{ is } \infty\text{-continuous} \iff f \text{ is topologically continuous},$$

where ∞-continuity means that f preserves ∞-joins (see Definition 6.10.4 (3)), and topological continuity means that $f^{-1}(V) \subseteq A$ is (Scott) open when $V \subseteq B$ is (Scott) open.

4 Consider the unit interval $[0, 1]$ as ∞-effect module, and thus implicitly as ∞-cpo. Check that the Scott open subsets are of the form $(r, 1] \subseteq [0, 1]$ for $r \geq 0$.

6.11 Updating and (co)variance

The final two section of this chapter contain miscellanious results about updating. First, the effect of updating in relation to (the sign of) covariance is described. This leads to several basic facts about updating and the natural logarithm $\ln(p)$ of a factor p. It results in upperbounds for the variance. The next section concentrates on update with e-powers e^p.

In Section 5.1 we stated informally that a positive covariance / correlation between two observables p, q means that they move together in the same direction. A negative covariance means that they move in opposite directions. The formulation below makes this more precise, when we restrict to factors. In words: when updating a distribution ω with p increases the validity of q, the covariance if positive, and when updating ω with p decreases the validity of q the covariance is negative.

Proposition 6.11.1 *Let $\omega \in \mathcal{D}(X)$ be a distribution with a factor $p\colon X \to \mathbb{R}_{\geq 0}$ having non-zero validity $\omega \models p$ and an observable $q\colon X \to \mathbb{R}$. Then:*

$$\begin{aligned} \mathit{Cov}(\omega, p, q) \geq 0 &\iff (\omega|_p \models q) \geq (\omega \models q) \\ \mathit{Cov}(\omega, p, q) \leq 0 &\iff (\omega|_p \models q) \leq (\omega \models q). \end{aligned} \tag{6.32}$$

Since $\mathrm{Cov}(\omega, p, q) = \mathrm{Cov}(\omega, q, p)$ there are dual formulations, when q is a factor with non-zero validity and p and observable.

Proof. Since:

$$\begin{aligned} \mathit{Cov}(\omega, p, q) &= (\omega \models p \mathbin{\&} q) - (\omega \models p) \cdot (\omega \models q) && \text{see Lemma 5.1.8} \\ &= (\omega \models p) \cdot \big((\omega|_p \models q) - (\omega \models q)\big) && \text{by Theorem 6.1.4 (1).} \end{aligned}$$

❑

We continue with general observations that provides a foundation for several subsequent validity increases and decreases.

Lemma 6.11.2 *Let $\omega \in \mathcal{D}(X)$ be a distribution with a factor $p\colon X \to \mathbb{R}_{\geq 0}$ that has a non-zero validity $\omega \models p$. Let $f\colon S \to \mathbb{R}$ be a function, defined on some subset $S \subseteq \mathbb{R}_{\geq 0}$ which includes the image of p, that is, with $\{p(x) \mid x \in X\} \subseteq S$.*

1 *If f is monotone / order-preserving / increasing, that is, $f(r) \leq f(s)$ holds for $r \leq s$ where $r, s \in S$, then there is an inequality:*

$$\omega|_p \models (f \circ p) \;\geq\; \omega \models (f \circ p).$$

The monotonicity requirement may be relaxed to:

$$\begin{cases} p(x) \leq (\omega \models p) \implies f(p(x)) \leq f(\omega \models p) \\ (\omega \models p) \leq p(x) \implies f(\omega \models p) \leq f(p(x)). \end{cases} \tag{6.33}$$

This clearly holds if f is monotone.

2 *If f is anti-monotone / order-reversing / decreasing, that is, $f(r) \leq f(s)$ holds for $r \geq s$, then there is an inequality:*

$$\omega|_p \models (f \circ p) \;\leq\; \omega \models (f \circ p).$$

The anti-monotonicity requirement may be relaxed to:

$$\begin{cases} p(x) \leq (\omega \models p) \implies f(p(x)) \geq f(\omega \models p) \\ (\omega \models p) \leq p(x) \implies f(\omega \models p) \geq f(p(x)). \end{cases} \tag{6.34}$$

Proof. 1 By Theorem 6.1.4 (1), the inequality $\omega|_p \models (f \circ p) \geq \omega \models (f \circ p)$ that we wish to prove is equivalent to:

$$\omega \models p \;\&\; (f \circ p) \geq (\omega \models p) \cdot (\omega \models (f \circ p)) = \omega \models (\omega \models p) \cdot \mathbf{1} \;\&\; (f \circ p).$$

By moving this right-hand-side to the left, our target becomes:

$$\omega \models \big(p - (\omega \models p) \cdot \mathbf{1}\big) \;\&\; (f \circ p) \;\geq\; 0. \tag{$*$}$$

We now define the following partition $X = X_- \cup X_+$.

$$X_- := \{x \in X \mid p(x) < (\omega \models p)\} \qquad X_+ := \{x \in X \mid p(x) \geq (\omega \models p)\}.$$

The relaxed property (6.33) of f can be reformulated as:

$$x \in X_- \;\Rightarrow\; -f(p(x)) \;\geq\; -f(\omega \models p) \qquad x \in X_+ \;\Rightarrow\; f(p(x)) \;\geq\; f(\omega \models p).$$

We can now prove the inequality $(*)$ as follows, where the inequality $\geq$ below uses these two implications.

$$\begin{aligned} &\omega \models \big(p - (\omega \models p) \cdot \mathbf{1}\big) \;\&\; (f \circ p) \\ &= \sum_{x \in X_+} \omega(x) \cdot \big(p(x) - (\omega \models p)\big) \cdot f(p(x)) - \sum_{x \in X_-} \omega(x) \cdot \big((\omega \models p) - p(x)\big) \cdot f(p(x)) \\ &\geq \sum_{x \in X_+} \omega(x) \cdot \big(p(x) - (\omega \models p)\big) \cdot f(\omega \models p) - \sum_{x \in X_-} \omega(x) \cdot \big((\omega \models p) - p(x)\big) \cdot f(\omega \models p) \\ &= f(\omega \models p) \cdot \sum_{x \in X} \omega(x) \cdot \big(p(x) - (\omega \models p)\big) \;=\; 0. \end{aligned}$$

2 By applying the previous result to $-f$. ❑

By taking f to be the identity function $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ one obtains an alternative proof of the validity increase of Theorem 6.1.6. But we can also obtain new results. When $p\colon X \to \mathbb{R}_{>0}$ is a non-zero factor, we write $\ln(p)\colon X \to \mathbb{R}$ for the observable given by the pointwise natural logarithm: $\ln(p)(x) := \ln\big(p(x)\big)$. In this context we recall the inequality $\ln\big(\omega \models p\big) \geq \omega \models \ln(p)$ from Exercise 4.1.6.

Proposition 6.11.3 *Fix a distribution $\omega \in \mathcal{D}(X)$ and a positive factor $p\colon X \to \mathbb{R}_{>0}$.*

1 *There is an inequality:*

$$\Big(\omega \models p\Big) \cdot \Big(\omega \models \frac{1}{p}\Big) \;\geq\; 1 \qquad \textit{and thus} \qquad \omega \models \frac{1}{p} \;\geq\; \frac{1}{\omega \models p}.$$

2 *This can be strengthened to the following inequality, for $q\colon X \to \mathbb{R}_{\geq 0}$ with non-zero validity $\omega \models q$.*

$$\Big(\omega \models q \mathbin{\&} p\Big) \cdot \Big(\omega \models q \mathbin{\&} \frac{1}{p}\Big) \;\geq\; \Big(\omega \models q\Big)^2.$$

3 *Updating with p gives a decrease of validity of $\frac{1}{p}$:*

$$\omega|_p \models \frac{1}{p} \;\leq\; \omega \models \frac{1}{p}.$$

By exchanging p and $\frac{1}{p}$ one immediately obtains: $\omega|_{\frac{1}{p}} \models p \;\leq\; \omega \models p$.

4 *When $p \leq \mathbf{1}$, so when p is a positive predicate,*

$$\omega|_p \models p + \frac{1}{p} \;\leq\; \omega \models p + \frac{1}{p}.$$

5 *Still when $p \leq \mathbf{1}$,*

$$\omega|_p \models p + \frac{\omega \models p}{p} \;\leq\; \omega \models p + \frac{\omega \models p}{p}.$$

6 *As a consequence, when $p \leq \mathbf{1}$, there is the following upperbound for the validity increase of Theorem 6.1.6,*

$$\omega|_p \models p \;-\; \omega \models p \;\leq\; \Big(\omega \models p\Big) \cdot \Big(\omega \models \frac{1}{p}\Big) - 1.$$

The right-hand-side is non-negative by item (1).

7 *Without restrictions on p,*

$$\omega|_p \models \ln(p) \;\geq\; \omega \models \ln(p).$$

8 *Again, with restriction $p \leq \mathbf{1}$,*

$$\omega|_p \models p - \ln(p) \;\leq\; \omega \models p - \ln(p).$$

By induction on $n \in \mathbb{N}$ we then get: $\omega|_{p^n} \models p - \ln(p) \;\leq\; \omega \models p - \ln(p)$.

Proof. 1 We use that the natural logarithm ln of the left-hand-side is above $0 = \ln(1)$, which suffices since the logartithm reflects the order.

$$\begin{aligned}
\ln\Big((\omega \models p) \cdot (\omega \models \tfrac{1}{p})\Big) &= \ln(\omega \models p) + \ln(\omega \models \tfrac{1}{p}) \\
&\geq (\omega \models \ln(p)) + (\omega \models \ln(\tfrac{1}{p})) && \text{by Jensen, see Exercise 4.1.6} \\
&= (\omega \models \ln(p)) + (\omega \models -\ln(p)) \\
&= \omega \models (\ln(p) - \ln(p)) \;=\; \omega \models \mathbf{0} \;=\; 0.
\end{aligned}$$

2 By Theorem 6.1.4 (1) and the previous point:

$$\Big(\omega \models q \;\&\; p\Big) \cdot \Big(\omega \models q \;\&\; \tfrac{1}{p}\Big) = \Big(\omega|_q \models p\Big) \cdot \Big(\omega|_q \models \tfrac{1}{p}\Big) \cdot \Big(\omega \models q\Big)^2 \geq \Big(\omega \models q\Big)^2.$$

3 The function $f(s) = \frac{1}{s}$ is anti-monotone, as function $f\colon \mathbb{R}_{>0} \to \mathbb{R}$. Hence, Lemma 6.11.2 (2) applies.

4 The function $f(s) = s + \frac{1}{s}$ is anti-monotone, as function $f\colon (0,1] \to \mathbb{R}$. Indeed, let $r \leq s$, for $r, s \in (0,1]$. Then:

$$\begin{aligned}
f(r) \geq f(s) \iff r + \frac{1}{r} \geq s + \frac{1}{s} &\iff \frac{s-r}{r \cdot s} = \frac{1}{r} - \frac{1}{s} \geq s - r \\
&\iff \frac{1}{r \cdot s} \geq 1 \iff r \cdot s \leq 1.
\end{aligned}$$

The latter clearly holds for $r, s \in (0,1]$. The assumption $s - r \geq 0$ is used for the third equivalence $\iff$, where it ensures that the order does not change when we divide by, or multiply with, $s - r$.

5 We now take $f(x) = s + \frac{1}{s} \cdot (\omega \models p)$ and show that the relaxed requirement (6.34) holds.

- Let $p(x) \leq (\omega \models p)$; then:

$$\begin{aligned}
f(p(x)) \geq f(\omega \models p) &\iff p(x) + \frac{\omega \models p}{p(x)} \geq (\omega \models p) + 1 \\
&\iff \frac{(\omega \models p) - p(x)}{p(x)} = \frac{\omega \models p}{p(x)} - 1 \geq (\omega \models p) - p(x) \\
&\iff \frac{1}{p(x)} \geq 1.
\end{aligned}$$

  The latter clearly holds since p is a predicate, so $p(x) \leq 1$.

- When $p(x) \geq (\omega \models p)$, we similarly get:

$$\begin{aligned}
f(p(x)) \leq f(\omega \models p) &\iff p(x) + \frac{\omega \models p}{p(x)} \leq (\omega \models p) + 1 \\
&\iff p(x) - (\omega \models p) \leq 1 - \frac{\omega \models p}{p(x)} = \frac{p(x) - (\omega \models p)}{p(x)} \\
&\iff 1 \leq \frac{1}{p(x)}.
\end{aligned}$$

6 We first rewrite the previous point and then use Theorem 6.1.4 (1) in:

$$\begin{aligned}
\omega|_p \models p \;-\; \omega \models p \;&=\; \omega \models \frac{\omega \models p}{p} \;-\; \omega|_p \models \frac{\omega \models p}{p} \\
&= \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) - \big(\omega \models p\big) \cdot \Big(\omega|_p \models \frac{1}{p}\Big) \\
&= \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) - \Big(\omega \models p \;\&\; \frac{1}{p}\Big) \\
&= \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) - \big(\omega \models \mathbf{1}\big) \;=\; \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) \;-\; 1.
\end{aligned}$$

7 Since the $\ln\colon \mathbb{R}_{>0} \to \mathbb{R}$ is monotone.

8 We use $f(s) = s - \ln(s)$ and note that $f\colon (0,1] \to \mathbb{R}$ is anti-monotone. Indeed, if $r \leq s$ for $r, s \in (0,1]$, then:

$$\begin{aligned}
f(r) \;\geq\; f(s) \;&\Longleftrightarrow\; r - \ln(r) \;\geq\; s - \ln(s) \\
&\Longleftrightarrow\; \ln\Big(\frac{s}{r}\Big) = \ln(s) - \ln(r) \;\geq\; s - r \\
&\Longleftarrow\; 1 - \frac{r}{s} \;\geq\; s - r \qquad \text{since } \ln(x) \geq 1 - \frac{1}{x}, \text{ see (2.77)} \\
&\Longleftrightarrow\; \frac{s-r}{s} = 1 - \frac{r}{s} \;\geq\; s - r \;\Longleftrightarrow\; \frac{1}{s} \;\geq\; 1 \;\Longleftrightarrow\; 1 \;\geq\; s. \qquad \square
\end{aligned}$$

We use these results to present some bounds for variances[5].

Proposition 6.11.4 *Let $\omega \in \mathcal{D}(X)$ be a distribution with a positive predicate $p\colon X \to (0,1]$.*

1 One has:

$$\mathrm{Var}(\omega, p) \;\leq\; \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) \;-\; 1.$$

2 This can be strengthened to, still for a positive predicate p,

$$\frac{\mathrm{Var}(\omega, p)}{\omega \models p} \;\leq\; \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) \;-\; 1.$$

3 One also has:

$$\mathrm{Var}(\omega, p) \;\leq\; \mathrm{Cov}(\omega, p, \ln(p)).$$

Proof. 1 We first rewrite the inequality $\omega|_p \models p + \frac{1}{p} \;\leq\; \omega \models p + \frac{1}{p}$ from Proposition 6.11.3 (4)

[5] After showing Proposition 6.11.4 (1), with a direct proof, to local colleagues, Dario Stein suggested a more general proof technique using the (anti-)monotone functions f, as in Lemma 6.11.2. Eric Cator came up with the strengthening in Proposition 6.11.4 (2), with a (different) proof that (also) works in the more general continuous setting.

as: $(\omega|_p \models p) - (\omega \models p) \leq (\omega \models \frac{1}{p}) - (\omega|_p \models \frac{1}{p})$. Then, via Theorem 6.1.4 (1),

$$\begin{aligned}
Var(\omega, p) \;=\; \big(\omega \models p^2\big) - \big(\omega \models p\big)^2 &= \big(\omega \models p\big) \cdot \big[(\omega|_p \models p) - (\omega \models p)\big] \\
&\leq \big(\omega \models p\big) \cdot \Big[\Big(\omega \models \frac{1}{p}\Big) - \Big(\omega|_p \models \frac{1}{p}\Big)\Big] \\
&= \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) - \Big(\omega \models p \;\&\; \frac{1}{p}\Big) \\
&= \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) - \big(\omega \models \mathbf{1}\big) \\
&= \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) - 1.
\end{aligned}$$

2 This is a direct consequence of Proposition 6.11.3 (6):

$$\frac{Var(\omega, p)}{\omega \models p} \;=\; \frac{(\omega \models p^2) - (\omega \models p)^2}{\omega \models p} \;=\; \big(\omega|_p \models p\big) - \big(\omega \models p\big) \leq \big(\omega \models p\big) \cdot \Big(\omega \models \frac{1}{p}\Big) - 1.$$

3 We now use Proposition 6.11.3 (8), with inequality $\omega|_p \models p - \ln(p) \;\leq\; \omega \models p - \ln(p)$ to get the required inequality:

$$\begin{aligned}
Var(\omega, p) \;=\; (\omega \models p^2) - (\omega \models p)^2 &= (\omega \models p) \cdot \big((\omega|_p \models p) - (\omega \models p)\big) \\
&\leq (\omega \models p) \cdot \big((\omega|_p \models \ln(p)) - (\omega \models \ln(p))\big) \\
&= \big(\omega \models p \;\&\; \ln(p)\big) - \big(\omega \models p\big) \cdot \big(\omega \models \ln(p)\big) \\
&= Cov(\omega, p, \ln(p)).
\end{aligned}$$

❑

Exercises

6.11.1 Let $\omega \in \mathcal{D}(X)$ and $p, q \colon X \to \mathbb{R}_{\geq 0}$, both with non-zero validity. Prove the following symmetry of validity increases.

$$(\omega|_p \models q) \;\geq\; (\omega \models q) \iff (\omega|_q \models p) \;\geq\; (\omega \models p).$$

6.11.2 Let $\omega \in \mathcal{D}(X)$ and $p \colon X \to \mathbb{R}_{\geq 0}$.

1 Show that for a function f as in Lemma 6.11.2 (1) one has:

$$Cov(\omega,\; p,\; f \circ p) \;\geq\; 0.$$

2 And for f as in Lemma 6.11.2 (2),

$$Cov(\omega,\; p,\; f \circ p) \;\leq\; 0.$$

6.11.3 Generalise Proposition 6.11.3 (1) to the following inequalities, for $n, m \in \mathbb{N}$,

1 $(\omega \models p^{n+m}) \cdot (\omega \models \frac{1}{p^n}) \;\geq\; \omega \models p^m$;

2 $(\omega \models p^n) \cdot (\omega \models \frac{1}{p^{n+m}}) \;\geq\; \omega \models \frac{1}{p^m}$

6.11.4 Find a counterexample showing that Proposition 6.11.3 (1) does *not* generalise to: $(\omega \models p) \cdot (\omega \models \frac{q}{p}) \geq (\omega \models q)$.

6.11.5 Prove that for $\omega \in \mathcal{D}(X)$ and $p \in Fact_*(X)$,

$$\big(\omega \models p\big) \cdot \big(\omega|_{\frac{1}{p}} \models \tfrac{1}{p}\big) \;\geq\; 1.$$

6.12 Updating with e-powers of observables

For an observable $p\colon X \to \mathbb{R}$ we can form the positive factor $e^p\colon X \to \mathbb{R}_{>0}$ given pointwise by $(e^p)(x) = e^{p(x)}$. Updating with such e-power factors e^p has several interesting properties and applications. This section gives an impression, showing for instance that updating with p works better than with e^p, when p is a predicate. It is shown how updating with an e-power is used to approximate the capacity of a channel.

To start, updating with e-powers forms an action for the (additive) group of observables, in analogue with the update action for the (multiplicative) group of positive factors in Proposition 6.1.10. Recall from (6.4) the isomorphism (of groups) between the additive group $Obs(X)$ and the multiplicative group $Fact_*(X)$.

Proposition 6.12.1 *The update assignment $(p, \omega) \mapsto \omega|_{e^p}$ forms an additive group action $upd\colon Obs(X) \times \mathcal{D}(X) \to \mathcal{D}(X)$.*

Proof. We use the fact that powers transform sums into multiplications and also Lemma 6.1.7 (3) and (1) in:

$$\begin{aligned} upd(p+q, \omega) &= \omega|_{e^{p+q}} = \omega|_{e^q \,\&\, e^p} = \omega|_{e^q}|_{e^p} = upd(p, upd(q, \omega)) \\ upd(\mathbf{0}, \omega) &= \omega|_{e^{\mathbf{0}}} = \omega|_{\mathbf{1}} = \omega. \end{aligned}$$

❑

With this group action e-updates with p can be undone via $-p$, as in:

$$\big(\omega|_{e^p}\big)\big|_{e^{-p}} = \omega|_{e^{p-p}} = \omega|_{e^{\mathbf{0}}} = \omega|_{\mathbf{1}} = \omega.$$

Updating with e-powers e^p has possible advantages.

- It applies to all observables, with outcomes in $\mathbb{R}$, and not just to factors, with outcomes in the non-negative reals $\mathbb{R}_{\geq 0}$, as used in the Bayesian update $\omega|_p$ from Definition 6.1.1.
- The validity $\omega \models e^p$ used for normalisation in $\omega|_{e^p}$ is never zero, since $e^p > \mathbf{0}$. This means that any worries about dividing by zero or positivity of factors disappear: the mapping $\omega \mapsto \omega|_{e^p}$ is a total function $\mathcal{D}(X) \to \mathcal{D}(X)$.
- The update action of Proposition 6.12.1 works additively.

A disadvantage of updating with e^p is that the validity increase of a predicate p is not as much as in updating with p, see Theorem 6.12.4 below. Thus one could refer to updating with e^p as slow(er) updating.

The following result (from Donsker and Varadhan (1983)) is known as the Donsker-Varadhan duality formula for KL-divergence. It is comparable to the different formulations for the Kantorovich distance between distributions on metric spaces, see *e.g.* Kantorovich and Rubinshtein (1958); Villani (2009); Jacobs (2025c). The proof uses an update with an exponent.

Proposition 6.12.2 *The KL-divergence between distributions $\omega, \rho \in \mathcal{D}(X)$ can be expressed as join over observable:*

$$D_{KL}(\omega, \rho) = \bigvee_{p\colon X \to \mathbb{R}} \big(\omega \models p\big) - \ln\big(\rho \models e^p\big).$$

Proof. We may assume $supp(\omega) \subseteq supp(\rho) = X$, see Definition 2.10.1. For the inequality ($\geq$),

consider for an arbitrary observable $p\colon X \to \mathbb{R}$ the distribution $\rho|_{e^p}$, involving an update with the exponent factor e^p. It satisfies:

$$\begin{aligned}
0 &\leq D_{KL}(\omega, \rho|_{e^p}) \\
&= \sum_{x \in X} \omega(x) \cdot \ln\left(\frac{\omega(x) \cdot (\rho \models e^p)}{\rho(x) \cdot e^p(x)}\right) \\
&= \sum_{x \in X} \omega(x) \cdot \ln\left(\frac{\omega(x)}{\rho(x)}\right) + \sum_{x \in X} \omega(x) \cdot \ln\big(\rho \models e^p\big) - \sum_{x \in X} \omega(x) \cdot \ln\big(e^{p(x)}\big) \\
&= D_{KL}(\omega, \rho) + \ln\big(\rho \models e^p\big) - \big(\omega \models p\big).
\end{aligned}$$

For the inequality ($\leq$) consider the observable $p\colon X \to \mathbb{R}$ defined as:

$$p(x) := \ln\left(\frac{\omega(x)}{\rho(x)}\right).$$

Then:

$$\begin{aligned}
\big(\omega \models p\big) - \ln\big(\rho \models e^p\big) &= \sum_{x \in X} \omega(x) \cdot \ln\left(\frac{\omega(x)}{\rho(x)}\right) - \ln\left(\sum_{x \in X} \rho(x) \cdot \frac{\omega(x)}{\rho(x)}\right) \\
&= D_{KL}(\omega, \rho) - \ln(1) = D_{KL}(\omega, \rho).
\end{aligned}$$

❑

We first collect some useful properties.

Lemma 6.12.3 *Let $\omega \in \mathcal{D}(X)$ be a distribution with a factor $p\colon X \to \mathbb{R}_{\geq 0}$. We collect some properties of the update $\omega|_{e^p} \in \mathcal{D}(X)$.*

1 *The validity of the e-power factor can be described as infinite sum:*

$$\omega \models e^p = \sum_{n \in \mathbb{N}} \frac{\omega \models p^n}{n!} = \sum_{n \in \mathbb{N}} \frac{\mathrm{mom}[n](\omega, p)}{n!},$$

where $\mathrm{mom}[n](\omega, p)$ is the n-moment, see Definition 5.5.1. As a result:

$$\omega \models e^p \,\&\, q = \sum_{n \in \mathbb{N}} \frac{\omega \models p^n \,\&\, q}{n!}. \qquad \omega \models e^p \,\&\, p^m = \sum_{n \in \mathbb{N}} \frac{\omega \models p^{n+m}}{n!}.$$

2 *For all $m, k \in \mathbb{N}$ there is an inequality between validities:*

$$\omega \models e^p \,\&\, p^{m+k} \geq (\omega \models p^m) \cdot (\omega \models e^p \,\&\, p^k).$$

3 *There are validity increases:*

$$\omega|_{p^m} \models e^p \geq \omega \models e^p \qquad \textit{and so} \qquad \omega|_{e^p} \models p^m \geq \omega \models p^m.$$

In particular, for $m = 1$, the last inequality tells that updating ω with e^p makes p 'more true'.

4 *If there is a bound b with $\omega|_{p^n} \models q \leq b$, for all $n \in \mathbb{N}$, then $\omega|_{e^p} \models q \leq b$. The same holds of $\geq$.*

5 *In the special case when p is a positive predicate,*

$$\omega|_{e^p} \models p - \ln(p) \leq \omega \models p - \ln(p).$$

6 *Without any restriction on the observable p,*

$$\ln(\omega \models e^p) - \ln(\omega \models p) \leq \omega|_{e^p} \models p - \ln(p).$$

Proof. 1 We use the familiar equation $e^a = \sum_n \frac{a^n}{n!}$ to prove the most general formulation:

$$\begin{aligned}\omega \models e^p \,\&\, q \;=\; \sum_{x\in X} \omega(x)\cdot e^{p(x)}\cdot q(x) \;&=\; \sum_{x\in X} \omega(x)\cdot\left(\sum_{n\in\mathbb{N}} \frac{p(x)^n}{n!}\right)\cdot q(x)\\ &=\; \sum_{n\in\mathbb{N}}\sum_{x\in X} \omega(x)\cdot\frac{p(x)^n}{n!}\cdot q(x)\\ &=\; \sum_{n\in\mathbb{N}} \frac{\omega\models p^n \,\&\, q}{n!}.\end{aligned}$$

2 We use the previous item twice in:

$$\begin{aligned}\omega \models e^p \,\&\, p^{m+k} \;&=\; \sum_{n\in\mathbb{N}} \frac{\omega\models p^{n+m+k}}{n!}\\ &\geq\; \sum_{n\in\mathbb{N}} \frac{(\omega\models p^m)\cdot(\omega\models p^{n+k})}{n!} \qquad \text{by Proposition 6.1.8 (3)}\\ &=\; (\omega\models p^m)\cdot(\omega\models e^p \,\&\, p^k).\end{aligned}$$

3 Since:

$$\begin{aligned}\omega|_{p^m} \models e^p \;&=\; \frac{\omega\models p^m \,\&\, e^p}{\omega\models p^m} \qquad \text{by Theorem 6.1.4 (1)}\\ &\geq\; \frac{(\omega\models p^m)\cdot(\omega\models e^p)}{\omega\models p^m} \qquad \text{by item (2)}\\ &=\; \omega\models e^p.\end{aligned}$$

The second inequality follows from Exercise 6.11.1, but there is also an easy direct argument:

$$\omega|_{e^p} \models p^m \;=\; \frac{\omega\models e^p \,\&\, p^m}{\omega\models e^p} \;\geq\; \frac{(\omega\models e^p)\cdot(\omega\models p^m)}{\omega\models e^p} \;=\; \omega\models p^m.$$

4 By Theorem 6.1.4 (1) the assumption $\omega|_{p^n} \models q \leq b$ can be rewritten as:

$$\omega\models p^n \,\&\, q \;\leq\; (\omega\models p^n)\cdot b.$$

But then, by item (1),

$$\begin{aligned}\omega \models e^p \,\&\, q \;=\; \sum_{n\in\mathbb{N}} \frac{\omega\models p^n \,\&\, q}{n!} \;&\leq\; \sum_{n\in\mathbb{N}} \frac{(\omega\models p^n)\cdot b}{n!}\\ &=\; b\cdot\sum_{n\in\mathbb{N}} \frac{\omega\models p^n}{n!} \;=\; b\cdot(\omega\models e^p)\end{aligned}$$

Hence:

$$\omega|_{e^p} \models q \;=\; \frac{\omega\models e^p \,\&\, q}{\omega\models e^q} \;\leq\; b.$$

5 Via the previous point and Proposition 6.11.3 (8).

6 The log-sum inequality of Exercise 2.10.8 gives:

$$\begin{aligned}(\omega\models e^p)\cdot\ln\left(\frac{\omega\models e^p}{\omega\models p}\right) \;&\leq\; \sum_{x\in X} \omega(x)\cdot e^{p(x)}\cdot\ln\left(\frac{\omega(x)\cdot e^{p(x)}}{\omega(x)\cdot p(x)}\right)\\ &=\; \sum_{x\in X} \omega(x)\cdot e^{p(x)}\cdot\Big(p(x)-\ln\big(p(x)\big)\Big)\\ &=\; \omega\models e^p \,\&\, (p-\ln(p)).\end{aligned}$$

As a result,

$$\begin{aligned}\ln(\omega \models e^p) \;-\; \ln(\omega \models p) &\leq \frac{\omega \models e^p \;\&\; (p - \ln(p))}{\omega \models e^p} \\ &= \omega|_{e^p} \models p - \ln(p). \qquad \square\end{aligned}$$

Updating with e^p increases the validity of an observable p, as we have seen in Lemma 6.12.3 (3) (for $m = 1$). The result below says that in the special where p is a predicate, updating with p outperforms updating with e^p, in the sense that it leads to a bigger increase of validity of p. Thus, with e-powers of predicates we do 'slow' updating. The second point shows that the divergence between the update with e^p and the original distribution is less than the divergence between the update with p and the original.

Theorem 6.12.4 *Let $\omega \in \mathcal{D}(X)$ be a distribution with a predicate $p\colon X \to [0,1]$.*

1 There are the following inequalities between validities:

$$\omega|_p \models p \;\geq\; \omega|_{e^p} \models p \;\geq\; \omega \models p.$$

2 There are also inequalities between divergences, in symmetric form:

$$D_{KL}(\omega|_p, \omega) \;\geq\; D_{KL}(\omega|_{e^p}, \omega) \qquad D_{KL}(\omega, \omega|_p) \;\geq\; D_{KL}(\omega, \omega|_{e^p}).$$

Proof. 1 We only prove the inequality $\omega|_p \models p \geq \omega|_{e^p} \models p$ on the left, since the inequality $\omega|_{e^p} \models p \geq \omega \models p$ on the right is a special case of Lemma 6.12.3 (3), for $m = 1$. We first show that for an arbitrary distribution σ and predicate p,

$$\sigma|_p \models p \;\&\; e^{-p} \;\geq\; \sigma \models p \;\&\; e^{-p}. \tag{a}$$

This holds by Lemma 6.11.2 (1), via the function $f(r) = r \cdot e^{-r}$. Its derivative $f'(r) = e^{-r} - r \cdot e^{-r} = (1 - r) \cdot e^{-r}$ is non-negative on $[0,1]$, so that f is increasing on $[0,1]$.

From (a) one gets the inequality below, via Exercise 6.11.1. We can also derive it explicitly via Bayes' rule, see Theorem 6.1.4 (2):

$$\sigma|_{p \,\&\, e^{-p}} \models p \;=\; \frac{(\sigma|_p \models p \;\&\; e^{-p}) \cdot (\sigma \models p)}{\sigma \models p \;\&\; e^{-p}} \;\geq\; \sigma \models p. \tag{b}$$

We can now prove the required inequality, via Lemma 6.1.7 (1) and (3), and by instantiating (b) with $\sigma = \omega|_{e^p}$, in:

$$\omega|_p \models p \;=\; \omega|_{e^p}|_{e^{-p}}|_p \models p \;=\; \omega|_{e^p}|_{p \,\&\, e^{-p}} \models p \;\geq\; \omega|_{e^p} \models p.$$

2 For the first inequality, we first unfold the relevant divergencies:

$$\begin{aligned}
D_{KL}(\omega|_p, \omega) &= \sum_{x \in X} \omega|_p(x) \cdot \ln\left(\frac{\omega|_p(x)}{\omega(x)}\right) \\
&= \sum_{x \in X} \omega|_p(x) \cdot \ln\left(\frac{p(x)}{\omega \models p}\right) \\
&= \sum_{x \in X} \omega|_p(x) \cdot \ln(p(x)) - \sum_{x \in X} \omega|_p(x) \cdot \ln(\omega \models p) \\
&= (\omega|_p \models \ln(p)) - \ln(\omega \models p) \\
D_{KL}(\omega|_{e^p}, \omega) &= \sum_{x \in X} \omega|_{e^p}(x) \cdot \ln\left(\frac{e^p(x)}{\omega \models e^p}\right) \\
&= \sum_{x \in X} \omega|_{e^p}(x) \cdot \ln(e^{p(x)}) - \sum_{x \in X} \omega|_p(x) \cdot \ln(\omega \models e^p) \\
&= (\omega|_{e^p} \models p) - \ln(\omega \models e^p).
\end{aligned}$$

By combining Lemma 6.12.3 (6) with the previous point we are done:

$$\begin{aligned}
\ln(\omega \models e^p) - \ln(\omega \models p) &\geq (\omega|_p \models p) - (\omega|_p \models \ln(p)) \\
&\geq (\omega|_{e^p} \models p) - (\omega|_p \models \ln(p)).
\end{aligned}$$

Thus:

$$\begin{aligned}
D_{KL}(\omega|_{e^p}, \omega) = (\omega|_{e^p} \models p) - \ln(\omega \models e^p) &\leq (\omega|_p \models \ln(p)) - \ln(\omega \models p) \\
&= D_{KL}(\omega|_p, \omega).
\end{aligned}$$

For the second inequality in item (2) we first note that:

$$\begin{aligned}
D_{KL}(\omega, \omega|_p) &= \sum_{x \in X} \omega(x) \cdot \ln\left(\frac{\omega(x) \cdot (\omega \models p)}{\omega(x) \cdot p(x)}\right) \\
&= \ln(\omega \models p) - (\omega \models \ln(p)) \\
D_{KL}(\omega, \omega|_{e^p}) &= \sum_{x \in X} \omega(x) \cdot \ln\left(\frac{\omega(x) \cdot (\omega \models e^p)}{\omega(x) \cdot e^{p(x)}}\right) \\
&= \ln(\omega \models e^p) - (\omega \models p).
\end{aligned}$$

Thus, the claimed inequality $D_{KL}(\omega, \omega|_p) \geq D_{KL}(\omega, \omega|_{e^p})$ amounts to:

$$\ln(\omega \models p) - (\omega \models \ln(p)) \geq \ln(\omega \models e^p) - (\omega \models p).$$

The combination of Lemma 6.12.3 (5) and (6) yields:

$$\ln(\omega \models e^p) - \ln(\omega \models p) \leq \omega \models p - \ln(p).$$

This does the job. ❑

Remark 6.12.5 In this section we have seen various results that involved a predicate, and not a more general factor. There is a simple example that demonstrates the necessity of this restriction to predicates. Consider

$$\omega = \tfrac{2}{5}|a\rangle + \tfrac{3}{5}|b\rangle \qquad \text{and} \qquad p = 2 \cdot \mathbf{1}_a + \mathbf{1}_b.$$

Clearly, p is not a predicate. With this pair (ω, p),

- all variance upperbounds in Proposition 6.11.4 fail;

- all inequlities comparing $\omega|_p$ and $\omega|_{e^p}$ in Theorem 6.12.4 fail.

6.12.1 The capacity of a channel

The capacity of a channel is an associated number in $\mathbb{R}_{\geq 0}$. It is the theoretical maximum rate at which information can be reliably transmitted over the channel. This number is expressed as the maximum mutual information, where the maximum is obtained by ranging over distributions on the domain of the channel. The mutual information is a certain Kullback-Leibler divergence. We quote from (Cover and Thomas, 2006, p. 183): "The characterization of the channel capacity (the logarithm of the number of distinguishable signals) as the maximum mutual information is the central and most famous success of information theory." This subsection considers the calculation of this channel capacity as a limit of distribution updates, following Blahut (1972) and Arimoto (1972). "In general, there is no closed-form solution for the capacity." (Cover and Thomas, 2006, p. 191). In this sequence of approximations the update with an e-power plays an important role.

The capacity of a channel is usually expressed in terms of the logarithm with base 2. Since this book standardly uses the natural logarithm (with base e), we shall do so here as well. This makes it possible to directly use earliers results, as in the following definition.

Definition 6.12.6 Let $c\colon X \to \mathcal{D}(Y)$ be channel, where X is a finite set. Its capacity $Cap(c) \in \mathbb{R}_{\geq 0}$ is the join given below over a mutual information expression that can be expanded as follows.

$$\begin{aligned} Cap(c) &= \bigvee_{\omega \in \mathcal{D}(X)} MI\Big(\langle id, c\rangle_*(\omega)\Big) \\ &= \bigvee_{\omega \in \mathcal{D}(X)} D_{KL}\Big(\langle id, c\rangle_*(\omega),\ \omega \otimes c_*(\omega)\Big) \\ &= \bigvee_{\omega \in \mathcal{D}(X)} \sum_{x \in X,\, y \in Y} \omega(x) \cdot c(x)(y) \cdot \ln\left(\frac{c(x)(y)}{c_*(\omega)(y)}\right) \\ &= \bigvee_{\omega \in \mathcal{D}(X)} \omega \models D_{KL}\Big(c(-), c_*(\omega)\Big) \\ &= \bigvee_{\omega \in \mathcal{D}(X)} H\Big(c_*(\omega)\Big) - \Big(\omega \models H \circ c\Big) \end{aligned} \tag{6.35}$$

In order to see that the above join exists, we obtain from the middle formulation in (6.35):

$$\begin{aligned}
Cap(c) &= \bigvee_{\omega\in\mathcal{D}(X)} \sum_{x\in X,\, y\in Y} \omega(x)\cdot c(x)(y)\cdot \ln\left(\frac{c(x)(y)}{c_*(\omega)(y)}\right) \\
&= \bigvee_{\omega\in\mathcal{D}(X)} \sum_{x\in X,\, y\in Y} \omega(x)\cdot c(x)(y)\cdot \ln\left(\frac{\omega(x)\cdot c(x)(y)}{\omega(x)\cdot c_*(\omega)(y)}\right) \\
&= \bigvee_{\omega\in\mathcal{D}(X)} \sum_{x\in X,\, y\in Y} \omega(x)\cdot c(x)(y)\cdot \ln\left(\frac{1}{\omega(x)}\right) \\
&\qquad + \sum_{x\in X,\, y\in Y} \omega(x)\cdot c^*(\mathbf{1}_y)(x)\cdot \ln\left(\frac{\omega(x)\cdot c^*(\mathbf{1}_y)(x)}{\omega \models c^*(\mathbf{1}_y)}\right) \\
&= \bigvee_{\omega\in\mathcal{D}(X)} H(\omega) \;+\; \sum_{x\in X,\, y\in Y} \big(\omega \models c^*(\mathbf{1}_y)\big)\cdot \omega|_{c^*(\mathbf{1}_y)}(x)\cdot \ln\big(\omega|_{c^*(\mathbf{1}_y)}(x)\big) \\
&= \bigvee_{\omega\in\mathcal{D}(X)} H(\omega) \;-\; \sum_{y\in Y} \big(\omega \models c^*(\mathbf{1}_y)\big)\cdot H\big(\omega|_{c^*(\mathbf{1}_y)}\big) \\
&\leq \bigvee_{\omega\in\mathcal{D}(X)} H(\omega) \\
&\leq \ln\big(|X|\big).
\end{aligned}$$

The latter inequality describes the maximal entropy of a distribution on a finite set X, see Lemma 2.10.8 (6).

Example 6.12.7 1 The identity channel $id : X \rightsquigarrow X$ on a finite set X has capacity $\ln(|X|)$ since:

$$\begin{aligned}
Cap(id) &= \bigvee_{\omega\in\mathcal{D}(X)} H(id_*(\omega)) - \big(\omega \models H\circ id\big) \\
&= \bigvee_{\omega\in\mathcal{D}(X)} H(\omega) - \sum_{x\in X} \omega(x)\cdot H(1|x\rangle) \\
&= \bigvee_{\omega\in\mathcal{D}(X)} H(\omega) - \sum_{x\in X} \omega(x)\cdot 0 \qquad \text{by Lemma 2.10.8 (1)} \\
&= \bigvee_{\omega\in\mathcal{D}(X)} H(\omega) \;=\; \ln\big(|X|\big).
\end{aligned}$$

2 The capacity of a distribution $\omega \in \mathcal{D}(X)$, considered as channel $\omega\colon \mathbf{1} \rightsquigarrow X$ is 0. The only distribution on the singleton set $\mathbf{1} = \{0\}$ is $1|0\rangle$, so:

$$Cap\big(\mathbf{1} \xrightarrow{\omega} X\big) \;=\; H\big(\omega_*(1|0\rangle)\big) - \big(1|0\rangle \models H\circ\omega\big) \;=\; H(\omega) - H(\omega) \;=\; 0.$$

3 A noisy typewriter is defined in (Cover and Thomas, 2006, §7.1.3) as a channel $tw\colon A \rightsquigarrow A$, where $A = \{a, b, c, \ldots, z\}$ is the set of 26 letters of the Latin alphabet. This typewrite channel tw is defined as:

$$tw(a) \;:=\; \tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle \qquad tw(b) \;:=\; \tfrac{1}{2}|b\rangle + \tfrac{1}{2}|c\rangle \qquad \cdots \qquad tw(z) \;:=\; \tfrac{1}{2}|z\rangle + \tfrac{1}{2}|a\rangle.$$

It is not hard to see that the uniform distribution $unif_A$ on A satisfies $tw_*(unif_A) = unif_A$. It yields the capacity $\ln(13)$ of the typewrite channel tw as:

$$\begin{aligned}
Cap(tw) &= H(unif_A) - (unif_A \models H\circ tw) \\
&= \ln(26) - \sum_{x\in A} \tfrac{1}{26}\cdot H(tw(x)) \\
&= \ln(26) - \Big(-\tfrac{1}{2}\cdot\ln\big(\tfrac{1}{2}\big) - \tfrac{1}{2}\cdot\ln\big(\tfrac{1}{2}\big)\Big) \;=\; \ln(26) - \ln(2) \;=\; \ln(13).
\end{aligned}$$

Intuitively, this number 13 is the maximum number of letters than can be transferred reliably, namely by transferring $a, c, e, \ldots$, each time skipping one letter. See Cover and Thomas (2006) for more details.

The following result is essentially a reformulation of (Blahut, 1972, Thm. 1). This connection is elaborated in Exercise 8.2.6.

Proposition 6.12.8 *Consider $\omega \in \mathcal{D}(X)$ and $c\colon X \rightsquigarrow Y$, where X is a finite set. Define the factor $p\colon X \to \mathbb{R}_{\geq 0}$ as:*

$$p(x) \coloneqq D_{KL}\Big(c(x),\, c_*(\omega)\Big),$$

so that $MI(\langle id, c\rangle_*(\omega)) = \omega \models p$, see (6.35). The e-power update $\omega' \coloneqq \omega|_{e^p}$ yields an increase of mutual information:

$$MI(\langle id, c\rangle_*(\omega')) \;\geq\; MI(\langle id, c\rangle_*(\omega)).$$

Proof. We first present the main line, and subsequently elaborate the marked inequalities.

$$\begin{aligned}
MI(\langle id, c\rangle_*(\omega')) &= \sum_{x\in X, y\in Y} \omega'(x)\cdot c(x)(y)\cdot \ln\left(\frac{c(x)(y)}{c_*(\omega')(y)}\right) \\
&\overset{(*)}{\geq} \sum_{x\in X, y\in Y} \omega'(x)\cdot c(x)(y)\cdot \ln\left(\frac{\omega(x)\cdot c(x)(y)}{\omega'(x)\cdot c_*(\omega)(y)}\right) \\
&\overset{(**)}{=} \ln\Big(\omega \models e^p\Big) \\
&\geq \omega \models p \qquad \text{by Exercise 4.1.6} \\
&= MI(\langle id, c\rangle_*(\omega)).
\end{aligned}$$

The marked inequality $\overset{(*)}{\geq}$ is obtained via the inequality $\ln(a) \geq 1 - \frac{1}{a}$ from (2.77).

$$\begin{aligned}
&\sum_{x\in X, y\in Y} \omega'(x)\cdot c(x)(y)\cdot \left[\ln\left(\frac{c(x)(y)}{c_*(\omega')(y)}\right) - \ln\left(\frac{\omega(x)\cdot c(x)(y)}{\omega'(x)\cdot c_*(\omega)(y)}\right)\right] \\
&= \sum_{x\in X, y\in Y} \omega'(x)\cdot c(x)(y)\cdot \ln\left(\frac{\omega'(x)\cdot c_*(\omega)(y)}{\omega(x)\cdot c_*(\omega')(y)}\right) \\
&\geq \sum_{x\in X, y\in Y} \omega'(x)\cdot c(x)(y)\cdot \left(1 - \frac{\omega(x)\cdot c_*(\omega')(y)}{\omega'(x)\cdot c_*(\omega)(y)}\right) \\
&= 1 - \sum_{x\in X, y\in Y} c(x)(y)\cdot \frac{\omega(x)\cdot c_*(\omega')(y)}{c_*(\omega)(y)} \;=\; 1 - \sum_{y\in Y} c_*(\omega')(y) \;=\; 1 - 1 \;=\; 0.
\end{aligned}$$

The inequality $\overset{(**)}{=}$ is obtained as follows.

$$\begin{aligned}
&\sum_{x\in X, y\in Y} \omega'(x)\cdot c(x)(y)\cdot \ln\left(\frac{\omega(x)\cdot c(x)(y)}{\omega'(x)\cdot c_*(\omega)(y)}\right)\\
&= \sum_{x\in X, y\in Y} \omega|_{e^p}(x)\cdot c(x)(y)\cdot \ln\left(\frac{c(x)(y)}{c_*(\omega)(y)}\cdot\frac{\omega\models e^p}{e^p(x)}\right)\\
&= \sum_{x\in X, y\in Y} \omega|_{e^p}(x)\cdot c(x)(y)\cdot \ln\left(\frac{c(x)(y)}{c_*(\omega)(y)}\right)\\
&\qquad + \sum_{x\in X, y\in Y} \omega|_{e^p}(x)\cdot c(x)(y)\cdot \ln\left(\frac{\omega\models e^p}{e^{p(x)}}\right)\\
&= \sum_{x\in X} \omega|_{e^p}(x)\cdot D_{KL}(c(x),\, c_*(\omega)) \;+\; \sum_{x\in X} \omega|_{e^p}(x)\cdot\Big(\ln(\omega\models e^p) - \ln(e^{p(x)})\Big)\\
&= \Big(\omega|_{e^p}\models p\Big) \;+\; \ln\Big(\omega\models e^p\Big) \;-\; \sum_{x\in X} \omega|_{e^p}(x)\cdot p(x)\\
&= \Big(\omega|_{e^p}\models p\Big) + \ln\Big(\omega\models e^p\Big) - \Big(\omega|_{e^p}\models p\Big) \;=\; \ln\Big(\omega\models e^p\Big).
\end{aligned}$$

❑

The following result is (Blahut, 1972, Thm. 3) and (Arimoto, 1972, Thm. 1). It uses the update construction from the previous lemma iteratively. We refer to Blahut (1972); Arimoto (1972) for proofs that the maximum is indeed reached in this manner. We do provide an illustration subsequently.

Theorem 6.12.9 *Let $c\colon X \rightsquigarrow Y$ be a channel, with a finite set X as domain. The capacity $Cap(c)\in\mathbb{R}_{\geq 0}$ of the channel is obtained as:*

$$Cap(c) \;=\; \lim_{n\to\infty} MI\Big(\langle id, c\rangle_*(\omega_n)\Big),$$

where:

$$\begin{cases} \omega_0 \;=\; unif_X & \textit{(or an arbitrary distribution on } X\textit{)}\\ \omega_{n+1} \;=\; \omega_n\big|_{e^p} & \textit{with } p(x) = D_{KL}(c(x), c_*(\omega_n)). \end{cases}$$

In Theorem 6.12.4 (1) we have seen that updating with p works better than with e^p, when p is a predicate. This may be applied in the situation of the previous theorem. Experimentally, this gives faster approximation of the capacity, but a proof that updating with p works better is missing. We do include an illustration.

Example 6.12.10 Consider the channel $c\colon \{0,1,2\} \rightsquigarrow \{a,b,c,d\}$ defined as:

$$\begin{aligned}
c(0) &= \tfrac{3}{10}|b\rangle + \tfrac{3}{10}|c\rangle + \tfrac{2}{5}|d\rangle\\
c(1) &= \tfrac{1}{10}|a\rangle + \tfrac{3}{10}|b\rangle + \tfrac{1}{5}|c\rangle + \tfrac{2}{5}|d\rangle\\
c(2) &= \tfrac{1}{10}|a\rangle + \tfrac{1}{2}|b\rangle + \tfrac{1}{10}|c\rangle + \tfrac{3}{10}|d\rangle.
\end{aligned}$$

Using the approach of Theorem 6.12.9 we obtain a channel capacity of 0.07755212, via the distribution $0.544|0\rangle + 0.456|2\rangle$. This capacity is reached in 409 iterations — starting from the uniform distribution on $\{0,1,2\}$.

When we use updating with p, instead of with the e-power e^p, we reach this capacity in only 26 iterations. Thus, we see a substantial difference — in this example.

Exercises

6.12.1 Let $\omega \in \mathcal{D}(X)$ and $p \in Pred(X)$. Check that updating ω with $e^{p^\perp}$ is the same as updating with e^{-p}.

6.12.2 Let $\omega \in \mathcal{D}(X)$ and $p \in Obs(X)$. Show that updating with e^{-p} decreases the validity of p^m, as in:

$$\omega|_{e^{-p}} \models p^m \;\leq\; \omega \models p^m.$$

Conclude that: $\omega|_{e^{-p}} \models e^p \;\leq\; \omega \models e^p$.

6.12.3 Let $\omega \in \mathcal{D}(X)$.

1 Prove the following generalisation of Exercise 4.1.6, for an observable $p\colon X \to \mathbb{R}$.

$$\omega \models p^m \;\&\; e^p \;\geq\; (\omega \models p^m) \cdot e^{(\omega \models p)}.$$

2 Show also that for any $m \in \mathbb{N}$, when p is a predicate,

$$\omega \models p^\perp \;\&\; p^m \;\&\; e^p \;\leq\; (\omega \models p^\perp) \cdot (\omega \models p^m \;\&\; e^p).$$

6.12.4 Let $\omega \in \mathcal{D}(X)$ be a distribution with a positive factor $p\colon X \to \mathbb{R}_{>0}$.

1 Use the 'log-sum inequality' from Exercise 2.10.8 to prove the first inequality in:

$$\omega|_p \models \ln(p) \;\geq\; \ln(\omega \models p) \;\geq\; \omega \models \ln(p).$$

The second inequality holds by Exercise 4.1.6.

2 Use Jensen's inequality to prove:

$$\omega|_p \models p - \ln(p) \;\leq\; \ln(\omega \models e^p) \;-\; \ln(\omega \models p).$$

6.12.5 Show for $\omega \in \mathcal{D}(X)$ and $p \in Obs(X)$,

$$\omega \models p \;\&\; e^p \;\geq\; (\omega \models e^p) \cdot \ln(\omega \models e^p).$$

Equivalently, for a positive factor q,

$$\omega \models q \;\&\; \ln(q) \;\geq\; (\omega \models q) \cdot \ln(\omega \models q).$$

That is, via Lemma 6.1.7 (3),

$$\omega|_q \models \ln(q) \;\geq\; \ln(\omega \models q).$$

Hint: The log-sum inequality works again.

6.12.6 Prove, or a distribution $\omega \in \mathcal{D}(X)$ and a factor $p\colon X \to \mathbb{R}_{\geq 0}$,

1 $\omega|_{e^{-p}} \models e^{-p} + e^p \;\leq\; \omega \models e^{-p} + e^p$;
Hint: Recall Proposition 6.11.3 (4).

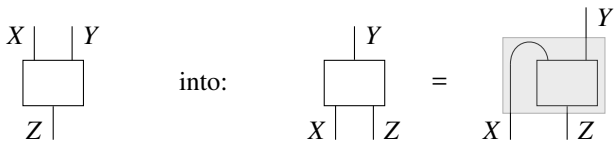
2 $\omega \models e^{-p} + e^p \;\leq\; \omega|_{e^p} \models e^{-p} + e^p$.

7

Daggers and disintegrations

In the previous chapter we have seen how distributions are updated, via incorporation of evidence, and how such updated distributions are used in probabilistic reasoning. In this chapter we use updating in a structural manner and show how it can be used to turn, under suitable conditions, a channel $X \multimap Y$ into a channel $Y \multimap X$ in the other direction. This reversal of channels is also called Bayesian inversion. In fact, because the Bayesian approach so crucially depends on this reversal, it is (or was) also called 'inverse probability'. It corresponds to turning a conditional probability $P(Y \mid X)$ into $P(X \mid Y)$. Earlier we have seen forward and backward inference along a channel. With this reversal of channels we can tie the notions together and show that forward inference along a channel c corresponds to backward reasoning along the reversed channel $c^\dagger$ — and vice-versa. We shall make extensive use of string diagrams for describing the properties of channel inversion (like in Cho and Jacobs (2019)).

Once we understand Bayesian inversion, we describe it as an instance of a more general technique, called disintegration. In essence it involves turning a channel:



Thus, disintegration changes an outgoing wire (here labeled X) into an incoming wire, by bending it around. Doing so in a probabilistic setting involves re-normalisation, as indicated by the shaded box. In the first three sections of this chapter we shall introduce this change of wires, ultimately in a form more general than described above.

This chapter illustrates, in Sections 7.4 and 7.5, how channel inversion and disintegration are used to solve practical problems in machine learning, like learning with missing data, naive Bayesian classification, and decision tree learning. Also, daggers express the essentials of Laplace's rule of succession, as shown in Section 7.6. This rule expresses what can be expected next, after a number observations. Laplace's original question was: what is the probability that the sun rises tomorrow, having seen n successive days with the sun rising. The answer to such (and similar) questions is expressed via the mean of a dagger.

In Section 7.7 the notion of sufficient statistic is introduced as a special form of disintegration without a particular dependence (wire). Such sufficient statistics express basic relationships that actually occurred earlier in several examples — for instance between accumulation and arrangement, in Theorem 3.3.1. Only now the general account is provided, in terms of disintegration. So-called Ewens and Bell distributions on multiset partitions and on set partitions, are elaborated as examples.

Section 7.8 describes what it means that a (joint) distribution 'matches' a string diagram, in the sense that all boxes in the diagram can be interpreted to form the distribution. Given a string diagram, one can try to 'factorise' the joint distribution, via marginalisation and disintegration, to find such interpretations. This is useful, since the shape of the string diagram gives information about how different parts in the distribution depend on each other (or not).

Next, dagger channels are used to describe a new update rule, named after Jeffrey. It is different from backward inference, which we call Pearl's rule, and produces quite different outcomes. What we will show is that the two update rules correspond to two intuitively different forms of learning, namely learning from what's right (via encouragement) and learning from what's wrong (via discouragement). This will be made mathematically precise: Pearl's rule increases validity, whereas Jeffrey's rule decreases (Kullback-Leibler) divergence. Thus, Jeffrey's rule can be understood as a correction mechanism, like in predictive coding theory, where the human mind is seen as a Bayesian prediction and correction engine. One can also see some resemblance with Karl Popper's emphasis on falsification, in the sense of learning from negative outcomes. In Section 7.9 we introduce the basics of Jeffrey's rule, with several illustrations. Further explanations and examples will be given in Chapter 8 on statistical learning.

As sketched above, disintegration can be described graphically via bending of wires and normalisation (using shaded boxes). Section 7.10 shows that such shaded boxes satisfy certain rules, including a rule for removal of such shaded boxes (from Jacobs et al. (2025)). This gives a calculus for graphically solving conditioning questions, as will be illustrated for Bayesian networks.

The final section of this chapter captures channel reversal in categorical terms via a 'dagger' functor $\mathbf{Krn} \to \mathbf{Krn}^{\mathrm{op}}$, on a category $\mathbf{Krn}$ of kernels (as in Clerc et al. (2017)). This abstract categorical perspectives demonstrates the analogy with reveral of relations.

7.1 Bayesian inversion: the dagger of a channel

In this section we make a construction explicit that we have already been using several times, namely the reversal of a channel. This reversal is familiar in the literature, under the name Bayesian inversion. It turns a conditional probability $P(y \mid x)$ into $P(x \mid y)$. We shall describe it as a construction that turns a channel $X \rightsquigarrow Y$ into a channel $Y \rightsquigarrow X$, in presence of a 'prior' distribution on X. A superscript-dagger notation $(-)^\dagger$ is often used for these Bayesian inversions since such inversions are similar to the adjoint-transpose of a linear operator A between Hilbert spaces, see Clerc et al. (2017). Such a transpose is typically written as $A^\dagger$, or also as A^*. The dagger notation is more common in quantum theory, and has been formalised in terms of dagger categories Coecke and Kissinger (2016). This categorical approach is sketched in Section 7.11 below.

Suppose we have a channel $c \colon X \rightsquigarrow Y$ and a distribution $\omega \in \mathcal{D}(X)$. How can we obtain a channel $Y \rightsquigarrow X$. Well, starting from an element $y \in Y$, we can form the point predicate $\mathbf{1}_y$ on Y, and turn it into a predicate $c^*(\mathbf{1}_y)$ on X via predicate transformation. The update $\omega|_{c^*(\mathbf{1}_y)}$ now gives a new distribution on X, via backward inference. In this way we have constructed a function $Y \to \mathcal{D}(X)$. There is one caveat: the update $\omega|_{c^*(\mathbf{1}_y)}$ must exist, that is we need to require the non-zero validity of:

$$\omega \models c^*(\mathbf{1}_y) \overset{(4.15)}{=} c_*(\omega) \models \mathbf{1}_y = c_*(\omega)(y).$$

Thus, in order to form the dagger, we need to require that the predicted, pushedforward distribu-

tion $c_*(\omega)$ has full support, so that $c_*(\omega)(y) \neq 0$ for each $y \in Y$. Such a full support requirement only makes sense if the set Y is finite. The definition below summarises the situation and introduces notation for the reversed channel.

Definition 7.1.1 Let $c\colon X \rightsquigarrow Y$ be a channel, where the codomain Y is a finite set, with a distribution $\omega \in \mathcal{D}(X)$ on its domain, such that the transformed / predicted state $c_*(\omega) \in \mathcal{D}(Y)$ has full support.

The reversed 'dagger' channel $c^\dagger_\omega\colon Y \rightsquigarrow X$ is defined as on $y \in Y$ as:

$$c^\dagger_\omega(y) \;=\; \omega|_{c^*(\mathbf{1}_y)} \;=\; \sum_{x \in X} \frac{\omega(x) \cdot c(x)(y)}{c_*(\omega)(y)} \,\big|x\big\rangle \;=\; \sum_{x \in X} \frac{\omega(x) \cdot c(x)(y)}{\sum_z \omega(z) \cdot c(z)(y)} \,\big|x\big\rangle. \tag{7.1}$$

In general, the pushforward $c_*(c^\dagger_\omega(y))$ of the inversion is called the posterior predictive distribution.

We have already implicitly used the dagger of a channel in many situations.

Example 7.1.2 In Example 6.2.3 we used a channel $c\colon \{H, T\} \rightsquigarrow \{W, B\}$ from the two sides of a coin to two colors of balls in an urn, together with a fair coin $\mathit{unif} \in \mathcal{D}(\{H, T\})$. We had evidence of a white ball and wanted to know the updated coin distribution. The outcome that we computed can be described via a dagger channel, since:

$$c^\dagger_{\mathit{unif}}(W) \;=\; \mathit{unif}|_{c^*(\mathbf{1}_W)} \;=\; \tfrac{22}{67}|H\rangle + \tfrac{45}{67}|T\rangle.$$

The same redescription in terms of Bayesian inversion can be used for Examples 6.2.4 and 6.2.11, 6.2.12. In Example 6.2.6 we can use Bayesian inversion for the point evidence case, but not for the illustration with soft evidence, *i.e.* with fuzzy predicates.

We consider a new illustration from the literature whose essence is expressed via a dagger of a channel.

Example 7.1.3 We quote from (Grinstead and Snell, 1998, Ex. 4.1.16).

A doctor is trying to decide if a patient has one of three diseases d_1, d_2, or d_3. Two tests are to be carried out, each of which results in a positive (+) or a negative (-) outcome. There are four possible test patterns ++, +−, −+, and −−. National records have indicated that, for 10,000 people having one of these three diseases, the distribution of diseases and test results are as in ...

... the following table, also copied from Grinstead and Snell (1998).

disease	**number having this disease**	++	+−	−+	−−
d_1	3215	2110	301	704	100
d_2	2125	396	132	1187	410
d_3	4660	510	3568	73	509
Total	10,000				

The aim formulated in Grinstead and Snell (1998) is to compute "posterior probabilities" for the diseases. We shall do this via a dagger of a channel with a prior distribution that are implicit in this table.

We formalise this situation with a set of diseases $D = \{d_1, d_2, d_3\}$ and a set of test patterns

$T = \{++, +-, -+, --\}$. The prior distribution $\omega \in \mathcal{D}(D)$ is obtained from the second column via normalisation:

$$\begin{aligned}\omega &= \mathit{Flrn}\big(3215|d_1\rangle + 2125|d_2\rangle + 4660|d_3\rangle\big)\\ &= \tfrac{3215}{10000}|d_1\rangle + \tfrac{2125}{10000}|d_2\rangle + \tfrac{4660}{10000}|d_3\rangle \;=\; \tfrac{643}{2000}|d_1\rangle + \tfrac{17}{80}|d_2\rangle + \tfrac{233}{500}|d_3\rangle.\end{aligned}$$

The four columns on the right in the above table determine a channel $c\colon D \rightarrowtriangle T$, namely:

$$\begin{aligned}c(d_1) &= \mathit{Flrn}\big(2110\big|{++}\rangle + 301\big|{+-}\rangle + 704\big|{-+}\rangle + 100\big|{--}\rangle\big)\\ &= \tfrac{422}{643}\big|{++}\rangle + \tfrac{301}{3215}\big|{+-}\rangle + \tfrac{704}{3215}\big|{-+}\rangle + \tfrac{20}{643}\big|{--}\rangle\\ c(d_2) &= \mathit{Flrn}\big(396\big|{++}\rangle + 132\big|{+-}\rangle + 1187\big|{-+}\rangle + 410\big|{--}\rangle\big)\\ &= \tfrac{396}{2125}\big|{++}\rangle + \tfrac{132}{2125}\big|{+-}\rangle + \tfrac{1187}{2125}\big|{-+}\rangle + \tfrac{82}{425}\big|{--}\rangle\\ c(d_3) &= \mathit{Flrn}\big(510\big|{++}\rangle + 3568\big|{+-}\rangle + 73\big|{-+}\rangle + 509\big|{--}\rangle\big)\\ &= \tfrac{51}{466}\big|{++}\rangle + \tfrac{892}{1165}\big|{+-}\rangle + \tfrac{73}{4660}\big|{-+}\rangle + \tfrac{509}{4660}\big|{--}\rangle\end{aligned}$$

In order to compute posterior disease probabilities for test patterns we need to reverse the channel $c\colon D \rightarrowtriangle T$. Let's write $f := c^{\dagger}_{\omega}\colon T \rightarrowtriangle D$ for the reversed dagger channel, with ω as prior. It produces four new disease distributions, associated with the four test patterns.

$$\begin{aligned}f(++) &= \omega|_{c^*(\mathbf{1}_{++})} = \tfrac{1055}{1508}|d_1\rangle + \tfrac{99}{754}|d_2\rangle + \tfrac{255}{1508}|d_3\rangle\\ &\approx 0.7|d_1\rangle + 0.131|d_2\rangle + 0.169|d_3\rangle\\ f(+-) &= \omega|_{c^*(\mathbf{1}_{+-})} = \tfrac{301}{4001}|d_1\rangle + \tfrac{132}{4001}|d_2\rangle + \tfrac{3568}{4001}|d_3\rangle\\ &\approx 0.0752|d_1\rangle + 0.033|d_2\rangle + 0.892|d_3\rangle\\ f(-+) &= \omega|_{c^*(\mathbf{1}_{-+})} = \tfrac{176}{491}|d_1\rangle + \tfrac{1187}{1964}|d_2\rangle + \tfrac{73}{1964}|d_3\rangle\\ &\approx 0.358|d_1\rangle + 0.604|d_2\rangle + 0.0372|d_3\rangle\\ f(--) &= \omega|_{c^*(\mathbf{1}_{--})} = \tfrac{100}{1019}|d_1\rangle + \tfrac{410}{1019}|d_2\rangle + \tfrac{509}{1019}|d_3\rangle\\ &\approx 0.0981|d_1\rangle + 0.402|d_2\rangle + 0.5|d_3\rangle.\end{aligned}$$

These updated disease distributions contain the decimal posterior probabilities that are listed in Table 4.1.2 in Grinstead and Snell (1998) — up to some small rounding differences.

We make some special cases of the dagger of a channel explicit.

Lemma 7.1.4 *Let X, Y be sets, where Y is finite.*

1 *When also X is finite, $c\colon X \rightarrowtriangle Y$ is a channel, and the uniform distribution $\mathrm{unif}_X = \sum_{x\in X} \frac{1}{|X|}|x\rangle$ is used as prior, then:*

$$c^{\dagger}_{\mathrm{unif}_X}(y) = \sum_{x\in X} \frac{c(x)(y)}{\sum_{z\in X} c(z)(y)}\,|x\rangle.$$

2 *When $f\colon X \to Y$ is a surjective function, used as deterministic channel $X \rightarrowtriangle Y$, then for an $\omega \in \mathcal{D}(X)$ such that $\mathcal{D}(f)(\omega)$ has full support,*

$$f^{\dagger}_{\omega}(y) = \sum_{x\in f^{-1}(y)} \frac{\omega(x)}{\mathcal{D}(f)(\omega)(y)}\,|x\rangle.$$

In this situation one has $f \circ f^{\dagger}_{\omega} = \mathrm{id}$, or, more explicitly, for $y \in Y$,

$$(f \circ f^{\dagger}_{\omega})(y) = f_*(f^{\dagger}_{\omega}(y)) = \mathcal{D}(f)\big(f^{\dagger}_{\omega}(y)\big) = 1\big|y\rangle. \tag{7.2}$$

3 *By combinining the previous two points we find that the dagger of a surjective function, with respect to the uniform prior, is the probabilistic inverse introduced in Definition 2.4.7:*

$$f^{\dagger}_{\mathit{unif}_X} = f^{\sim 1}.$$

The surjectivity of the function in the last to items is needed for the full support requirement that is required for daggers.

Proof. 1 This follows directly from the definitions.

2 Recall that $\langle f \rangle = \mathit{unit} \circ f \colon X \to \mathcal{D}(Y)$ is the deterministic channel corresponding to f, where the $\langle \cdot \rangle$ brackets are often omitted. Thus:

$$\begin{aligned} f^{\dagger}_{\omega}(y) &\overset{(7.1)}{=} \sum_{x \in X} \frac{\omega(x) \cdot \langle f \rangle(x)(y)}{\sum_{z \in X} \omega(z) \cdot \langle f \rangle(z)(y)} \big|x\rangle = \sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\sum_{z \in f^{-1}(y)} \omega(z)} \big|x\rangle \\ &= \sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\mathcal{D}(f)(\omega)(y)} \big|x\rangle \qquad \text{see Lemma 2.1.3.} \end{aligned}$$

Then:

$$\mathcal{D}(f)\big(f^{\dagger}_{\omega}(y)\big) = \sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\mathcal{D}(f)(\omega)(y)} \big|f(x)\rangle = \sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\mathcal{D}(f)(\omega)(y)} \big|y\rangle = 1\big|y\rangle.$$

3 We now have, for a surjective function $f \colon X \twoheadrightarrow Y$,

$$\begin{aligned} f^{\dagger}_{\mathit{unif}_X}(y) &= \sum_{x \in f^{-1}(y)} \frac{\mathit{unif}_X(x)}{\sum_{z \in f^{-1}(y)} \mathit{unif}_X(z)} \big|x\rangle = \sum_{x \in f^{-1}(y)} \frac{1/|X|}{\sum_{z \in f^{-1}(y)} 1/|X|} \big|x\rangle \\ &= \sum_{x \in f^{-1}(y)} \frac{1}{|f^{-1}(y)|} \big|x\rangle \overset{(2.40)}{=} f^{\sim 1}(y). \end{aligned}$$

❑

Theorem 6.3.2 describes updating when a joint distribution can be written in two different ways as graph. In the special case (6.10) when the update involves a point predicate, the above dagger formula (7.1) emerges. We shall soon see (in Definition 7.1.1) that daggers indeed arise in such a graph situation. In the meantime we re-describe some more earlier examples, now as daggers / reversals of channels.

Example 7.1.5 In Example 6.3.3 we have seen several instances of backward inference $\omega|_{c^*(\mathbf{1}_y)}$ that we now recognise as daggers of the channel c.

Originally, in Example 2.4.6 we have introduced the arrangement channel $\mathit{arr} \colon \mathcal{N}[K](X) \rightsquigarrow X^K$ as the probabilistic inverse of the accumulation function $\mathit{acc} \colon X^K \to \mathcal{N}[K](X)$, see Definition 2.4.7. Using Lemma 7.1.4 (3) we see that we can now write:

$$\mathit{arr} = \mathit{acc}^{\sim 1} = \mathit{acc}^{\dagger}_{\mathit{unif}_{X^K}},$$

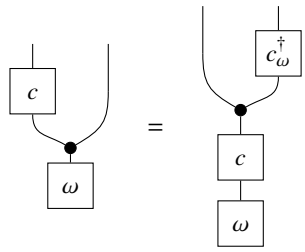
where the latter description as dagger works if the set X is finite.

Accumulation can be described as a dagger too, namely of arrangement. In Example 6.3.3 we have done so for the multinomial distribution as prior. But it also works for the uniform distribution, when the set X is finite, since by Lemma 7.1.4 (1),

$$\mathit{arr}^{\dagger}_{\mathit{unif}_{\mathcal{N}[K](X)}}(\vec{x}) = \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\mathit{arr}(\varphi)(\vec{x})}{\sum_{\psi \in \mathcal{N}[K](X)} \mathit{arr}(\psi)(\vec{x})} \big|\varphi\rangle = 1\big|\mathit{acc}(\vec{x})\rangle = \langle \mathit{acc} \rangle(\vec{x}).$$

The following result gives the string-diagrammatic characterisation of dagger channels that is used in Cho and Jacobs (2019).

Theorem 7.1.6 *Let $c\colon X \rightsquigarrow Y$ be a channel, with finite codomain Y, together with a distribution $\omega \in \mathcal{D}(X)$ such that $c_*(\omega)$ has full support. The resulting dagger channel $c^\dagger_\omega$ is the unique channel $Y \rightsquigarrow X$ satisfying:*



(7.3)

In particular, by marginalising out the first wires we get:

$$\omega \;=\; (c^\dagger_\omega)_*(c_*(\omega)). \tag{7.4}$$

This is in essence the law of total probability, see Exercise 7.1.10 below.

Proof. We first check that (7.3) holds.

$$\begin{aligned}
\langle id, c^\dagger_\omega\rangle_*(c_*(\omega)) &= \sum_{y\in Y}\sum_{x\in X} c_*(\omega)(y)\cdot c^\dagger_\omega(y)(x)\,\big|y,x\rangle \\
&\overset{(7.1)}{=} \sum_{y\in Y}\sum_{x\in X} c_*(\omega)(y)\cdot \frac{\omega(x)\cdot c(x)(y)}{c_*(\omega)(y)}\,\big|y,x\rangle \\
&= \sum_{y\in Y}\sum_{x\in X} \omega(x)\cdot c(x)(y)\,\big|y,x\rangle \;=\; \langle c, id\rangle_*(\omega).
\end{aligned}$$

For uniqueness, let Equation (7.3) hold with dagger $d\colon Y \rightsquigarrow X$ instead of $c^\dagger_\omega$. Then we can apply Equation (6.10) from Theorem 6.3.2, giving for each $y \in Y$,

$$d(y) \;=\; \omega|_{c^*(\mathbf{1}_y)} \overset{(7.1)}{=} c^\dagger_\omega(y). \qquad \square$$

A consequence (from Jacobs (2019)) of this result is that backward inference along a channel c can be expressed as forward inference along the reversal of c, and vice-versa. This demonstrates that the directions of inference and of channels are closely connected.

Corollary 7.1.7 *In the situation of Theorem 7.1.6, write $\tau := c_*(\omega)$.*

1 *Given a factor q on Y, we can express backward inference as forward inference with the dagger:*

$$\omega|_{c^*(q)} \;=\; (c^\dagger_\omega)_*(\tau|_q).$$

2 *Given a factor p on X, we can express forward inference as backward inference with the dagger:*

$$c_*(\omega)|_p \;=\; \tau|_{(c^\dagger_\omega)^*(p)}.$$

Proof. We combine Theorems 7.1.6 and 6.3.2. Item (1) follows directly from Equation (6.9). Also item (2) follows from Equation (6.9), after left-right mirroring of the two string diagrams in (7.3). $\square$

We continue this section with a few technical facts about daggers, for future use. First we describe several compositionality properties of the dagger, demonstrating that it interacts nicely with the basic constructions on channels.

Lemma 7.1.8 *For the following points we assume that daggers are well-defined, i.e. that the relevant pushforwards have full support, see Definition 7.1.1.*

1 For channels $c\colon X \rightsquigarrow U$ and $d\colon Y \rightsquigarrow V$, with distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$,

$$(c \otimes d)^{\dagger}_{\omega\otimes\rho} \;=\; c^{\dagger}_{\omega} \otimes d^{\dagger}_{\rho}.$$

2 For composable channels $X \overset{c}{\rightsquigarrow} Y \overset{d}{\rightsquigarrow} Z$ and a distribution $\omega \in \mathcal{D}(X)$, there is the following 'chain rule'.

$$(d \circ c)^{\dagger}_{\omega} \;=\; c^{\dagger}_{\omega} \circ d^{\dagger}_{c_*(\omega)}.$$

3 Identity channels are stable under reversal:

$$id^{\dagger}_{\omega} \;=\; id.$$

4 Double reversal returns the original channel:

$$\left(c^{\dagger}_{\omega}\right)^{\dagger}_{c_*(\omega)} \;=\; c.$$

Proof. 1 For elements $u \in U$ and $v \in V$ we have:

$$\begin{aligned}
(c \otimes d)^{\dagger}_{\omega\otimes\rho}(u,v) \;&=\; \sum_{(x,y)\in X\times Y} \frac{(\omega\otimes\rho)(x,y)\cdot(c\otimes d)(x,y)(u,v)}{(c\otimes d)_*(\omega\otimes\rho)(u,v)}\,\big|x,y\big\rangle \\
&=\; \sum_{x\in X,\,y\in Y} \frac{\omega(x)\cdot\rho(y)\cdot c(x)(u)\cdot d(y)(v)}{(c_*(\omega)\otimes d_*(\rho))(u,v)}\,\big|x,y\big\rangle \\
&=\; \sum_{x\in X,\,y\in Y} \frac{\omega(x)\cdot c(x)(u)}{c_*(\omega)(u)}\cdot\frac{\rho(y)\cdot d(y)(v)}{d_*(\rho)(v)}\,\big|x,y\big\rangle \\
&=\; \sum_{x\in X,\,y\in Y} c^{\dagger}_{\omega}(u)(x)\cdot d^{\dagger}_{\rho}(u)(y)\,\big|x,y\big\rangle \;=\; c^{\dagger}_{\omega}(u)\otimes d^{\dagger}_{\rho}(u) \;=\; \left(c^{\dagger}_{\omega}\otimes d^{\dagger}_{\rho}\right)(u,v).
\end{aligned}$$

2 For $z \in Z$ we have:

$$\begin{aligned}
\left(c^{\dagger}_{\omega}\circ d^{\dagger}_{c_*(\omega)}\right)(z) \;=\; (c^{\dagger}_{\omega})_*(d^{\dagger}_{c_*(\omega)}(z)) \;&=\; \sum_{x\in X}\sum_{y\in Y} d^{\dagger}_{c_*(\omega)}(z)(y)\cdot c^{\dagger}_{\omega}(y)(x)\,\big|x\big\rangle \\
&=\; \sum_{x\in X}\sum_{y\in Y} \frac{c_*(\omega)(y)\cdot d(y)(z)}{d_*(c_*(\omega))(z)}\cdot\frac{\omega(x)\cdot c(x)(y)}{c_*(\omega)(y)}\,\big|x\big\rangle \\
&=\; \sum_{x\in X} \frac{\omega(x)\cdot\sum_{y\in Y} c(x)(y)\cdot d(y)(z)}{(d\circ c)_*(\omega)(z)}\,\big|x\big\rangle \\
&=\; \sum_{x\in X} \frac{\omega(x)\cdot(d\circ c)(x)(z)}{(d\circ c)_*(\omega)(z)}\,\big|x\big\rangle \;=\; (d\circ c)^{\dagger}_{\omega}
\end{aligned}$$

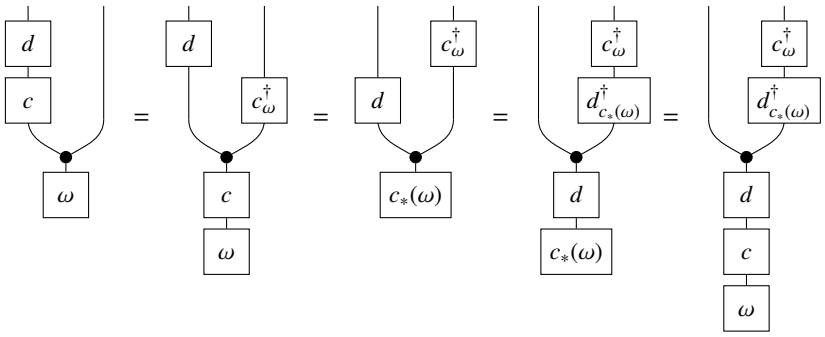
3 For an element $x \in X$,

$$id^{\dagger}_{\omega}(x) \;=\; \sum_{z\in X} \frac{\omega(z)\cdot id(z)(x)}{id_*(\omega)(x)}\,\big|z\big\rangle \;=\; \frac{\omega(x)}{\omega(x)}\,\big|x\big\rangle \;=\; 1\big|x\big\rangle \;=\; id(x).$$

4 Finally, for $x \in X$,

$$\begin{aligned}\left(c_\omega^\dagger\right)^\dagger_{c_*(\omega)}(x) &= \sum_{y \in Y} \frac{c_*(\omega)(y) \cdot c_\omega^\dagger(y)(x)}{(c_\omega^\dagger)_*(c_*(\omega))(x)} \big|y\rangle \overset{(7.4)}{=} \sum_{y \in Y} \frac{c_*(\omega)(y)}{\omega(x)} \cdot \frac{\omega(x) \cdot c(x)(y)}{c_*(\omega)(y)} \big|y\rangle \\ &= \sum_{y \in Y} c(x)(y) \big|y\rangle = c(x). \qquad \square\end{aligned}$$

One can also prove these results via equational reasoning with string diagrams, using uniqueness of daggers. For instance preservation of composition $\circ$ by the dagger, in item (2), follows by uniqueness from:



Daggers play a role in the basic concepts in computational neuroscience. We start with an illustration and then make these concepts explicit. This will be continued in Remark 7.9.6.

Example 7.1.9 The situation that we will describe now comes from the setting of active inference (Parr et al., 2022, Fig. 2.1). It is about distinguishing an apple (a) from a frog (f), where there is a prior distribution $\sigma = \frac{9}{10}|a\rangle + \frac{1}{10}|f\rangle$ about the state of affairs in the outside world. The criterion is wheter the object can jump (j) or not ($j^\perp$), which is given via a channel $c\colon \{a, f\} \rightsquigarrow \{j, j^\perp\}$ with:

$$c(a) = \tfrac{1}{100}|j\rangle + \tfrac{99}{100}|j^\perp\rangle \qquad\qquad c(f) = \tfrac{81}{100}|j\rangle + \tfrac{19}{100}|j^\perp\rangle.$$

This give a predicted jump distribution $c_*(\sigma) = \frac{9}{100}|j\rangle + \frac{91}{100}|j^\perp\rangle$. The dagger $c_\sigma^\dagger\colon \{j, j^\perp\} \rightsquigarrow \{a, f\}$ can be computed as:

$$\begin{aligned}c_\sigma^\dagger(j) &= \frac{\sigma(a) \cdot c(a)(j)}{c_*(\sigma)(j)}|a\rangle + \frac{\sigma(f) \cdot c(f)(j)}{c_*(\sigma)(j)}|f\rangle \\ &= \frac{9/10 \cdot 1/100}{9/100}|a\rangle + \frac{1/10 \cdot 81/100}{9/100}|f\rangle = \tfrac{1}{10}|a\rangle + \tfrac{9}{10}|f\rangle \\ c_\sigma^\dagger(j^\perp) &= \frac{\sigma(a) \cdot c(a)(j^\perp)}{c_*(\sigma)(j^\perp)}|a\rangle + \frac{\sigma(f) \cdot c(f)(j^\perp)}{c_*(\sigma)(j^\perp)}|f\rangle \\ &= \frac{9/10 \cdot 99/100}{91/100}|a\rangle + \frac{1/10 \cdot 19/100}{91/100}|f\rangle = \tfrac{891}{910}|a\rangle + \tfrac{19}{910}|f\rangle.\end{aligned}$$

Two kinds of 'surprise' are defined in Parr et al. (2022), for an element of the codomain of the channel. In this case, the element is jump j. What is called the 'ordinary' suprise in Parr et al. (2022), but what is more appropriately called the prediction surprise, is the negative logarithm of the jump prediction:

$$-\ln\big(c_*(\sigma)(j)\big) \approx 2.41.$$

There is also what is called the 'Bayesian' suprise in Parr et al. (2022), but what we prefer to call the update suprise. It is expressed via the dagger, as:

$$D_{KL}\Big(c^{\dagger}_{\sigma}(j),\, \sigma\Big) \approx 1.76.$$

This is the divergence between the posterior $c^{\dagger}_{\sigma}(j) = \sigma|_{c^*(\mathbf{1}_j)}$ and the prior σ. These two surprise values are high. This makes sense if we have in mind that the thing can jump and the prior involves a high apple probability. In contrast, the two surprises for no-jump are much lower:

$$-\ln\Big(c_*(\sigma)(j^{\perp})\Big) \approx 0.094 \qquad\qquad D_{KL}\Big(c^{\dagger}_{\sigma}(j^{\perp}),\, \sigma\Big) \approx 0.05.$$

We notice that one can write the prediction and update surprises for jump — and similarly for no-jump — as:

$$-\ln\Big(c_*(\sigma)(j)\Big) = D_{KL}\Big(1|\,j\rangle,\, c_*(\sigma)\Big) \qquad\qquad D_{KL}\Big(c^{\dagger}_{\sigma}(j),\, \sigma\Big) = D_{KL}\Big((c^{\dagger}_{\sigma})_*(1|\,j\rangle),\, \sigma\Big).$$

It makes sense to generalise these surprises from a point distribution $1|\,j\rangle$ to an arbitrary distribution τ on the domain of the channel. This then gives prediction and update surprises with respect to τ of the form:

$$D_{KL}\Big(\tau,\, c_*(\sigma)\Big) \qquad\qquad D_{KL}\Big((c^{\dagger}_{\sigma})_*(\tau),\, \sigma\Big). \tag{7.5}$$

The update surprise (on the right) involves the divergence between what we later call the Jeffrey posterior and its prior. Using the inequality from Proposition 2.10.6 (3) we see that the update surprise is always below the prediction surprise:

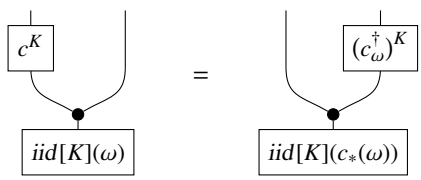
$$D_{KL}\Big((c^{\dagger}_{\sigma})_*(\tau),\, \sigma\Big) \overset{(7.4)}{=} D_{KL}\Big((c^{\dagger}_{\sigma})_*(\tau),\, (c^{\dagger}_{\sigma})_*(c_*(\sigma))\Big) \leq D_{KL}\Big(\tau,\, c_*(\sigma)\Big).$$

This prediction surprise (or error) on the left in (7.5) will play an important role in the sequel. This error is decreased by Jeffrey's update rule, see Theorem 7.9.3 (2). Also, expectation maximisation is all about decreasing such prediction errors, see Chapter 8.

The next result (from Jacobs and Stein (2023c)) shows that the dagger (also) interacts nicely with the extension $\mathcal{N}[K]\colon \mathbf{Chan} \to \mathbf{Chan}$ of the multinomial functor to the category of channels, see Corollary 3.8.7.

Theorem 7.1.10 *Let $c\colon X \to Y$ be a channel to a finite set Y, with a distribution $\omega \in \mathcal{D}(X)$ such that $c_*(\omega)$ has full support. Fix a number $K \in \mathbb{N}$.*

1 *The K-fold tensor product of channels $c^K = c \otimes \cdots \otimes c\colon X^K \rightsquigarrow Y^K$ satisfies the following equation.*

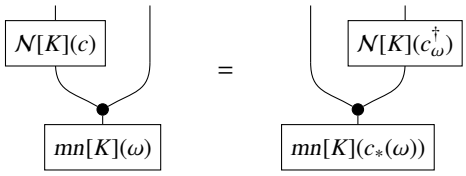


(7.6)

As a result, the dagger commutes with K-fold tensor product:

$$(c^K)^{\dagger}_{iid[K](\omega)} = (c^{\dagger}_{\omega})^K. \tag{7.7}$$

2 *The multiset extension of channels* $\mathcal{N}[K](c)\colon \mathcal{N}[K](X) \rightsquigarrow \mathcal{N}[K](Y)$ *satisfies:*



(7.8)

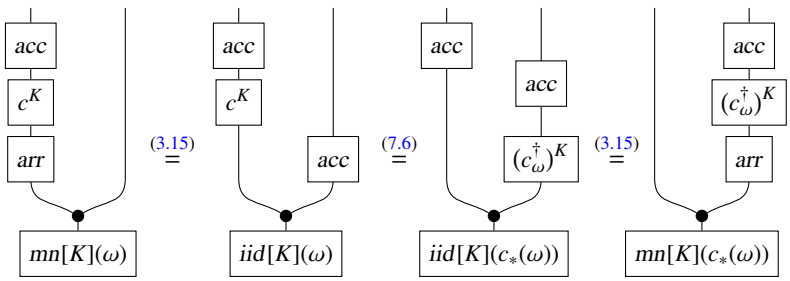
We can now conclude that the dagger commutes with multiset-extension:

$$\mathcal{N}[K](c)^\dagger_{mn[K](\omega)} = \mathcal{N}[K](c^\dagger_\omega). \tag{7.9}$$

Proof. 1 We start on the right-hand-side of (7.6):

$$\begin{aligned}
\langle id, (c^\dagger_\omega)^K\rangle_*(iid[K](c_*(\omega))) &= \sum_{\vec{y}\in Y^K}\sum_{\vec{x}\in X^K} (c^\dagger_\omega)^K(\vec{y})(\vec{x}) \cdot iid[K](c_*(\omega)(\vec{y}))\,\big|\vec{y},\vec{x}\rangle \\
&= \sum_{\vec{y}\in Y^K}\sum_{\vec{x}\in X^K}\prod_{1\leq i\leq K} c^\dagger_\omega(y_i)(x_i)\cdot c_*(\omega)(y_i)\,\big|\vec{y},\vec{x}\rangle \\
&= \sum_{\vec{y}\in Y^K}\sum_{\vec{x}\in X^K}\prod_{1\leq i\leq K} \frac{\omega(x_i)\cdot c(x_i)(y_i)}{c_*(\omega)(y_i)}\cdot c_*(\omega)(y_i)\,\big|\vec{y},\vec{x}\rangle \\
&= \sum_{\vec{y}\in Y^K}\sum_{\vec{x}\in X^K}\prod_{1\leq i\leq K} c(x_i)(y_i)\cdot\omega(x_i)\,\big|\vec{y},\vec{x}\rangle \\
&= \sum_{\vec{y}\in Y^K}\sum_{\vec{x}\in X^K} c^K(\vec{x})(\vec{y})\cdot iid[K](\omega)(\vec{x})\,\big|\vec{y},\vec{x}\rangle \\
&= \langle id, c^K\rangle_*(iid[K](\omega)).
\end{aligned}$$

2 We recall that $\mathcal{N}[K](c) = acc \circ c^K \circ arr\colon \mathcal{N}[K](X) \rightsquigarrow X^K \rightsquigarrow Y^K \rightsquigarrow \mathcal{N}[K](Y)$, see Corollary 3.8.7. We use the previous point, in combination with the *acc*-*arr* equation (3.15).



A direct proof, in the style of item (1), is much harder. ❑

Notice that the naturality of the $iid[K]$ and $mn[K]$ channels — with respect product and multiset functors extended to the category **Chan**, in Exercise 2.4.10 (4) and in Theorem 3.8.11 — is a consequence of Equations (7.6) and (7.8), via marginalising out the second wires. Concretely, in this way one obtains the first equation in (3.44). In fact, marginalising out the first wires also gives naturality equations, in combination with (7.4).

Exercises

7.1.1 Consider the sets $X = \{1, 2\}$ and $Y = \{a, b, c\}$ with channel $c\colon X \rightsquigarrow Y$ given by:

$$c(1) = \tfrac{1}{2}|a\rangle + \tfrac{1}{3}|b\rangle + \tfrac{1}{6}|c\rangle \qquad\qquad c(2) = \tfrac{1}{4}|a\rangle + \tfrac{3}{8}|b\rangle + \tfrac{1}{8}|c\rangle.$$

Consider the following two distributions on X.

$$\omega_1 \;=\; \tfrac{1}{2}|1\rangle + \tfrac{1}{2}|2\rangle \qquad\qquad \omega_1 \;=\; \tfrac{1}{3}|1\rangle + \tfrac{2}{3}|2\rangle.$$

1 Check that $c_*(\omega_1)$ and $c_*(\omega_2)$ both have full support.
2 Compute the dagger channels $c^\dagger_{\omega_1}$ and $c^\dagger_{\omega_2}$.

7.1.2 In Exercises 6.2.1 and 6.2.2 we have seen a channel $c\colon \{d, d^\perp\} \rightsquigarrow \{p, n\}$ that combines the sensitivity and specificity of a medical test, in a situation with a prior / prevalence of 1% for the disease. Show that the associated Positive Prediction Value (PPV) and Negative Predication Value (NPV) can be expressed via a dagger as:

$$\begin{aligned} PPV \;&=\; c^\dagger_\omega(p)(d) \;=\; \omega|_{c^*(\mathbf{1}_p)}(d) \;=\; \tfrac{18}{117} \\ NPV \;&=\; c^\dagger_\omega(n)(d^\perp) \;=\; \omega|_{c^*(\mathbf{1}_n)}(d^\perp) \;=\; \tfrac{1881}{1883}, \end{aligned}$$

where $\omega = \frac{1}{100}|d\rangle + \frac{99}{100}|d^\perp\rangle$ is the prior distribution.

7.1.3 Let $f\colon X \to Y$ be a function, considered as deterministic channel. Show that for arbitrary distributions $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$, when everything is well-defined,

$$\left(f^\dagger_\sigma\right)^\dagger_\tau \;=\; f.$$

First from Lemma 7.1.4 (2) we get:

$$f^\dagger_\sigma(y) \;=\; \sum_{x \in f^{-1}(y)} \frac{\sigma(x)}{\mathcal{D}(f)(\sigma)(y)}\,\big|x\rangle.$$

7.1.4 We have introduced daggers for $\mathcal{D}$-channels, but they work as well in the infinite case, for $\mathcal{D}_\infty$-channels. For $\lambda \in \mathbb{R}_{\geq 0}$ define the shift-Poisson channel $sp[\lambda]\colon \mathbb{N} \to \mathcal{D}_\infty(\mathbb{N})$ as:

$$sp[\lambda](n) \;=\; shift(n, pois[\lambda]) \;=\; 1|n\rangle + pois[\lambda] \;=\; \sum_{m \in \mathbb{N}} e^{-\lambda} \cdot \frac{\lambda^m}{m!}\,\big|n+m\rangle.$$

1 Check, for $\mu \in \mathbb{R}_{\geq 0}$, like in Exercise 2.7.8, the first equation in:

$$sp[\lambda]_*(pois[\mu]) \;=\; pois[\lambda] + pois[\mu] \;=\; pois[\lambda + \mu].$$

The second equation comes from Proposition 2.7.9.

2 Now prove that the binomial distribution arises via the dagger of the shift-Poisson channel, as:

$$sp[\lambda]^\dagger_{pois[\mu]}(m) \;=\; bn[m]\big(\tfrac{\mu}{\lambda+\mu}\big).$$

This is not a very familiar result, but it can be found for instance in the online Cambridge lecture notes[1] of Richard Weber. Here it is reformulated in terms of a dagger channel. Similar daggers appear in Section 7.6.

7.1.5 Let $c\colon X \rightsquigarrow Y$ be a channel with distribution $\omega \in \mathcal{D}(X)$, where $c_*(\omega)$ has full support. Consider the following distribution of distributions:

$$\Omega \;\coloneqq\; \sum_{y \in Y} c_*(\omega)(y)\big|c^\dagger_\omega(y)\rangle \;\in\; \mathcal{D}(\mathcal{D}(X)).$$

[1] Specifically, in Example 5.2 in: www.statslab.cam.ac.uk/~rrw1/stats/S05.pdf.

Recall the probabilistic 'flatten' operation from Section 2.4 and show:

$$\mathit{flat}(\Omega) \;=\; \omega.$$

This says that Ω is 'Bayes plausible' in the terminology of Kamenica and Gentzkow (2011), in the context of Bayesian persuasion. The construction of Ω is one part of the bijective correspondence in Exercise 7.2.5.

7.1.6 Consider the draw-delete channel $DD\colon \mathcal{N}[K{+}1](X) \rightsquigarrow \mathcal{N}[K](X)$ from Definition 3.2.1 (2). Let X be a finite set, say with n elements. Consider the the uniform state $\upsilon := \mathit{unif}_{\mathcal{N}[K+1](X)}$ on $\mathcal{N}[K+1](X)$, see Exercise 2.4.5. Show that DD's dagger, with respect to υ, can be described on $\varphi \in \mathcal{N}[K](X)$ as:

$$DD^{\dagger}_{\upsilon}(\varphi) \;=\; \sum_{x \in X} \frac{\varphi(x)+1}{K+n} \,\big|\varphi + 1|x\rangle\big\rangle.$$

This dagger differs from the draw-add map $DA : \mathcal{N}[K](X) \rightsquigarrow \mathcal{N}[K+1](X)$.

7.1.7 Use (7.9) and Exercise 3.8.6 to prove:

$$\mathit{Flrn} \circ \mathcal{N}[K](c)^{\dagger}_{mn[K](\omega)} \;=\; c^{\dagger}_{\omega} \circ \mathit{Flrn}.$$

7.1.8 Let $c\colon X \rightsquigarrow Y$ be a channel with distribution $\omega \in \mathcal{D}(X)$. Recall the description (2.79) of mutual information $MI(\tau)$ of a joint distribution τ as the KL-divergence between τ and the product of its marginals.

1 Consider the graph distribution $\langle id, c\rangle_*(\omega) \in \mathcal{D}(X \times Y)$ and show that its mutual information can be expressed in the following four (equal) ways.

$$\begin{aligned} MI\Big(\langle id, c\rangle_*(\omega)\Big) &= H(c_*(\omega)) - \Big(\omega \models H \circ c\Big) \\ &= H(\omega) - \Big(c_*(\omega) \models H \circ c^{\dagger}_{\omega}\Big) \\ &= \omega \models D_{KL}(c(-),\, c_*(\omega)) \\ &= c_*(\omega) \models D_{KL}(c^{\dagger}_{\omega}(-),\, \omega). \end{aligned}$$

This is sometimes called the mutual information between the channel input and channel output, see Blahut (1972); Arimoto (1972). These equations occur in different notation for instance in (Bishop and Bishop, 2024, Eqn. (2.110)). The maximum over all $\omega \in \mathcal{D}(X)$ is the capacity of the channel, see Subsection 6.12.1 and *e.g.* Cover and Thomas (2006).

2 Check that:

$$MI\Big(\langle id, c\rangle_*(\omega)\Big) \;=\; MI\Big(\langle c, id\rangle_*(\omega)\Big).$$

3 Prove the following equality of mutual informations, for a channel and its dagger.

$$MI\Big(\langle id, c\rangle_*(\omega)\Big) \;=\; MI\Big(\langle id, c^{\dagger}_{\omega}\rangle_*(c_*(\omega))\Big).$$

7.1.9 Let p be a predicate on X, considered as a channel $p\colon X \rightsquigarrow \mathbf{2}$, and let $\omega \in \mathcal{D}(X)$ be a distribution. Compute the dagger $p^{\dagger}_{\omega}\colon \mathbf{2} \rightsquigarrow X$ and show that:

$$p^{\dagger}_{\omega}(1) \;=\; \omega|_p \qquad \text{and} \qquad p^{\dagger}_{\omega}(0) \;=\; \omega|_{p^{\perp}}. \tag{7.10}$$

7.1.10 Let $\omega \in \mathcal{D}(X)$ be a state.

1 Consider a predicate $p\colon X \to [0,1]$ as a channel, like in the previous exercise. Show that the binary version of the law of total probability, see Exercise 6.1.6, can be expressed as:

$$\omega \;=\; (p^{\dagger}_{\omega})_{*}\Big((\omega \models p)|1\rangle + (\omega \models p^{\perp})|0\rangle\Big).$$

2 Let $c\colon X \rightsquigarrow n$ be a channel and define n predicates on X by $p_i := c^{*}(\mathbf{1}_i)$. Show that these $p_0, \ldots, p_{n-1}$ form a test:

$$\bigvee_i p_i \;=\; \mathbf{1} \qquad \text{with} \qquad c^{\dagger}_{\omega}(i) \;=\; \omega|_{p_i}.$$

3 Conclude that the equation $(c^{\dagger}_{\omega})_{*}(c_{*}(\omega)) = \omega$ from Equation 7.4 can also be understood as the (n-ary version of the) law of total probability.

7.1.11 Let $c\colon X \times Y \rightsquigarrow Z$ be a channel with two distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ such that $c_{*}(\omega \otimes \rho) \in \mathcal{D}(Z)$ has full support. Write $c_{\rho} = c \circ (id \otimes \rho)\colon X \cong X \times \mathbf{1} \rightsquigarrow X \times Y \rightsquigarrow Z$, where ρ is used as channel $\rho\colon \mathbf{1} \rightsquigarrow Y$. Show that the following two channels $Z \rightsquigarrow X$ are equal.

$$\begin{tikzcd} Z \arrow[r, bend left, "c^{\dagger}_{\omega\otimes\rho}"] \arrow[rr, bend right, "(c_{\rho})^{\dagger}_{\omega}"'] & X \times Y \arrow[r, bend left, "\pi_1"] & X \end{tikzcd}$$

7.1.12 Let $c\colon X \rightsquigarrow Y$ and $d\colon X \rightsquigarrow Z$ be channels with a distribution $\omega \in \mathcal{D}(X)$ on their domain such that both pushforwards $c_{*}(\omega)$ and $d_{*}(\omega)$ have full support. Show that:

$$\begin{aligned} \langle c, d\rangle^{\dagger}_{\omega}(y,z) \;&=\; \omega|_{c^{*}(\mathbf{1}_y)\,\&\,d^{*}(\mathbf{1}_z)} \;=\; d^{\dagger}_{c^{\dagger}_{\omega}(y)}(z) \\ &=\; \omega|_{d^{*}(\mathbf{1}_z)\,\&\,c^{*}(\mathbf{1}_y)} \;=\; c^{\dagger}_{d^{\dagger}_{\omega}(z)}(y). \end{aligned}$$

7.1.13 In (Goy and Petrişan, 2020, Lem. 3.3) and also in (Fritz and Perrone, 2020, Prop. 6.4) it is shown that the following naturality diagram for flattening is a so-called weak pullback.

$$\begin{tikzcd} \mathcal{D}^2(X) \arrow[r, "\mathcal{D}^2(f)"] \arrow[d, "flat"'] & \mathcal{D}^2(Y) \arrow[d, "flat"] \\ \mathcal{D}(X) \arrow[r, "\mathcal{D}(f)"] & \mathcal{D}(Y) \end{tikzcd}$$

This means that given $\omega \in \mathcal{D}(X)$ and $\Omega \in \mathcal{D}^2(Y)$ with $\mathcal{D}(f)(\omega) = flat(\Omega)$, there is a $\Psi \in \mathcal{D}^2(X)$ with $flat(\Psi) = \omega$ and $\mathcal{D}^2(f)(\Psi) = \Omega$.

Check that one can take $\Psi = \mathcal{D}((f^{\dagger}_{\omega})_{*})(\Omega)$. This is essentially as in Fritz and Perrone (2020) but the dagger formulation is not used there.

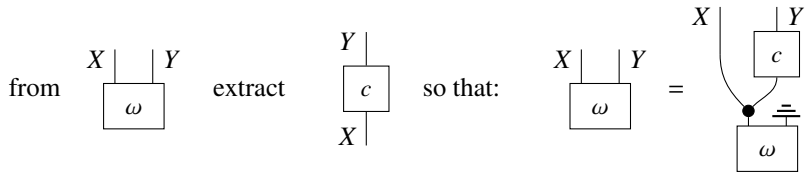
7.1.14 In Definition 6.7.4 we have seen filtering for hidden Markov models, starting from a sequence of factors $\vec{p}$ as observations. In practice these factors are often point predicates $\mathbf{1}_{y_i}$ for a sequence of elements $\vec{y}$ of the visible space Y. Show that in that case one can describe filtering via Bayesian inversion as follows, for a hidden Markov model with transition channel $t\colon X \rightsquigarrow X$, emission channel $e\colon X \rightsquigarrow Y$ and initial state $\sigma \in \mathcal{D}(X)$.

$$\sigma_1 \;=\; \sigma \qquad \text{and} \qquad \sigma_{i+1} \;=\; (t \circ e^{\dagger}_{\sigma_i})(y_i).$$

7.2 Disintegration of joint distributions

In Subsections 1.5.3 and 1.6.3 we have seen how a binary relation $R \in \mathcal{P}(A \times B)$ on $A \times B$ corresponds to a $\mathcal{P}$-channel $A \to \mathcal{P}(B)$, and similarly how a multiset $\psi \in \mathcal{M}(A \times B)$ corresponds to an $\mathcal{M}$-channel $A \to \mathcal{M}(B)$. The phenomenon was called: extraction of a channel from a joint state. This section describes the analogue for probabilistic binary / joint states and channels (like in Fritz (2020)). It is called disintegration. It turns out to be more subtle (than for $\mathcal{P}$ and $\mathcal{M}$) because probabilistic extraction requires normalisation — in order to ensure the unitality requirement of a $\mathcal{D}$-channel: multiplicities must add up to one.

In this section we look at disintegration, first for joint distributions. In the next section we formulate a more general version, for channels, with multiple incoming and outgoing wires. The graphical language of string diagrams is useful for understanding the essentials: disintegration for joint states involves:



(7.11)

It turns out that the extracted channel c can be obtained via a dagger of a projection. This will be shown first.

Theorem 7.2.1 (Disintegration) *Consider the two projections $X \xleftarrow{\pi_1} X \times Y \xrightarrow{\pi_2} Y$ as deterministic channels. Let $\omega \in \mathcal{D}(X \times Y)$ be a joint distribution, whose two marginals $\omega_1 := \omega[1,0] = (\pi_1)_*(\omega) \in \mathcal{D}(X)$ and $\omega_2 := \omega[0,1] = (\pi_2)_*(\omega) \in \mathcal{D}(Y)$ both have full support.*

Extract from the joint distribution ω the two channels:

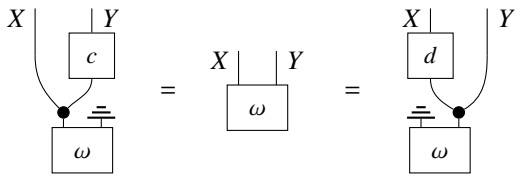
$$c := \Big(X \xrightarrow{(\pi_1)^\dagger_\omega} X \times Y \xrightarrow{\pi_2} Y\Big) \qquad\qquad d := \Big(Y \xrightarrow{(\pi_2)^\dagger_\omega} X \times Y \xrightarrow{\pi_1} X\Big)$$

Then:

1 *These channels c and d can be described explicitly via crossover update as:*

$$\begin{aligned} c(x) &:= (\omega|_{\mathbf{1}_x \otimes \mathbf{1}})[0,1] \\ &= \sum_{y \in Y} \frac{\omega(x,y)}{\omega_1(x)} \,|y\rangle \end{aligned} \qquad\qquad \begin{aligned} d(y) &:= (\omega|_{\mathbf{1} \otimes \mathbf{1}_x})[1,0] \\ &= \sum_{x \in X} \frac{\omega(x,y)}{\omega_2(y)} \,|x\rangle. \end{aligned} \tag{7.12}$$

2 *The joint distribution ω is the graph of both these channels: $\langle id, c\rangle_*(\omega_1) = \omega = \langle d, id\rangle_*(\omega_2)$, that is:*



(7.13)

3 *The extracted channels are each other's daggers:*

$$c^\dagger_{\omega_1} = d \qquad \text{and} \qquad d^\dagger_{\omega_2} = c.$$

We sometimes use the following notation for these extracted channels:

$$\omega[0,1\,|\,1,0] = c : X \twoheadrightarrow Y \qquad \textit{and} \qquad \omega[1,0\,|\,0,1] = d : Y \twoheadrightarrow X. \tag{7.14}$$

This notation with two masks with a bar | inbetween mimics the notation for marginalisation from Subsection 2.3.2, with a single mask. The mask numbers $1,0$ on the right of the bar, for the channel c, describe the domain of the extracted channel in relation to the product space $X \times Y$ of the joint distribution: the X is kept as first component, but not the Y. Similarly, the mask $0,1$ on the left of the bar refer to the codomain of the channel and indicates that X is dropped but Y remains.

Proof. For all three items we only do the first equations, since the second ones follow by symmetry.

1 By assumption $(\pi_1)_*(\omega) = \omega_1$ has full support, so we can form the dagger channel $c := (\pi_1)^\dagger_\omega$. It is:

$$c(x) \overset{(7.1)}{=} \sum_{x' \in X} \sum_{y \in Y} \frac{\omega(x',y) \cdot \langle \pi_1 \rangle(x',y)(x)}{(\pi_1)_*(\omega)(x)} \,|y\rangle = \sum_{y \in Y} \frac{\omega(x,y)}{\omega_1(x)} \,|y\rangle.$$

2 Using this formulation, we get, for $x \in X$ and $y \in Y$,

$$\langle id, c \rangle_*(\omega_1)(x,y) = \omega_1(x) \cdot c(x)(y) = \omega(x,y).$$

3 Similarly,

$$\begin{aligned} c^\dagger_{\omega_1}(y)(x) &\overset{(7.1)}{=} \frac{\omega_1(x) \cdot c(x)(y)}{c_*(\omega_1)(y)} \overset{(7.12)}{=} \frac{\omega(x,y)}{\sum_{x \in X} \omega_1(x) \cdot c(x)(y)} = \frac{\omega(x,y)}{\sum_{x \in X} \omega(x,y)} \\ &= \frac{\omega(x,y)}{\omega_2(y)} \overset{(7.12)}{=} d(y)(x). \end{aligned}$$

❑

From a joint distribution $\omega \in \mathcal{D}(X \times Y)$ we can extract two channels, in both direction, written as $\omega[0,1\,|\,1,0] \colon X \twoheadrightarrow Y$ and $\omega[1,0\,|\,0,1] \colon Y \twoheadrightarrow X$, see (7.14). The direction of the channels is thus in a certain sense arbitrary, and does not reflect any form of causality, see Chapter 9.

Example 7.2.2 The following question is copied from (Grinstead and Snell, 1998, Exc. 4.1.29).

A student is applying to Harvard and Dartmouth. He estimates that he has a probability of .5 of being accepted at Dartmouth and .3 of being accepted at Harvard. He further estimates the probability that he will be accepted by both is .2. What is the probability that he is accepted by Dartmouth if he is accepted by Harvard? Is the event "accepted at Harvard" independent of the event "accepted at Dartmouth"?

We formalise the situation via two sets $\{H, H^\perp\}$ and $\{D, D^\perp\}$ for being accepted or not at Harvard and Dartmouth. There is a joint acceptance distribution $\alpha \in \mathcal{D}(\{H, H^\perp\} \times \{D, D^\perp\})$ which we write as:

$$\alpha = \tfrac{1}{5}|H,D\rangle + r_1|H,D^\perp\rangle + r_2|H^\perp,D\rangle + r_3|H^\perp,D^\perp\rangle.$$

It is given that $\frac{1}{5} + r_2 = \frac{1}{2}$, so $r_2 = \frac{3}{10}$. Similarly, $\frac{1}{5} + r_1 = \frac{3}{10}$, so $r_1 = \frac{1}{10}$. But then $r_3 = \frac{2}{5}$. Thus we fully know the acceptance distribution as:

$$\alpha = \tfrac{1}{5}|H,D\rangle + \tfrac{1}{10}|H,D^\perp\rangle + \tfrac{3}{10}|H^\perp,D\rangle + \tfrac{2}{5}|H^\perp,D^\perp\rangle.$$

We can now extract a channel $c := \alpha[0,1 \,|\, 1,0] \colon \{H, H^\perp\} \rightsquigarrow \{D, D^\perp\}$. Following (7.12) we get:

$$\begin{aligned}
c(H) &= \frac{\alpha(H,D)}{\alpha(H,D)+\alpha(H,D^\perp)}|D\rangle + \frac{\alpha(H,D^\perp)}{\alpha(H,D)+\alpha(H,D^\perp)}|D^\perp\rangle \\
&= \frac{1/5}{1/5+1/10}|D\rangle + \frac{1/10}{1/5+1/10}|D\rangle \;=\; \tfrac{2}{3}|D\rangle + \tfrac{1}{3}|D^\perp\rangle \\
c(H^\perp) &= \frac{\alpha(H^\perp,D)}{\alpha(H^\perp,D)+\alpha(H^\perp,D^\perp)}|D\rangle + \frac{\alpha(H^\perp,D^\perp)}{\alpha(H^\perp,D)+\alpha(H^\perp,D^\perp)}|D^\perp\rangle \\
&= \frac{3/10}{3/10+2/5}|D\rangle + \frac{2/5}{3/10+2/5}|D\rangle \;=\; \tfrac{3}{7}|D\rangle + \tfrac{4}{7}|D^\perp\rangle.
\end{aligned}$$

Thus, the probability $c(H)(D)$ that the student is accepted by Dartmouth if he is accepted by Harvard is $\frac{2}{3}$. This is often written as a conditional probability $P(D \mid H)$.

The question about the independence of the events "accepted at Harvard" and "accepted at Dartmouth" translates into the question of non-entwinedness of the joint distribution α. This is not the case, since the first and second marginals of α are

$$\alpha[1,0] \;=\; \tfrac{3}{10}|H\rangle + \tfrac{7}{10}|H^\perp\rangle \qquad\qquad \alpha[0,1] \;=\; \tfrac{1}{2}|D\rangle + \tfrac{1}{2}|D^\perp\rangle.$$

The joint distribution α differs from the product of its marginals, which is:

$$\alpha[1,0] \otimes \alpha[0,1] \;=\; \tfrac{3}{20}\big|H,D\rangle + \tfrac{3}{20}\big|H,D^\perp\rangle + \tfrac{7}{20}\big|H^\perp,D\rangle + \tfrac{7}{10}\big|H^\perp,D^\perp\rangle.$$

Exercise 7.2.4 describes non-entwinedness of a joint distribution in terms of its extracted channels being constant. Indeed, the above channel c is not constant, that is $c(H) \neq c(H^\perp)$.

We have seen that disintegration can be defined in terms of daggers. This also works the other way around: daggers can be obtained via disintegration.

Corollary 7.2.3 *Let $c \colon X \rightsquigarrow Y$ be channel with a distribution $\sigma \in \mathcal{D}(X)$ such that $c_*(\sigma)$ has full support. Write $\omega := \langle id, c\rangle_*(\sigma) \in \mathcal{D}(X \times Y)$ for the graph distribution. The channel $\omega[1,0 \,|\, 0,1] \colon Y \rightsquigarrow X$ extracted via disintegration is then the dagger $c^\dagger_\sigma$.*

Proof. The distribution $(\pi_2)_*(\omega) = c_*(\sigma) \in \mathcal{D}(Y)$ has full support. The extraction $\omega[1,0 \,|\, 0,1] \colon Y \rightsquigarrow X$ is thus well-defined. It is the dagger of c by Theorem 7.2.1 (2) and (a swapped version of) Theorem 7.1.6. ❑

Example 7.2.4 We shall look at two examples, involving spaces $A = \{a, a^\perp\}$ and $B = \{b, b^\perp\}$.

1 Consider the following state $\omega \in \mathcal{D}(A \times B)$,

$$\omega \;=\; \tfrac{1}{4}|a,b\rangle + \tfrac{1}{2}|a,b^\perp\rangle + \tfrac{1}{4}|a^\perp,b^\perp\rangle.$$

We have as first marginal $\omega_1 := \omega[1,0] = \frac{3}{4}|a\rangle + \frac{1}{4}|a^\perp\rangle$ with full support. The extracted channel $c := \omega[0,1 \,|\, 1,0] \colon A \rightsquigarrow B$ is given by:

$$\begin{aligned}
c(a) &\overset{(7.12)}{=} \frac{\omega(a,b)}{\omega_1(a)}|b\rangle + \frac{\omega(a,b^\perp)}{\omega_1(a)}|b^\perp\rangle \;=\; \frac{1/4}{3/4}|b\rangle + \frac{1/2}{3/4}|b^\perp\rangle \;=\; \tfrac{1}{3}|b\rangle + \tfrac{2}{3}|b^\perp\rangle \\
c(a^\perp) &\overset{(7.12)}{=} \frac{\omega(a^\perp,b)}{\omega_1(a^\perp)}|b\rangle + \frac{\omega(a^\perp,b^\perp)}{\omega_1(a^\perp)}|b^\perp\rangle \;=\; \frac{0}{1/4}|b\rangle + \frac{1/4}{1/4}|b^\perp\rangle \;=\; 1|b^\perp\rangle.
\end{aligned}$$

Then indeed, $\langle id, c\rangle_*(\omega_1) = \omega$.

2 Now let us start from a different joint distribution:

$$\omega = \tfrac{1}{3}|a,b\rangle + \tfrac{2}{3}|a,b^{\perp}\rangle.$$

Then $\omega_1 := \omega[1,0] = 1|a\rangle$. It does not have full support. Let $\tau \in \mathcal{D}(B)$ be an arbitrary distribution. We define $c\colon A \rightsquigarrow B$ as:

$$c(a) = \tfrac{1}{3}|b\rangle + \tfrac{2}{3}|b^{\perp}\rangle \qquad \text{and} \qquad c(a^{\perp}) = \tau.$$

We then still get an equation $\langle id, c\rangle_*(\omega_1) = \omega$, no matter what τ is.

More generally, if we do not have full support, disintegrations may still exist, but they are not unique. We generally avoid such non-uniqueness by requiring full support.

Remark 7.2.5 For a joint distribution $\tau \in \mathcal{D}(X \times Y)$ there is a notion of 'conditional entropy'. It is defined in terms of the extracted channel $\tau[1,0\,|\,0,1]\colon X \rightsquigarrow Y$ as:

$$-\sum_{x\in X,\, y\in Y} \tau(x,y) \cdot \ln\Big(\tau[1,0\,|\,0,1](x)(y)\Big).$$

When we write $\tau = \langle id, c\rangle_*(\omega)$, for $c := \tau[1,0\,|\,0,1]\colon X \rightsquigarrow Y$ and $\omega = \omega[1,0] \in \mathcal{D}(X)$, then we can reformulate this conditional entropy as validity:

$$\omega \models (H \circ c),$$

where the observable $H \circ c\colon X \to \mathbb{R}_{\geq 0}$ is obtained via the (ordinary) entropy function $H\colon \mathcal{D}(Y) \to \mathbb{R}_{\geq 0}$. Indeed:

$$\begin{aligned}
\omega \models (H \circ c) = \sum_{x\in X} \omega(x) \cdot H(c(x)) &= \sum_{x\in X} \omega(x) \cdot -\sum_{y\in Y} c(x)(y) \cdot \ln\big(c(x)(y)\big) \\
&= -\sum_{x\in X,\, y\in Y} \omega(x) \cdot c(x)(y) \cdot \ln\big(c(x)(y)\big) \\
&= -\sum_{x\in X,\, y\in Y} \tau(x,y) \cdot \ln\Big(\tau[1,0\,|\,0,1](x)(y)\Big).
\end{aligned}$$

Example 7.2.6 The *natural join* $\bowtie$ is a basic construction in database theory that makes it possible to join two databases which coincide on their overlap. Such natural joins are used in a probabilistic setting in Abramsky (2016); Bruza and Abramsky (2016) — in particular in relation to Bell tables, see Exercise 2.4.19.

Let $\omega \in \mathcal{D}(X \otimes Y)$ and $\rho \in \mathcal{D}(X \otimes Z)$ be two joint distributions with equal first marginals: $\omega[1,0] = \rho[1,0]$. A *natural join*, if it exists, is a distribution $\omega \bowtie \rho \in \mathcal{D}(X \otimes Y \otimes Z)$ which marginalises both to ω and two ρ, as in:

$$(\omega \bowtie \rho)[1,1,0] = \omega \qquad \text{and} \qquad (\omega \bowtie \rho)[1,0,1] = \rho.$$

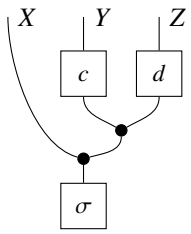
One can understand such a natural join as a parameterised form of coupling.

We show how such natural joins can be constructed via disintegration. Let's assume we have joint distributions $\omega \in \mathcal{D}(X \times Y)$ and $\rho \in \mathcal{D}(X \times Z)$ as described above, with common marginal written as $\sigma := \omega[1,0] = \rho[1,0]$. We extract channels:

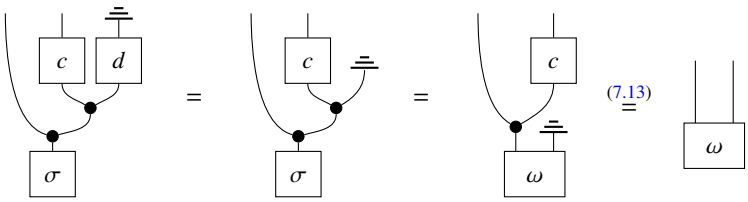
$$c := \omega[0,1\,|\,1,0] : X \rightsquigarrow Y \qquad\qquad d := \rho[0,1\,|\,1,0] : X \rightsquigarrow Z.$$

Now we define:

$$\omega \bowtie \rho \quad := \tag{7.15}$$



The equation $(\omega \bowtie \rho)[1,1,0] = \omega$ can be obtained easily via diagrammatic reasoning:



Similarly one proves $(\omega \bowtie \rho)[1,0,1] = \rho$. For a concrete instantiation of this construction, see Exercise 7.2.3. Such natural joins are typically non-trivial, but the above construction (7.15) gives a clear recipe. The graphical approach emphasises the relevant flows and is especially useful in more complicated situations, with multiple distributions which agree on multiple marginals.

The next result illustrates two bijective correspondences resulting from disintegration. The first bijective correspondence is basically a reformulation of disintegration itself (for states). The second one is new and involves distributions over distributions — sometimes called hyperdistributions McIver et al. (2015, 2014).

Theorem 7.2.7 *Let X, Y be two sets, where Y is finite.*

1 *There is a bijective correspondence between:*

$$\frac{\tau \in \mathcal{D}(X \times Y) \text{ where } \tau[0,1] \text{ has full support}}{\omega \in \mathcal{D}(X) \text{ and } c \colon X \rightsquigarrow Y \text{ such that } c_*(\omega) \text{ has full support}}$$

2 *For each natural number $N \geq 1$ there is a bijective correspondence between:*

$$\frac{\Omega \in \mathcal{D}(\mathcal{D}(X)) \text{ with } \big|supp(\Omega)\big| = N}{\rho \in \mathcal{D}(X \times \boldsymbol{N}) \text{ where } \rho[0,1] \text{ has full support}}$$

Proof. 1 In the downward direction, starting from a joint distribution $\tau \in \mathcal{D}(X \times Y)$ we take $\omega := \tau[1,0] \in \mathcal{D}(X)$ as first marginal and we extract a channel $c := \tau[0,1 \,\big|\, 1,0] \colon X \rightsquigarrow Y$ via disintegration. The latter exists because the second marginal $\tau[0,1]$ has full support. In the upward direction we transform a state-channel pair ω, c to the joint state $\tau := \langle id, c\rangle_*(\omega) \in \mathcal{D}(X \times Y)$. By assumption, $\tau[0,1] = c_*(\omega)$ has full support. Doing these transformations twice, both down-up and up-down, yields the orginal data, by definition of disintegration.

2 Let distribution $\Omega \in \mathcal{D}(\mathcal{D}(X))$ have a support with N elements, say $supp(\Omega) = \{\omega_0, \ldots, \omega_{N-1}\}$ with $\omega_i \in \mathcal{D}(X)$. We take:

$$\rho(x,i) := \Omega(\omega_i) \cdot \omega_i(x). \tag{7.16}$$

This yields a distribution $\rho \in \mathcal{D}(X \times \boldsymbol{N})$ since these probabilities add up to one:

$$\sum_{x \in X,\, i \in \boldsymbol{N}} \rho(x,i) \coloneqq \sum_{i \in \boldsymbol{N}} \Omega(\omega_i) \cdot \sum_{x \in X} \omega_i(x) = \sum_{i \in \boldsymbol{N}} \Omega(\omega_i) = 1.$$

Furthermore, ρ's second marginal has full support, since for each $i \in \boldsymbol{N}$,

$$\rho[0,1](i) = \sum_{x \in X} \rho(x,i) = \Omega(\omega_i) \cdot \sum_{x \in X} \omega_i(x) = \Omega(\omega_i) > 0.$$

In the upward direction, starting from $\rho \in \mathcal{D}(X \times \boldsymbol{N})$ we from the channel $d \coloneqq \rho[1,0 \,|\, 0,1] \colon \boldsymbol{N} \rightsquigarrow X$ by disintegration and use it to define $\Omega \in \mathcal{D}(\mathcal{D}(X))$ as:

$$\Omega \coloneqq \sum_{i \in \boldsymbol{N}} \rho[0,1](i) \,|\, d(i) \rangle. \tag{7.17}$$

Since $\rho[0,1](i) > 0$ for each $i \in \boldsymbol{N}$, the support of Ω has N elements.

Starting from Ω with support $\{\omega_0, \ldots, \omega_{N-1}\}$, we can get ρ as in (7.16) with extracted channel d satisfying:

$$d(i)(x) = \frac{\rho(x,i)}{\rho[0,1](i)} = \frac{\Omega(\omega_i) \cdot \omega_i(x)}{\Omega(\omega_i)} = \omega_i(x).$$

Hence the definition (7.17) yields the original state $\Omega \in \mathcal{D}(\mathcal{D}(X))$:

$$\sum_{i \in \boldsymbol{N}} \rho[0,1](i) \,|\, d(i) \rangle = \sum_{i \in \boldsymbol{N}} \Omega(\omega_i) \,|\, \omega_i \rangle = \Omega.$$

In the other direction, starting from a joint state $\rho \in \mathcal{D}(X \times \boldsymbol{N})$ whose second marginal has full support, we can form Ω as in (7.17) and turn it into a joint state again via (7.16), whose probability at (x,i) is:

$$\Omega(d(i)) \cdot d(i)(x) = \rho[0,1](i) \cdot \rho[1,0 \,|\, 0,1](i)(x) = \langle id, \rho[1,0 \,|\, 0,1] \rangle_*(\rho[0,1])(x,i) = \rho(x,i).$$

This last equation holds by disintegration. ❑

By combining the two correspondences in this theorem we can bijectively relate a 'hyper' distribution $\Omega \in \mathcal{D}(\mathcal{D}(X))$ and a state-channel pair $\omega \in \mathcal{D}(X)$, $c \colon X \rightsquigarrow Y$, see Exercise 7.2.5 below. This corresponce is used in Bayesian persuasion, see Kamenica and Gentzkow (2011), where the channel c is called a signal.

With disintegration well-understood we can formulate a follow-up of crossover results in Theorem 6.3.1. There we looked at *marginalisation* after update. Here we look at *extraction* of a channel after such an update, at the same position of the original channel. The newly extracted channel can be expressed in terms of an updated channel, see Definition 6.1.1 (3).

Theorem 7.2.8 *Let $c \colon X \rightsquigarrow Y$ be a channel with state $\omega \in \mathcal{D}(X)$ on its domain, and let $p \in \mathsf{Fact}(X)$ and $q \in \mathsf{Fact}(Y)$ be factors.*

1 The extraction on an updated graph state yields:

$$\langle id, c \rangle_*(\omega)\big|_{p \otimes q}[0,1 \,|\, 1,0] = c|_q \quad \textit{where} \quad c|_q(x) \coloneqq c(x)|_q.$$

2 And as a result:

$$\langle id, c \rangle_*(\omega)\big|_{p \otimes q} = \langle id, c|_q \rangle_*(\omega|_{p \,\&\, c^*(q)}).$$

Proof. 1 For $x \in X$ and $y \in Y$ we have:

$$\begin{aligned}
\langle id, c\rangle_*(\omega)\big|_{p\otimes q}[0,1 \,\big|\, 1,0](x)(y) &\overset{(7.12)}{=} \frac{\langle id, c\rangle_*(\omega)\big|_{p\otimes q}(x,y)}{\langle id, c\rangle_*(\omega)\big|_{p\otimes q}[1,0](x)} \\
&= \frac{\langle id, c\rangle_*(\omega)(x,y)\cdot(p\otimes q)(x,y)}{\sum_v \langle id, c\rangle_*(\omega)(x,v)\cdot(p\otimes q)(x,v)} \\
&= \frac{\omega(x)\cdot c(x)(y)\cdot p(x)\cdot q(y)}{\sum_v \omega(x)\cdot c(x)(v)\cdot p(x)\cdot q(v)} \\
&= \frac{c(x)(y)\cdot q(y)}{\sum_v c(x)(v)\cdot q(v)} \\
&= \frac{c(x)(y)\cdot q(y)}{c(x)\models q} \;=\; c(x)|_q(y) \;=\; c|_q(x)(y).
\end{aligned}$$

2 An (updated) joint state like $\langle id, c\rangle_*(\omega)|_{p\otimes q}$ can always be written as graph $\langle id, d\rangle_*(\tau)$. The channel d is obtained by distintegration of the joint state, and equals $c|_q$ by the previous point. The state τ is the first marginal of the joint state. Hence we are done by charactersing this first marginal, in:

$$\begin{aligned}
\langle id, c\rangle_*(\omega)\big|_{p\otimes q}[1,0] &= \langle id, c\rangle_*(\omega)\big|_{(\mathbf{1}\otimes q)\&(p\otimes\mathbf{1})}[1,0] && \text{by Exercise 4.5.9} \\
&= \langle id, c\rangle_*(\omega)\big|_{\mathbf{1}\otimes q}\big|_{p\otimes\mathbf{1}}[1,0] && \text{by Lemma 6.1.7 (3)} \\
&= \langle id, c\rangle_*(\omega)\big|_{\mathbf{1}\otimes q}[1,0]\big|_p && \text{by Lemma 6.1.7 (6)} \\
&= \omega|_{c^*(q)}|_p && \text{by Theorem 6.3.1} \\
&= \omega|_{p\,\&\,c^*(q)} && \text{by Lemma 6.1.7 (3).}
\end{aligned}$$

❑

Remark 7.2.9 In the preface to this book — to be precise, on page ix — we distinguished two approaches to crossover effects in updating. They were labeled there as 'weaken-update-marginalise' and 'extract-infer'. We can now make this more precise. In both cases we start from a joint distribution $\omega \in \mathcal{D}(X\times Y)$ with evidence in the form of a factor $q\colon Y \to \mathbb{R}_{\geq 0}$ on the second component.

- The 'weaken-update-marginalise' approach first weakens q to a factor $\mathbf{1}\otimes q = (\pi_2)^*(q)$ on the product space $X\times Y$. The types now match, so that $\mathbf{1}\otimes q$ can be used to update the joint distribution $\omega \in \mathcal{D}(X\times Y)$, giving $\omega|_{\mathbf{1}\otimes q} \in \mathcal{D}(X\times Y)$. Its first marginal $\omega|_{\mathbf{1}\otimes q}[1,0] \in \mathcal{D}(X)$ then gives the effect in X of updating in Y.
- The 'extract-infer' approach works differently: it first extracts a channel $c := \omega[0,1 \,\big|\, 1,0]\colon X \twoheadrightarrow Y$ from ω, via disintegration, such that $\omega = \langle id, c\rangle_*(\sigma)$, where $\sigma = \omega[1,0] \in \mathcal{D}(X)$ is ω's first marginal. One then obtains the effect in X of updating in Y via backward inference, as $\sigma|_{c^*(q)} \in \mathcal{D}(X)$.

From Theorem 6.3.1 we know that both approaches give the same outcome.

Exercises

7.2.1 Let $\sigma \in \mathcal{D}(X)$ have full support and consider $\omega := \sigma\otimes\tau$ for some $\tau \in \mathcal{D}(Y)$. Check that the channel $X \twoheadrightarrow Y$ extracted from ω by disintegration is the constant function $x \mapsto \tau$. Give a string diagrammatic account of this situation.

7.2.2 Disintegrate the distribution $Flrn(\tau) \in \mathcal{D}(\{H, L\} \times \{1, 2, 3\})$ in Subsection 1.6.1 to a channel $\{H, L\} \rightsquigarrow \{1, 2, 3\}$.

7.2.3 Consider sets $X = \{x_1, x_2\}$, $Y = \{y_1, y_2, y_3\}$, $Z = \{z_1, z_2, z_3\}$ with distribution $\omega \in \mathcal{D}(X \times Y)$ given by:

$$\tfrac{1}{4}|x_1, y_1\rangle + \tfrac{1}{4}|x_1, y_2\rangle + \tfrac{1}{6}|x_2, y_1\rangle + \tfrac{1}{6}|x_2, y_2\rangle + \tfrac{1}{6}|x_2, y_3\rangle$$

and $\rho \in \mathcal{D}(X \times Z)$ by:

$$\tfrac{1}{12}|x_1, z_1\rangle + \tfrac{1}{3}|x_1, z_2\rangle + \tfrac{1}{12}|x_1, z_3\rangle + \tfrac{1}{8}|x_2, z_1\rangle + \tfrac{1}{8}|x_2, z_2\rangle + \tfrac{1}{4}|x_2, z_3\rangle.$$

1 Check that ω and ρ have equal X-marginals.

2 Explicitly describe the extracted channels $c := \omega[0, 1 \,|\, 1, 0] \colon X \rightsquigarrow Y$ and $d := \omega[0, 1 \,|\, 1, 0] \colon X \rightsquigarrow Z$.

3 Show that the natural join $\omega \bowtie \rho \in \mathcal{D}(X \times Y \times Z)$ according to (7.15) is:

$$\begin{aligned}
&\tfrac{1}{24}|x_1, y_1, z_1\rangle + \tfrac{1}{6}|x_1, y_1, z_2\rangle + \tfrac{1}{24}|x_1, y_1, z_3\rangle + \tfrac{1}{24}|x_1, y_2, z_1\rangle + \tfrac{1}{6}|x_1, y_2, z_2\rangle \\
&\quad + \tfrac{1}{24}|x_1, y_2, z_3\rangle + \tfrac{1}{24}|x_2, y_1, z_1\rangle + \tfrac{1}{24}|x_2, y_1, z_2\rangle + \tfrac{1}{12}|x_2, y_1, z_3\rangle + \tfrac{1}{24}|x_2, y_2, z_1\rangle \\
&\quad + \tfrac{1}{24}|x_2, y_2, z_2\rangle + \tfrac{1}{12}|x_2, y_2, z_3\rangle + \tfrac{1}{24}|x_2, y_3, z_1\rangle + \tfrac{1}{24}|x_2, y_3, z_2\rangle + \tfrac{1}{12}|x_2, y_3, z_3\rangle.
\end{aligned}$$

7.2.4 Consider a joint distribution $\omega \in \mathcal{D}(X \times Y)$ and write its two marginals as $\omega_1 = \omega[1, 0] \in \mathcal{D}(X)$ and $\omega_2 = \omega[0, 1] \in \mathcal{D}(Y)$. Also write $c = \omega[0, 1 \,|\, 1, 0] \colon X \rightsquigarrow Y$ and $d = \omega[1, 0 \,|\, 0, 1] \colon Y \rightsquigarrow X$ for the two extracted channels. Prove the equivalences:

$$\begin{aligned}
&\omega \text{ is non-entwined, that is, } \omega = \omega_1 \otimes \omega_2 \\
&\iff c \text{ is constant, } \textit{i.e.}, c(x) = \tau \text{ for all } x \in X, \text{ for some } \tau \in \mathcal{D}(Y) \\
&\iff d \text{ is constant, } \textit{i.e.}, d(y) = \sigma \text{ for all } y \in Y, \text{ for some } \sigma \in \mathcal{D}(X).
\end{aligned}$$

The distributions σ and τ, if they exist, are the first and second marginal of ω. The equivalences correspond to the traditional way of expressing indepence in X and Y as $P(y \mid x) = P(y)$ or as $P(x \mid y) = P(x)$.

7.2.5 Combining the two items of Theorem 7.2.7 gives a bijective correspondence between:

$$\frac{\Omega \in \mathcal{D}(\mathcal{D}(X)) \text{ with } \big|supp(\Omega)\big| = N}{\omega \in \mathcal{D}(X) \text{ and } c \colon X \rightsquigarrow \boldsymbol{N} \text{ such that } c_*(\omega) \text{ has full support}}$$

Define this correspondence in detail and check that it is bijective.

7.2.6 Prove the following possibilitistic analogues of Theorem 7.2.7.

1 There is a bijective correspondence between:

$$\frac{R \in \mathcal{P}(X \times Y)}{U \in \mathcal{P}(X) \text{ with } f \colon U \to \mathcal{P}_*(Y)}$$

where $\mathcal{P}_*$ is used for the the subset of *non-empty* subsets.

2 For each number $N \geq 1$ there is a bijective correspondence:

$$\frac{A \in \mathcal{P}(\mathcal{P}(X)) \text{ with } |A| = N}{R \in \mathcal{P}(X \times N) \text{ with } \forall i \neq j.\, \exists x.\, \neg(R(x, i) \Leftrightarrow R(x, j))}$$

7.2.7 Prove that the equation

$$\langle id, c\rangle_*(\sigma)\big|_{p \otimes \mathbf{1}}[0, 1 \,|\, 1, 0] \;=\; c.$$

can be obtained both from Theorem 6.3.4 and from Theorem 7.2.8.

7.2.8 Prove that for a joint distribution $\omega \in \mathcal{D}(X \times Y)$ with full support one has:

$$\omega[1,0 \,|\, 0,1] \;=\; \big(\omega[0,1 \,|\, 1,0]\big)^{\dagger} \;:\; Y \rightsquigarrow X.$$

7.2.9 Let $\omega \in \mathcal{D}(X \times Y)$ be a joint state, whose two marginals $\omega_1 := \omega[1,0] \in \mathcal{D}(X)$ and $\omega_2 := \omega[0,1] \in \mathcal{D}(Y)$ have full support. Let $p \in \mathit{Pred}(X)$ and $q \in \mathit{Pred}(Y)$ be arbitrary predicates.

Prove that there are predicates q_1 on X and p_2 on Y such that:

$$\omega_1 \models p \,\&\, q_1 \;=\; \omega \models p \otimes q \;=\; \omega_2 \models p_1 \,\&\, q.$$

This is a discrete version of (Panangaden, 1998, Prop. 6.7).

Hint: Disintegrate!

7.2.10 Show that extraction for multisets (from Subsection 1.6.3) and disintegration commute with frequentist learning: ignoring non-emptyness requirements, the following diagram commutes.

$$\begin{tikzcd}
\mathcal{M}(X \times Y) \arrow[r, "extr_1"] \arrow[d, "Flrn"'] & \mathcal{M}(Y)^X \arrow[d, "Flrn^X"] \\
\mathcal{D}(X \times Y) \arrow[r, "dis_1"'] & \mathcal{D}(Y)^X
\end{tikzcd}$$

7.2.11 Let X, Y, Z be sets, where Z is finite. We use the *ad hoc* notation $\mathcal{D}_{fs2}(X \times Z) \hookrightarrow \mathcal{D}(X \times Z)$ for the subset of joint distributions whose second marginal, in $\mathcal{D}(Z)$, has full support. In this way, the disintegration / extraction function $e_X \colon \mathcal{D}_{fs2}(X \times Z) \to \mathcal{D}(X)^Z$, given by $e_X(\tau) = \tau[1,0 \,|\, 0,1] \colon Z \to \mathcal{D}(X)$, is well-defined.

Show that this extraction e_X is natural in X, with respect to channels. This means that for a channel $c \colon X \rightsquigarrow Y$ the following diagram commutes.

$$\begin{tikzcd}
\mathcal{D}_{fs2}(X \times Z) \arrow[r, "e_X"] \arrow[d, "(c \otimes id)_*"'] & \mathcal{D}(X)^Z \arrow[d, "c_*)^Z"] \\
\mathcal{D}_{fs2}(Y \times Z) \arrow[r, "e_Y"] & \mathcal{D}(Y)^Z
\end{tikzcd}$$

7.3 Disintegration, in general form

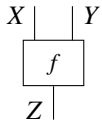
In the previous section we have introduced disintegration for joint distributions $\omega \in \mathcal{D}(X \times Y)$. We are now generalising it first to channels $Z \rightsquigarrow X \times Y$, and then more specifically to channels $X \rightsquigarrow Y_1 \times \cdots \times Y_n$. In the latter situation we shall disintegrate for a subset of the outgoing wires, as indicated by the Y_i. This gives considerable flexibility. The graphical language of string diagrams is useful to see what happens.

In essence, the more general form of disintegration in this section involves additional parameters. Traditionally, the situation at hand involves turing a conditional probability $P(X, Y \mid Z)$ into $P(Y \mid X, Z)$ such that

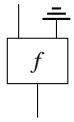
$$P(X, Y \mid Z) \;=\; P(Y \mid X, Z) \cdot P(X \mid Z).$$

This is captured diagrammatically in (7.18) below.

Definition 7.3.1 Consider a channel / box $f\colon Z \rightsquigarrow X \times Y$

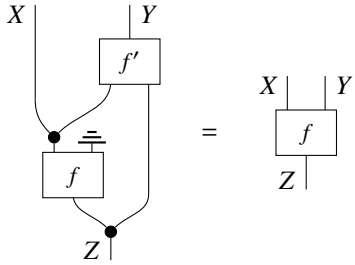


for which $f[1,0] =$



has full support.

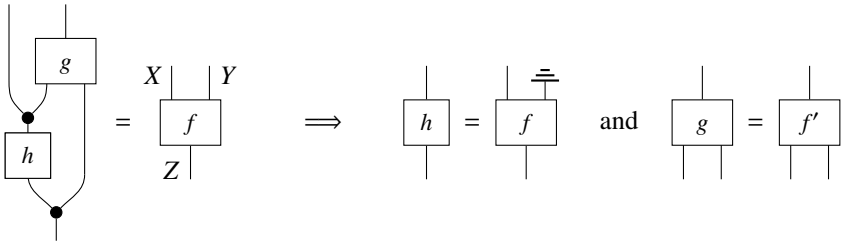
We say that f admits *disintegration* if there is a unique 'disintegration' channel $f'\colon X \times Z \rightsquigarrow Y$ satisfying the equation:



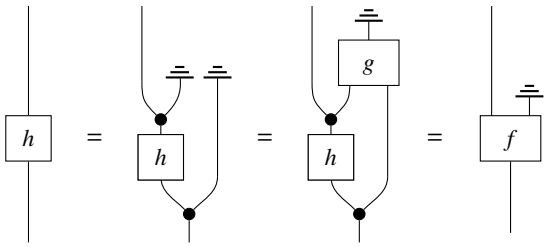
(7.18)

Graphically, one can think of the disintegration f' as being obtained via bending the X-wire downward, as described below in (7.20).

Uniqueness of the channel f' in this definition means: for each channel $g\colon X \times Z \rightsquigarrow Y$ and for each channel h from Z to B with full support one has:



The first equation $h = f[1,0]$ in the conclusion is obtained simply by applying discard $\overline{\top}$ to two right-most outgoing wires in the assumption — on the left and on the right of the equation — and using that g is unital:



The second equation $g = f'$ in the conclusion expresses uniqueness of the disintegration box f'.

We shall use the notation $f[0,1 \,|\, 1,0]$ for this disintegration box f'. This notation will be explained in greater generality below.

The disintegration channel $f'\colon X \times Z \rightsquigarrow Y$ is a pointwise version of the formulation for joint distributions on $X \times Y$ from (7.12):

$$f'(x,z) \coloneqq (f(z)|_{\mathbf{1}_X \otimes \mathbf{1}})[0,1] = \sum_{y \in Y} \frac{f(z)(x,y)}{f[1,0](z)(x)} \,|y\rangle. \tag{7.19}$$

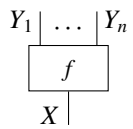
The full-support assumption for the marginalised channel $f[1,0]$ says that the marginal distribution $f[1,0](z) = f(z)[1,0] \in \mathcal{D}(X)$ has full support, for each $z \in Z$. More explicitly, this means that for each $x \in X$, the sum $\sum_{y \in Y} f(z)(x,y)$ is non-zero. Implicitly, this means that the set X is finite.

Without the full support requirement, disintegrations may still exist, but they are not unique, see Example 7.2.4 (2). We first describe an illustration involving multinomial and hypergeometric channels from Chapter 3.

Example 7.3.2 In Diagram (3.23) in Theorem 3.3.9 one can recognise how the sum of numbers function $sum \colon \mathbb{N}^m \to \mathbb{N}$ is disintegrated, with m-many parallel poisson channels $pois[-] \otimes \cdots \otimes pois[-]$ as prior. In this diagram one can ignore the *Freq* boxes / channels, since they are isomorphisms. The relevant disintegration channel is $mn[-](-) \circ (id \otimes Flrn)$. It has a 'dependent type' $\prod_{K \colon \mathbb{N}} \mathcal{D}(\boldsymbol{m}) \twoheadrightarrow \mathcal{N}[K](\boldsymbol{m})$.

Next we elaborate on the notation that we will use for disintegration.

Remark 7.3.3 In Subsection 2.3.2 we have been introduced the mask syntax for marginalisation, where a mask is a list containing as only elements 1 or 0, corresponding to the components (variables) that remain (1) and those that are marginalised out and disappear (0). We like to use a similar mask-style notation for disintegration, mimicking the traditional notation $\omega(x_1, x_4 | x_2, x_3)$. This requires *two* masks, separated by the sign '|' for conditional probability. We sketch how it works for a box



from X to $Y_1, \ldots, Y_n$. We introduce the following notation for a disintegration channel, with two masks M and N.

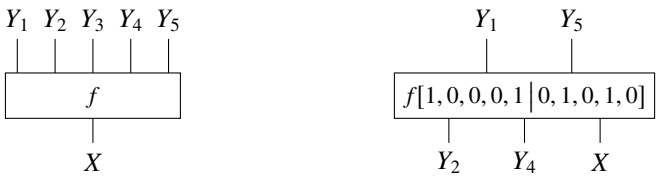
$$f[N \mid M]$$

The idea is to change the wires Y_i with a 1 at position i in M from output into input wires in the box $f[N \mid M]$. This notation $f[N \mid M]$ will be used in the following manner:

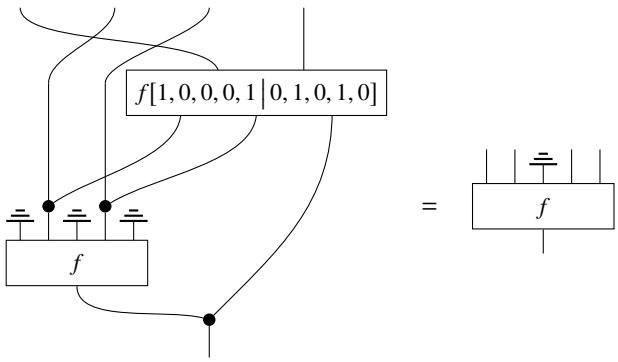
1 masks M, N must both be of length n;
2 M, N must be *disjoint*: there is no position i with a 1 both in M and in N;
3 the marginal $f[M]$ must have full support;
4 the domain of the disintegration box $f[N|M]$ is $\vec{Y} \cap M, X$, where $\vec{Y} \cap M$ contain precisely those Y_i with a 1 in M at position i;
5 the codomain of $f[N \mid M]$ is $\vec{Y} \cap N$.
6 $f[N|M]$ is unique in satisfying an "obvious" adaptation of Equation (7.18), in which $f[M \cup N]$ is equal to a string diagram consisting of $f[M]$ suitably followed by $f[N \mid M]$.

How this really works is best illustrated via a concrete example. Consider the box f on the left below, with five output wires. We elaborate the disintegration $f[1,0,0,0,1 \,\big|\, 0,1,0,1,0]$, as on

the right.



The disintegration box $f[1,0,0,0,1 \,|\, 0,1,0,1,0]$ is unique in satisfying:



In traditional notation one could express this equation as:

$$f(Y_1, Y_5 \mid Y_2, Y_4, X) \cdot f(Y_2, Y_4 \mid X) \;=\; f(Y_1, Y_2, Y_4, Y_5 \mid X).$$

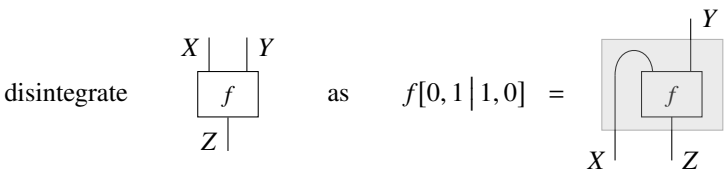
The wiring then remains implicit.

Two points are still worth noticing.

- It is not required that at position i there is a 1 either in mask M or in mask N when we form $f[N \mid M]$. If there is a 0 at i both in M and in N, then the wire at i is discarded altogether. This happens in the above illustration for the third wire: the string diagram on the right-hand side of the equation is $f[M \cup N] = f[1,1,0,1,1]$.
- The above disintegration $f[1,0,0,0,1 \,|\, 0,1,0,1,0]$ can be obtained from the 'simple', one-wire version of disintegration in Definition 7.3.1 by first suitably rearranging wires and combining them via products. How to do this precisely is left as an exercise (see below).

There is one more notational addition. This one is a non-standard construction in string diagrams.

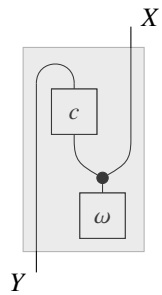
Remark 7.3.4 Disintegration is a not a 'compositional' operation that can be obtained by combining other string diagrammatic primitives. The reason is that disintegration involves normalisation, via division in (7.19). Still it would be nice to be able to use disintegration in diagrammatic form. For this purpose one may use a trick: in the setting of Definition 7.3.1 we 'bend' the relevant wire downwards and put a gray 'normalisation' box around the result in order to indicate that its interior is closed off and has become inaccessible. Thus:



(7.20)

This 'shaded box' notation can also be used for more complicated forms of disintegration, as described above. Using such shading we can describe the dagger of a channel $c\colon X \twoheadrightarrow Y$ as:

$$c^{\dagger}_{\omega} \quad = \tag{7.21}$$



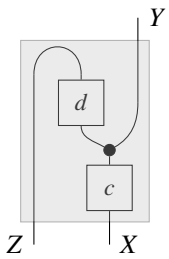
This shaded notation is useful for expressing basic properties of disintegration[2]. In the remainder of this section we will start using these shaded boxes more systematically. They will be elaborated further in Section 7.10.

In the beginning of this chapter we have introduced daggers of the form $d^{\dagger}_{\omega}$, where ω is the prior distribution on the domain of the channel d. In practice, this ω may be of the form $c(x)$, where c is a channel. Thus we like to generalise to daggers of the form $d^{\dagger}_{c}$, with the channel outputs $c(x)$ as priors, giving an additional parameter. This construction turns out to be useful in disintegration of Bayesian networks.

Definition 7.3.5 Suppose we have two composable channels $X \overset{c}{\twoheadrightarrow} Y \overset{d}{\twoheadrightarrow} Z$, such that $(d \circ c)(x) = d_*(c(x))$ has full support for each $x \in X$. Implicitly, this means that the set Z is finite.

We shall write $d^{\dagger}_{c}\colon Z \times X \twoheadrightarrow Y$ for the channel obtained as:

$$d^{\dagger}_{c} \quad := \tag{7.22}$$



We have the following characterisation.

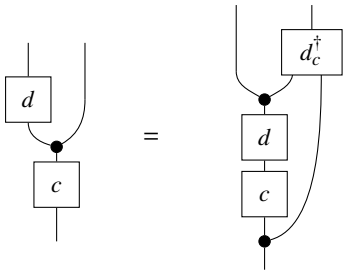
Lemma 7.3.6 *In the situation of Definition 7.3.5,*

1 The dagger $d^{\dagger}_{c}\colon Z \times X \twoheadrightarrow Y$ can be described concretely as:

$$d^{\dagger}_{c}(z, x) = \sum_{y \in Y} \frac{c(x)(y) \cdot d(y)(z)}{(d \circ c)(x)(z)} \left|y\right\rangle = c(x)\big|_{d^*(\mathbf{1}_z)} \overset{(7.1)}{=} d^{\dagger}_{c(x)}(z). \tag{7.23}$$

[2] It was introduced in the 2021 online version of this book, as graphical description for disintegration, but it developed into a calculus, see Lorenz and Tull (2023); Jacobs et al. (2025). Some authors write dashed instead of shaded boxes.

2 *It satisfies the following equation between channels* $X \rightsquigarrow Y \times Z$.



(7.24)

This is a special case of Equation (7.18)*, where the channel* f *is given by the string diagram on the left in* (7.24)*.*

Proof. 1 We check the middle equation of (7.23):

$$\begin{aligned} c(x)\big|_{d^*(\mathbf{1}_z)} &= \sum_{y \in Y} \frac{c(x)(y) \cdot d^*(\mathbf{1}_z)(y)}{\sum_y c(x)(y) \cdot d^*(\mathbf{1}_z)(y)} \big|y\rangle \\ &= \sum_{y \in Y} \frac{c(x)(y) \cdot d(y)(z)}{\sum_y c(x)(y) \cdot d(y)(z)} \big|y\rangle \;=\; \sum_{y \in Y} \frac{c(x)(y) \cdot d(y)(z)}{(d \circ c)(x)(z)} \big|y\rangle. \end{aligned}$$

2 Using this characterisation:

$$\begin{aligned} \Big((id \otimes d_c^\dagger) \circ \langle \Delta \circ d \circ c, id \rangle\Big)(x)(z, y) \;&=\; d_c^\dagger(z, x)(y) \cdot (d \circ c)(x)(z) \;=\; c(x)(y) \cdot d(y)(z) \\ &=\; \Big(\langle d, id \rangle \circ c\Big)(x)(z, y). \qquad \square \end{aligned}$$

We elaborate an example of a parameterised dagger in a situation is relevant in statistical mechanics, see Jacobs (2025b).

Example 7.3.7 This illustration builds on the multiset coefficient distribution $mc[\boldsymbol{N}, K] \in \mathcal{D}(\mathcal{N}[K](\boldsymbol{N}))$ from (2.5), for numbers $N, K \in \mathbb{N}_{>0}$ and on the nomial distribution $nom[N, K] \in \mathcal{D}(\{0, \ldots, (N-1) \cdot K\})$ from (2.3). From Exercise 2.1.4 we know that the latter can be described as image distribution of the first: $nom[N, K] = \mathcal{D}(sum)(mc[\boldsymbol{N}, K])$, via the sum of multisets function $sum \colon \mathcal{N}[K](\boldsymbol{N}) \to \{0, \ldots, (N-1) \cdot K\}$, with $sum(\varphi) = \sum_i \varphi(i) \cdot i$.

In this illustration we take the dagger of this *sum* function, with multiset coefficient distribution as prior. It yields what is called the 'Boltzmann on multisets' distribution in Jacobs (2025b), defined as: $BoM[N, K] := sum^\dagger_{mc[N,K]}$. Explicitly, via Lemma 7.1.4 (2), on a number $0 \leq i \leq (N-1) \cdot K$,

$$\begin{aligned} BoM[N, K](i) \;=\; sum^\dagger_{mc[N,K]}(i) \;&=\; \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{N}),\, sum(\varphi) = i} \frac{mc[\boldsymbol{N}, K](\varphi)}{\mathcal{D}(sum)(mc[\boldsymbol{N}, K])(i)} \big|\varphi\rangle \\ &\overset{(2.5)}{=} \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{N}),\, sum(\varphi) = i} \frac{\binom{\boldsymbol{N}}{\varphi}}{N^K \cdot nom[N, K](i)} \big|\varphi\rangle \\ &\overset{(2.3)}{=} \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{N}),\, sum(\varphi) = i} \frac{\binom{\boldsymbol{N}}{\varphi}}{N^K} \cdot \frac{N^K}{C_N(K, i)} \big|\varphi\rangle \\ &=\; \sum_{\varphi \in \mathcal{N}[K](\boldsymbol{N}),\, sum(\varphi) = i} \frac{\binom{\boldsymbol{N}}{\varphi}}{C_N(K, i)} \big|\varphi\rangle. \end{aligned} \tag{7.25}$$

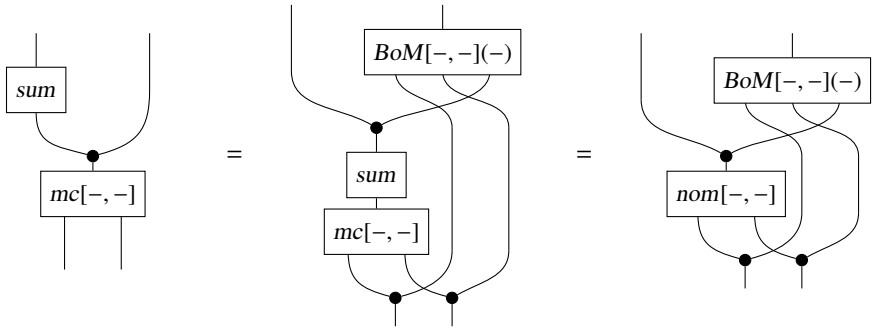
The latter formulation yields a proper distribution, with probabilities adding up to one, by the definition of the N-nomial (1.55) in terms of multisets.

For instance, for $N = 3$, $K = 9$ and $i = 6$ we get:

$$\begin{aligned} BoM[\mathbf{3}, 9](6) \;=\; & \tfrac{2}{37}\Big|\, 3|0\rangle + 6|1\rangle \Big\rangle + \tfrac{15}{37}\Big|\, 4|0\rangle + 4|1\rangle + 1|2\rangle \Big\rangle \\ & + \tfrac{18}{37}\Big|\, 5|0\rangle + 2|1\rangle + 2|2\rangle \Big\rangle + \tfrac{2}{37}\Big|\, 6|0\rangle + 3|2\rangle \Big\rangle . \end{aligned}$$

Notice that the multisets inside the big kets all have size 9 and sum 6. We get into physics territory: this number $i = 6$ may be seen as the total energy of $K = 9$ particles, distributed over three energy levels $0, 1, 2 \in \mathbf{3}$.

This dagger (7.25) involves parameter $N, K \in \mathbb{N}_{>0}$ and can be described in a situation as in (7.24).



The joint distributions described these three string diagrams are equal. The equation on the right follows from Exercise 2.1.4, and for the equation on the left we elaborate:

$$\begin{aligned} \langle sum, id\rangle_*(mc[N, K]) &= \sum_{\varphi \in \mathcal{N}[K](N)} \frac{(\varphi)}{N^K} \,\big|\, sum(\varphi), \varphi\rangle \\ &= \sum_{0 \le i \le (N-1)\cdot K} \;\sum_{\varphi \in \mathcal{N}[K](N),\, sum(\varphi) = i} \frac{(\varphi)}{C_N(K, i)} \cdot \frac{C_N(K, i)}{N^K} \,\big|\, i, \varphi\rangle \\ &= \sum_{0 \le i \le (N-1)\cdot K,\, \varphi \in \mathcal{N}[K](N)} BoM[N, K](i)(\varphi) \cdot nom[N, K](i) \,\big|\, i, \varphi\rangle \\ &= \langle id, BoM[N, K]\rangle_*(nom[N, K]). \end{aligned}$$

Exercises

7.3.1 Consider an extracted channel $\omega[1, 0, 0 \,|\, 0, 0, 1]$ for some state ω.

1 Write down the defining equation for this channel, as string diagram.

2 Check that $\omega[1, 0, 0 \,|\, 0, 0, 1]$ is the same as $\omega[1, 0, 1][1, 0 \,|\, 0, 1]$.

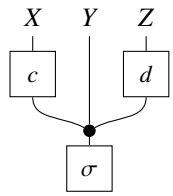
7.3.2 Check that a marginalisation ωM can also be described as disintegration $\omega[M \,|\, 0, \ldots, 0]$ where the number of 0's equals the length of the mask / list M.

7.3.3 Consider the channel / box f in Remark 7.3.3. Write down the equation for the disintegration $f[0, 1, 0, 1, 0 \,|\, 1, 0, 1, 0, 1]$. Formulate also what uniqueness means.

7.3.4 Show how to obtain the disintegration $f[0, 0, 0, 1, 1 \,|\, 0, 1, 1, 0, 0]$ in Remark 7.3.3 from the formulation in Definition 7.3.1 via rearranging and combining wires (via $\times$).

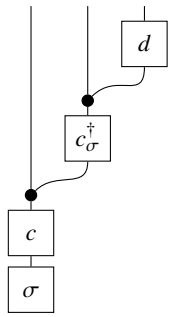
7.3.5 Consider the distribution $\tau \in \mathcal{D}(X \times Y \times Z)$ given by the following string diagram.

$\tau \quad =$

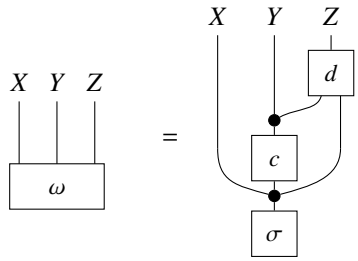


1 Extract $e := \tau[0,0,1 \,|\, 1,0,0] \colon X \rightarrow Z$ and show that $e = d \circ c_\sigma^\dagger$.

2 Show that τ is also be described by the following string diagram.



7.3.6 1 Show that each ternary distribution $\omega \in \mathcal{D}(X \times Y \times Z)$ can be factorised as below — assuming the relevant distributions have full support.

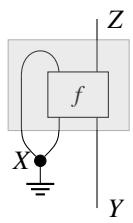


Give a similar factorisation for a quaternary distribution. How does this generalise to the n-ary case, for $n \geq 2$?

2 Check that the extracted channel $e := \omega[0,0,1 \,|\, 0,1,0] \colon Y \rightarrow Z$ can be described as: $e = d \circ \langle id, c_\sigma^\dagger \rangle$.

7.3.7 Let $f \colon X \times Y \rightarrow X \times Z$ be a channel, where X, Y, Z are finite sets. Define the 'probabilistic trace' $prtr(f) \colon Y \rightarrow Z$ as:

$prtr(f) \quad :=$



(7.26)

where $\perp\!\!\!\perp$ is used for the uniform distribution.

1 Check that:

$$prtr(f)(y) \;=\; \sum_{z\in Z}\left(\sum_{x\in X}\frac{1}{|X|}\cdot\frac{f(x,y)(x,z)}{\sum_{y'\in Y} f(x,y)(x,y')}\right)\big|z\rangle$$

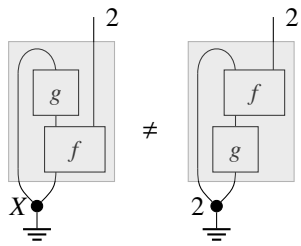
A categorically oriented reader might now ask if the above definition $prtr(f)$ yields a proper 'trace' construction in the symmetric monoidal Kleisli category $\mathbf{Chan}(\mathcal{D})$ of probabilistic channels, but this is *not* the case: the so-called dinaturality condition fails. This will be illustrated in the next few points.

2 Take space $X = \{a,b,c\}$ and $2 = \{0,1\}$ with channels $f\colon X \rightsquigarrow 2\times 2$ and $g\colon 2 \rightsquigarrow X$ defined by:

$$\begin{aligned} f(a) &= \tfrac{1}{4}|0,0\rangle + \tfrac{3}{4}|1,0\rangle \\ f(b) &= \tfrac{2}{5}|0,0\rangle + \tfrac{3}{5}|1,1\rangle \\ f(c) &= \tfrac{1}{2}|0,1\rangle + \tfrac{1}{2}|1,0\rangle \end{aligned} \qquad\qquad \begin{aligned} g(0) &= \tfrac{1}{3}|a\rangle + \tfrac{2}{3}|c\rangle \\ g(1) &= \tfrac{1}{2}|a\rangle + \tfrac{1}{2}|b\rangle. \end{aligned}$$

The aim is to show an inequality of states:

$$prtr((g\otimes id)\circ f) \;=$$



$$= prtr(f\circ g)$$

Both sides are an instance of (7.26) with $B = 1$.

3 Check that $(g\otimes id)\circ f\colon X \rightsquigarrow X\times 2$ is:

$$\begin{aligned} a &\mapsto \tfrac{11}{24}|a,0\rangle + \tfrac{3}{8}|b,0\rangle + \tfrac{1}{6}|c,0\rangle \\ b &\mapsto \tfrac{2}{15}|a,0\rangle + \tfrac{3}{10}|a,1\rangle + \tfrac{3}{10}|b,1\rangle + \tfrac{4}{15}|c,0\rangle \\ c &\mapsto \tfrac{1}{4}|a,0\rangle + \tfrac{1}{6}|a,1\rangle + \tfrac{1}{4}|b,0\rangle + \tfrac{1}{3}|c,1\rangle. \end{aligned}$$

And that $f\circ g\colon 2 \rightsquigarrow 2\times 2$ is:

$$\begin{aligned} 0 &\mapsto \tfrac{1}{12}|0,0\rangle + \tfrac{1}{3}|0,1\rangle + \tfrac{7}{12}|1,0\rangle \\ 1 &\mapsto \tfrac{13}{40}|0,0\rangle + \tfrac{3}{8}|1,0\rangle + \tfrac{3}{10}|1,1\rangle. \end{aligned}$$

4 Now show that the disintegration $((g\otimes id)\circ f)[0,1\,|\,1,0]\colon X\times X \rightsquigarrow 2$ is:

$$\begin{aligned} (a,a) &\mapsto 1|0\rangle & (b,a) &\mapsto \tfrac{4}{13}|0\rangle + \tfrac{9}{13}|1\rangle & (c,a) &\mapsto \tfrac{3}{5}|0\rangle + \tfrac{2}{5}|1\rangle \\ (a,b) &\mapsto 1|0\rangle & (b,b) &\mapsto 1|1\rangle & (c,b) &\mapsto 1|0\rangle \\ (a,c) &\mapsto 1|0\rangle & (b,c) &\mapsto 1|0\rangle & (c,c) &\mapsto 1|1\rangle. \end{aligned}$$

And that $(f\circ g)[0,1\,|\,1,0]\colon 2\times 2 \rightsquigarrow 2$ is:

$$\begin{aligned} (0,0) &\mapsto \tfrac{1}{5}|0\rangle + \tfrac{4}{5}|1\rangle & (1,0) &\mapsto 1|0\rangle \\ (0,1) &\mapsto 1|0\rangle & (1,1) &\mapsto \tfrac{5}{9}|0\rangle + \tfrac{4}{9}|1\rangle. \end{aligned}$$

5 Conclude that:

$$prtr((g\otimes id)\circ f) \;=\; \tfrac{1}{3}|0\rangle + \tfrac{2}{3}|1\rangle,$$

whereas:

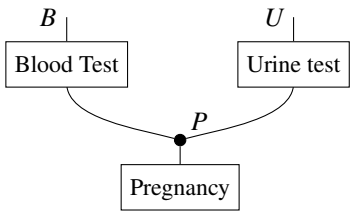
$$prtr(f \circ g) \;=\; \tfrac{17}{45}|0\rangle + \tfrac{28}{45}|1\rangle.$$

7.4 Disintegration for learning with missing data

This section consideres learning from tables in which some entries are missing and illustrates how disintegration can be used to extract relevant channels. We do not treat this topic exhaustively, but instead we elaborate two examples from the literature in order to give an impression of the issues involved. The examples differ with respect to the presence (or not) of prior information. The learning methodology is the same in both cases.

7.4.1 Learning with missing data, without prior

We look at an example from (Jensen and Nielsen, 2007, §6.2.1). It involves pregnancy of cows, which can be deduced from a urine test and a blood test. A simple Bayesian network structure is assumed, which we write as string diagram:



with sets
$$\begin{cases} P \;=\; \{p, p^\perp\} \\ B \;=\; \{b, b^\perp\} \\ U \;=\; \{u, u^\perp\}. \end{cases} \tag{7.27}$$

The elements p and $p^\perp$ represent 'pregnancy' and 'no pregnancy', respectively. Similarly, $b, b^\perp$ and $u, u^\perp$ represent a positive and negative blood / urine test.

The data in Jensen and Nielsen (2007) involves 5 cases, as given in the following table, where a question mark is used for a missing item.

case	**Pregnancy**	**Blood test**	**Urine test**
1	?	b	u
2	p	$b^\perp$	u
3	p	b	?
4	p	b	$u^\perp$
5	?	$b^\perp$	?

(7.28)

We wish to learn a distribution on P and two channels $P \twoheadrightarrow B$ and $P \twoheadrightarrow U$ as interpretation of the Bayesian network (7.27). The state and two channels can be obtained from a joint distribution on $P \times B \times U$ by marginalisation and disintegration (channel extraction). Our aim is thus to first learn such a joint distribution from the table.

Earlier, with the Medicine-Blood Table (1.31), without missing data, we could directly translate the table into a multiset and then turn it into a joint distribution via frequentist learning. That approach does not work when some of the data points are missing.

Here, instead, we shall translate each of the rows i in Table (7.28) into a predicate q_i on $P \times B \times U$, using a point predicate when all data are there, like in the second and fourth row, so:

$$q_2 \;\coloneqq\; \mathbf{1}_{(p,b^\perp,u)} \;=\; \mathbf{1}_p \otimes \mathbf{1}_{b^\perp} \otimes \mathbf{1}_u \qquad\qquad q_4 \;\coloneqq\; \mathbf{1}_{(p,b,u^\perp)} \;=\; \mathbf{1}_p \otimes \mathbf{1}_b \otimes \mathbf{1}_{u^\perp}.$$

For cases 1,3,5 we cannot use a point predicate where there is a missing item ?. There we use the truth predicate $\mathbf{1}$, in the parallel conjunction formulation $\otimes$. This indicates that all options are equally likely. Thus:

$$q_1 \coloneqq \mathbf{1} \otimes \mathbf{1}_b \otimes \mathbf{1}_u \qquad q_3 \coloneqq \mathbf{1}_p \otimes \mathbf{1}_b \otimes \mathbf{1} \qquad q_5 \coloneqq \mathbf{1} \otimes \mathbf{1}_{b^\perp} \otimes \mathbf{1}.$$

We can now update the uniform state *unif* on $P \times B \times U$ with these predicates q_i. We take the convex sum over these five Bayesian updates, corresponding to five rows in Table 7.28.

$$\begin{aligned}
\tau &\coloneqq \tfrac{1}{5} \cdot \mathit{unif}|_{q_1} + \tfrac{1}{5} \cdot \mathit{unif}|_{q_2} + \tfrac{1}{5} \cdot \mathit{unif}|_{q_3} + \tfrac{1}{5} \cdot \mathit{unif}|_{q_4} + \tfrac{1}{5} \cdot \mathit{unif}|_{q_5} \\
&= \tfrac{1}{10}|p,b,u\rangle + \tfrac{1}{10}|p^\perp,b,u\rangle + \tfrac{1}{5}|p,b^\perp,u\rangle + \tfrac{1}{10}|p,b,u\rangle + \tfrac{1}{10}|p,b,u^\perp\rangle \\
&\qquad + \tfrac{1}{5}|p,b,u^\perp\rangle + \tfrac{1}{20}|p,b^\perp,u\rangle + \tfrac{1}{20}|p^\perp,b^\perp,u\rangle + \tfrac{1}{20}|p,b^\perp,u^\perp\rangle + \tfrac{1}{20}|p^\perp,b^\perp,u^\perp\rangle \\
&= \tfrac{1}{5}|p,b,u\rangle + \tfrac{3}{10}|p,b,u^\perp\rangle + \tfrac{1}{4}|p,b^\perp,u\rangle + \tfrac{1}{20}|p,b^\perp,u^\perp\rangle \\
&\qquad + \tfrac{1}{10}|p^\perp,b,u\rangle + \tfrac{1}{20}|p^\perp,b^\perp,u\rangle + \tfrac{1}{20}|p^\perp,b^\perp,u^\perp\rangle.
\end{aligned}$$

Later on we shall see that this is an instance of Jeffrey's updating with evidence, see Definition 8.3.5 (1).

The first marginal $\tau[1,0,0] = \frac{4}{5}|p\rangle + \frac{1}{5}|p^\perp\rangle \in \mathcal{D}(P)$ is the learned pregnancy distribution. The two 'Blood test' and 'Urine Test' channels $P \rightsquigarrow B$ and $P \rightsquigarrow U$ in (7.27) can be extracted via disintegration.

- We get $c \coloneqq \tau[0,1,0 \,|\, 1,0,0] = \tau[1,1,0][0,1 \,|\, 1,0] \colon P \rightsquigarrow B$ with:

$$\begin{aligned}
c(p) &= \frac{\tau(p,b,u) + \tau(p,b,u^\perp)}{\tau(p,b,u) + \tau(p,b,u^\perp) + \tau(p,b^\perp,u) + \tau(p,b^\perp,u^\perp)}|b\rangle \\
&\qquad + \frac{\tau(p,b^\perp,u) + \tau(p,b^\perp,u^\perp)}{\tau(p,b,u) + \tau(p,b,u^\perp) + \tau(p,b^\perp,u) + \tau(p,b^\perp,u^\perp)}|b^\perp\rangle \\
&= \frac{1/5 + 3/10}{1/5 + 3/10 + 1/4 + 1/20}|b\rangle + \frac{1/4 + 1/20}{1/5 + 3/10 + 1/4 + 1/20}|b^\perp\rangle = \tfrac{5}{8}|b\rangle + \tfrac{3}{8}|b^\perp\rangle. \\
c(p^\perp) &= \cdots = \tfrac{1}{2}|b\rangle + \tfrac{1}{2}|b^\perp\rangle.
\end{aligned}$$

  We see that a pregnant cow has a slightly higher probability than a non-pregnant cow of getting a positive blood test.
- In the same way we get $d \coloneqq \tau[0,0,1 \,|\, 1,0,0] = \tau[1,0,1][0,1 \,|\, 1,0] \colon P \rightsquigarrow U$ with:

$$d(p) = \tfrac{9}{16}|u\rangle + \tfrac{7}{16}|u^\perp\rangle \qquad \text{and} \qquad d(p^\perp) = \tfrac{3}{4}|u\rangle + \tfrac{1}{4}|u^\perp\rangle.$$

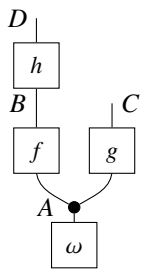
  Apparantly, the urine test is more likely to be positive for a non-pregnant cow.

These channels c and d coincide with the conditional probabilities computed in Jensen and Nielsen (2007), without explicitly using disintegration.

7.4.2 Learning with missing data, with prior

In the previous example we used a uniform distribution for the missing data items — corresponding to the fact that we assumed no prior knowledge. Here we look into an example, taken from (Darwiche, 2009, §17.3.1), where we do have prior information, namely in the form of a

Bayesian network, involving four sets $A = \{a, a^\perp\}$, $B = \{b, b^\perp\}$, $C = \{c, c^\perp\}$ and $D = \{d, d^\perp\}$, in the string diagram:



with

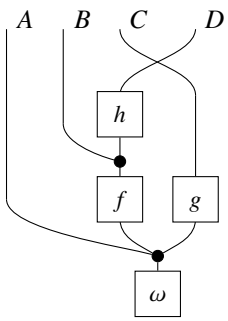
$$\begin{cases} \omega = \frac{1}{5}|a\rangle + \frac{4}{5}|a^\perp\rangle \\ f(a) = \frac{3}{4}|b\rangle + \frac{1}{4}|b^\perp\rangle \\ f(a^\perp) = \frac{1}{10}|b\rangle + \frac{9}{10}|b^\perp\rangle \\ h(b) = \frac{1}{5}|d\rangle + \frac{4}{5}|d^\perp\rangle \\ h(b^\perp) = \frac{7}{10}|d\rangle + \frac{3}{10}|d^\perp\rangle \\ g(a) = \frac{1}{2}|c\rangle + \frac{1}{2}|c^\perp\rangle \\ g(a^\perp) = \frac{1}{4}|c\rangle + \frac{3}{4}|c^\perp\rangle \end{cases} \tag{7.29}$$

We first turn this Bayesian network into a joint distribution $\tau \in \mathcal{D}(A \times B \times C \times D)$, as interpretation of the accessible string diagram on the left below — obtained from the string diagram in (7.29).

$$\begin{aligned} \tau \coloneqq \Big(& (id \otimes id \otimes swap) \circ (id \otimes id \otimes h \otimes id) \\ & \circ (id \otimes \Delta_2 \otimes id) \circ (id \otimes f \otimes g) \circ \Delta_3 \Big)_*(\omega) \end{aligned}$$

That is:

$\tau \;=$



$$\begin{aligned} = \; & \tfrac{3}{200}|a,b,c,d\rangle + \tfrac{3}{50}|a,b,c,d^\perp\rangle + \\ & \tfrac{3}{200}|a,b,c^\perp,d\rangle + \tfrac{3}{50}|a,b,c^\perp,d^\perp\rangle + \\ & \tfrac{7}{400}|a,b^\perp,c,d\rangle + \tfrac{3}{400}|a,b^\perp,c,d^\perp\rangle + \\ & \tfrac{7}{400}|a,b^\perp,c^\perp,d\rangle + \tfrac{3}{400}|a,b^\perp,c^\perp,d^\perp\rangle + \\ & \tfrac{1}{250}|a^\perp,b,c,d\rangle + \tfrac{2}{125}|a^\perp,b,c,d^\perp\rangle + \\ & \tfrac{3}{250}|a^\perp,b,c^\perp,d\rangle + \tfrac{6}{125}|a^\perp,b,c^\perp,d^\perp\rangle + \\ & \tfrac{63}{500}|a^\perp,b^\perp,c,d\rangle + \tfrac{27}{500}|a^\perp,b^\perp,c,d^\perp\rangle + \\ & \tfrac{189}{500}|a^\perp,b^\perp,c^\perp,d\rangle + \tfrac{81}{500}|a^\perp,b^\perp,c^\perp,d^\perp\rangle. \end{aligned}$$

This distribution τ will be used as prior.

The table with data provided in Darwiche (2009) is:

case	A	B	C	D
1	?	b	$c^\perp$	?
2	?	b	?	$d^\perp$
3	?	$b^\perp$	c	d
4	?	$b^\perp$	c	d
5	?	b	?	$d^\perp$

(7.30)

At this stage we do not use uniform distributions as fill-in. Instead, we use the Bayesian network to infer probabilities for the missing data. For instance, in the first case we have point predicates $\mathbf{1}_b$ and $\mathbf{1}_{c^\perp}$. From the crossover update result Theorem 6.3.1 we know that this inference can also be done via the joint state τ, namely as:

$$\tau_1 \coloneqq \tau\big|_{\mathbf{1} \otimes \mathbf{1}_b \otimes \mathbf{1}_{c^\perp} \otimes \mathbf{1}}.$$

The missing probabilities for A, D are then obtained via the marginal, as:

$$\begin{aligned} \tau_1[1,0,0,1] &= \tfrac{1}{9}|a,d\rangle + \tfrac{4}{9}|a,d^\perp\rangle + \tfrac{4}{45}|a^\perp,d\rangle + \tfrac{16}{45}|a^\perp,d^\perp\rangle \\ &= 0.1111|a,d\rangle + 0.4444|a,d^\perp\rangle + 0.08889|a^\perp,d\rangle + 0.3556|a^\perp,d^\perp\rangle. \end{aligned}$$

This distribution appears in (Darwiche, 2009, Fig. 17.4 (a)).

In a similar way we define:

$$\tau_2 \coloneqq \tau\big|_{\mathbf{1}\otimes\mathbf{1}_b\otimes\mathbf{1}\otimes\mathbf{1}_{d^\perp}} \qquad \tau_3 \coloneqq \tau\big|_{\mathbf{1}\otimes\mathbf{1}_{b^\perp}\otimes\mathbf{1}_c\otimes\mathbf{1}_d} \qquad \tau_4 \coloneqq \tau_3 \qquad \tau_5 \coloneqq \tau_2.$$

These five states $\tau_1, \ldots, \tau_5$ are combined in a convex combination, like in the previous subsection, giving a new joint state:

$$\tau' \coloneqq \tfrac{1}{5}\cdot\tau_1 + \tfrac{1}{5}\cdot\tau_2 + \tfrac{1}{5}\cdot\tau_3 + \tfrac{1}{5}\cdot\tau_4 + \tfrac{1}{5}\cdot\tau_5 = \tfrac{1}{5}\cdot\tau_1 + \tfrac{2}{5}\cdot\tau_2 + \tfrac{2}{5}\cdot\tau_3.$$

From τ' we can obtain a newly learned interpretation of the Bayesian network (7.29), via marginalisation and disintegration. In detail, the new distribution ω' on A is:

$$\omega' \coloneqq \tau'[1,0,0,0] = \tfrac{3571}{8487}|a\rangle + \tfrac{4916}{8487}|a^\perp\rangle \approx 0.4208|a\rangle + 0.5792|a^\perp\rangle. \tag{7.31}$$

We get as newly learned channels:

$$f' \coloneqq \tau'[0,1,0,0\,|\,1,0,0,0] \qquad \begin{cases} f'(a) = \frac{3157}{3571}|b\rangle + \frac{414}{3571}|b^\perp\rangle \\ \qquad\quad \approx 0.8841|b\rangle + 0.1159|b^\perp\rangle \\ f'(a^\perp) = \frac{2419}{6145}|b\rangle + \frac{3726}{6145}|b^\perp\rangle \\ \qquad\quad \approx 0.3937|b\rangle + 0.6063|b^\perp\rangle \end{cases}$$

$$g' \coloneqq \tau'[0,0,1,0\,|\,1,0,0,0] \qquad \begin{cases} g'(a) = \frac{1521}{3571}|c\rangle + \frac{2050}{3571}|c^\perp\rangle \\ \qquad\quad \approx 0.4259|c\rangle + 0.5741|c^\perp\rangle \\ g'(a^\perp) = \frac{819}{1229}|c\rangle + \frac{410}{1229}|c^\perp\rangle \\ \qquad\quad \approx 0.6664|c\rangle + 0.3336|c^\perp\rangle \end{cases}$$

$$h' \coloneqq \tau'[0,0,0,1\,|\,0,1,0,0] \qquad \begin{cases} h'(b) = \frac{1}{15}|d\rangle + \frac{14}{15}|d^\perp\rangle \\ \qquad\quad \approx 0.06667|d\rangle + 0.9333|d^\perp\rangle \\ h'(b^\perp) = 1|d\rangle. \end{cases}$$

These outcomes are as in (Darwiche, 2009, Fig. 17.5) — up to some small differences, probably due to rounding.

Interestingly, Table (7.30) gives no information about A, but still we learn a new distribution ω' on A, basically from the prior information τ.

Exercises

7.4.1 Check that the pregnancy distribution in Subsection 7.4.1, obtained from Table (7.28), is $\frac{4}{5}|p\rangle + \frac{1}{5}|p^\perp\rangle$.

7.4.2 Calculate in detail that, in Subsection 7.4.2,

$$\tau_1[1,0,0,1] = \tfrac{1}{9}|a,d\rangle + \tfrac{4}{9}|a,d^\perp\rangle + \tfrac{4}{45}|a^\perp,d\rangle + \tfrac{16}{45}|a^\perp,d^\perp\rangle.$$

7.4.3 Check that the new state ω' in (7.31) can also be obtained as:

$$\omega' = \tfrac{1}{5}\cdot\omega|_{f^*(\mathbf{1}_b)\,\&\,g^*(\mathbf{1}_{c^\perp})} + \tfrac{2}{5}\cdot\omega|_{f^*(\mathbf{1}_b)\,\&\,h^*(\mathbf{1}_{d^\perp})} + \tfrac{2}{5}\cdot\omega|_{f^*(\mathbf{1}_{b^\perp}\,\&\,h^*(\mathbf{1}_c))\,\&\,g^*(\mathbf{1}_c)}.$$

Argue why this works.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Figure 7.1 Weather and play data, copied from (Witten et al., 2011, Table 1.2).

7.5 Disintegration and inversion in machine learning

This section illustrates the role of disintegration and Bayesian inversion in two fundamental techniques in machine learning, namely in naive Bayesian classification and in decision tree learning. These applications will be explained via examples from the literature.

7.5.1 Naive Bayesian classification

We illustrate the use of both disintegration and Bayesian inversion in an example of 'naive' Bayesian classification from Witten et al. (2011); we follow the analysis of Cho and Jacobs (2019). Instead of trying to explain what a naive Bayesian classification is or does, we demonstrate via this example how it works. In the end, in Remark 7.5.1 we give a more general description.

Consider the table in Figure 7.1. It collects data about certain weather conditions and whether or not there is playing (outside). The question asked in Witten et al. (2011) is: given this table, what can be said about the probability of playing if the outlook is *sunny*, the temperature is *cold*, the humidity is *high* and it is *windy*? This is a typical Bayesian update question, starting from (point) evidence. We will first analyse the situation in terms of channels.

We start by extracting the underlying spaces for the columns / categories in the table in Figure 7.1. We choose obvious abbreviations for the entries in the table:

$$O = \{s, o, r\} \qquad T = \{h, m, c\} \qquad H = \{h, n\}. \qquad W = \{t, f\} \qquad P = \{y, n\}$$

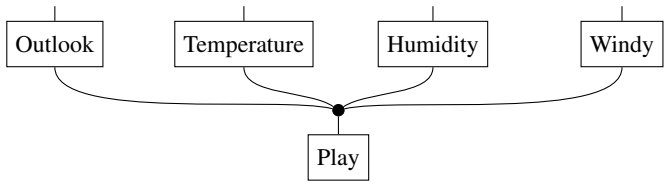
These sets are joined into a single product space:

$$S \coloneqq O \times T \times H \times W \times P.$$

It combines the five columns in Figure 7.1. The table itself can now be considered as a multiset in $\mathcal{M}(S)$ with 14 elements, each with multiplicity one. We will turn it immediately into an empirical distribution — formally via frequentist learning. It yields a distribution $\tau \in \mathcal{D}(S)$, with 14 entries, each with the same probability, written as:

$$\tau = \tfrac{1}{14}|s, h, h, f, n\rangle + \tfrac{1}{14}|s, h, h, t, n\rangle + \cdots + \tfrac{1}{14}|r, m, h, t, n\rangle.$$

We use a 'naive' Bayesian model in this situation, which means that we assume that all weather features are independent. This assumption can be visualised via the following string diagram:



(7.32)

This model with separate channels oversimplifies the situation, but still it often leads to good (enough) outcomes.

We take the above perspective on the distribution τ, that is, we 'factorise' τ according to this string diagram (7.32). Obviously, the play state $\pi \in \mathcal{D}(P)$ is obtained as the last marginal:

$$\pi \coloneqq \tau[0,0,0,0,1] = \tfrac{9}{14}|y\rangle + \tfrac{5}{14}|n\rangle.$$

Next, we extract four channels c_O, c_T, c_H, c_W via appropriate disintegrations, from the Play column to the Outlook / Temperature / Humidity / Windy columns.

$$\begin{aligned} c_O &\coloneqq \tau[1,0,0,0,0 \,|\, 0,0,0,0,1] : P \longrightarrow O \\ c_T &\coloneqq \tau[0,1,0,0,0 \,|\, 0,0,0,0,1] : P \longrightarrow T \\ c_H &\coloneqq \tau[0,0,1,0,0 \,|\, 0,0,0,0,1] : P \longrightarrow H \\ c_W &\coloneqq \tau[0,0,0,1,0 \,|\, 0,0,0,0,1] : P \longrightarrow W \end{aligned} \tag{7.33}$$

For instance the 'outlook' channel $c_O \colon P \to O$ looks as follows.

$$c_O(y) = \tfrac{2}{9}|s\rangle + \tfrac{4}{9}|o\rangle + \tfrac{3}{9}|r\rangle \qquad\qquad c_O(n) = \tfrac{3}{5}|s\rangle + \tfrac{2}{5}|r\rangle. \tag{7.34}$$

It is analysed in greater detail in Exercise 7.5.1 below.

Now we can form the tuple channel of these extracted channels, called c in:

$$P \xrightarrow{c \,\coloneqq\, \langle c_O, c_T, c_H, c_W \rangle} O \times T \times H \times W$$

Recall the question that we started from: what is the probability of playing if the outlook is *sunny*, the temperature is *Cold*, the humidity is *High* and it is Windy? These features can be translated into an element (s, c, h, w) of the codomain $O \times T \times H \times W$ of this tuple channel — and thus into a point predicate. Hence our answer can be obtained by Bayesian inversion of the tuple channel, as:

$$c_\pi^\dagger(s,c,h,t) \overset{(7.1)}{=} \pi\big|_{c^*(\mathbf{1}_{(s,c,h,t)})} = \tfrac{125}{611}|y\rangle + \tfrac{486}{611}|n\rangle \approx 0.2046|y\rangle + 0.7954|n\rangle.$$

This corresponds to the probability 20.5% calculated in Witten et al. (2011) — without any disintegration or Bayesian inversion.

The classification that we have just performed works via what it called a naive Bayesian *classifier*. In our set-up this classifier is the dagger channel:

$$O \times T \times H \times W \xrightarrow{c_\pi^\dagger} P$$

It predicts playing for a 4-tuple in $O \times T \times H \times W$.

In the end one can reconstruct a joint state with space S via the extracted channel, as graph:

$$\langle c_O, c_T, c_H, c_W, id \rangle_*(\pi).$$

This state differs considerably from the original table / state τ. It shows that the shape (7.32) does not really fit the data that we have in Figure 7.1. But recall that this approach is called naive. We shall soon look closer into such matters of shape in Section 7.8.

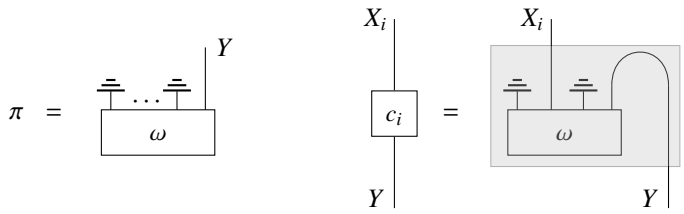
Remark 7.5.1 Now that we have seen the above illustration we can give a more abstract recipe of how to obtain a Bayesian classifier. The starting point is a joint distribution $\omega \in \mathcal{D}(X_1 \times \cdots \times X_n \times Y)$ where $X_1, \ldots, X_n$ are the sets describing the 'input features' and Y is the set of 'target features' that are used in classification: its elements represent the different classes. The recipe involves the following steps, in which disintegration and Bayesian inversion play a prominent role.

1 Compute the prior classification probability $\pi := \omega[0, \ldots, 0, 1] \in \mathcal{D}(Y)$.
2 Extract n-channels $c_i \colon Y \rightsquigarrow X_i$ via disintegration:

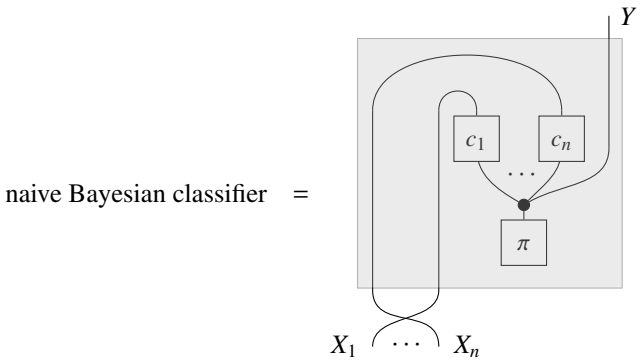
$$\begin{aligned} c_1 &:= \omega[1, 0, \ldots, 0 \,|\, 0, \ldots, 0, 1] \;:\; Y \rightsquigarrow X_1 \\ c_2 &:= \omega[0, 1, 0 \ldots, 0 \,|\, 0, \ldots, 0, 1] \;:\; Y \rightsquigarrow X_2 \\ &\;\;\vdots \\ c_n &:= \omega[0, \ldots, 0, 1, 0 \,|\, 0, \ldots, 0, 1] \;:\; Y \rightsquigarrow X_n. \end{aligned}$$

3 Form the tuple channel $c := \langle c_1, \ldots, c_n \rangle \colon Y \rightsquigarrow X_1 \times \cdots \times X_n$.
4 Take the Bayesian inversion (dagger channel) $c_\pi^\dagger \colon X_1 \times \cdots \times X_n \rightsquigarrow Y$, as classifier. It gives for each n-tuple of input features $x_1, \ldots, x_n$ a distribution $c_\pi^\dagger(x_1, \ldots, x_n) \in \mathcal{D}(Y)$ on the set Y of classes (or target features). The distribution gives the probability of the tuple belonging to class $y \in Y$.

Graphically, these steps can be described via the shaded box notation from Remark 7.3.4. First take:



The classification channel $X_1 \times \cdots \times X_n \rightsquigarrow Y$ is then obtained via the dagger construction:



The twisting at the bottom happens implicitly when we consider the product $X_1 \times \cdots \times X_n$ as a single set and take the dagger with respect to this set.

Author	Thread	Length	Where read	User action
known	new	long	home	skip
unknown	new	short	work	read
unknown	follow-up	long	work	skip
known	follow-up	long	home	skip
known	new	short	home	read
known	follow-up	long	work	skip
unknown	follow-up	short	work	skip
unknown	new	short	work	read
known	follow-up	long	home	skip
known	new	long	work	skip
unknown	follow-up	short	home	skip
known	new	long	work	skip
known	follow-up	short	home	read
known	new	short	work	read
known	new	short	home	read
known	follow-up	short	work	read
known	new	short	home	read
unknown	new	short	work	read

Figure 7.2 Data about circumstances for reading or skipping of articles, copied from Poole and Mackworth (2017).

7.5.2 Decision tree learning

We proceed as before, by first going trough an example from the literature, and then taking a step back to describe more abstractly what is going on. We start with a table of data in Figure 7.2, from (Poole and Mackworth, 2017, Fig. 7.1). It describes user actions (read or skip) for articles that are "posted to a threaded discussion website depending on whether the author is known or not, whether the article started a new thread or was a follow-up, the length of the article, and whether it is read at home or at work."

We formalise the table in Figure 7.2 via the following five sets, with elements corresponding in an obvious way to the entries in the table: k = known, u = unknown, *etc.*.

$$A = \{k, u\} \qquad T = \{n, f\} \qquad L = \{l, s\} \qquad W = \{h, w\} \qquad U = \{s, r\}.$$

We interprete the table as a joint distribution $\omega \in \mathcal{D}(A \times T \times L \times W \times U)$ with the same probability for each of the 18 entries in the table:

$$\omega = \tfrac{1}{18}\big|k, n, l, h, s\big\rangle + \tfrac{1}{18}\big|u, n, s, w, r\big\rangle + \cdots + \tfrac{1}{18}\big|u, n, s, w, r\big\rangle. \tag{7.35}$$

So far there is no real difference with the naive Bayesian classification example in the previous subsection. But the aim now is not *classification* but *deciding*: given a 4-tuple of input features $(x_1, x_2, x_3, x_4) \in A \times T \times L \times W$ we wish to decide as quickly as possible whether this tuple leads to a read or a skip action.

The way to take such a decision is to use a *decision tree* as described in Figure 7.3. It puts Length as dominant feature on top and tells that a long article is skipped immediately. Indeed, this is what we see in the table in Figure 7.2: we can quickly take this decision without inspecting the other features. If the article is short, the next most relevant feature, after Length, is Thread. Again we can see in Figure 7.2 that a short and new article is read. Finally, we see in Figure 7.3 that if the article is short and a follow-up, then it is read if the author is known, and skipped if

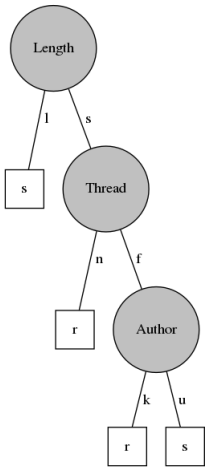


Figure 7.3 Decision tree for reading (r) or skipping (s) derived from the data in Figure 7.2.

the author is unkown. Apparently the location of the user (the Where feature) is irrelevant for the read / skip decision.

We see that such decision trees provide a (visually) clear and efficient method for reaching a decision about the target feature, starting from input features. The question that we wish to address here is: how to learn (derive) such a decision tree (in Figure 7.3) from the table (in Figure 7.2)?

The recipe (algorithm) for learning the tree involves iteratively going through the following three steps, acting on a joint state σ.

1 Check if we are done, that is, if the U-marginal (for User action) of the current joint state σ is a point state; if so, we are done and write this point state $1|x\rangle$ in a box as leaf in the tree.

2 If we are not done, determine the 'dominant' feature X for σ, and write X as a circle in the tree. This 'domination' is formulated in terms of lowest conditional entropy, see Remark 7.2.5, giving the lowest uncertainty, *i.e.* the highest certainty.

3 For each element of $x \in X$, update (and marginalise) σ with the point predicate $\mathbf{1}_x$ and re-start with step 1.

We shall describe these steps in more detail, starting from the state ω in (7.35), corresponding to the table in Figure 7.2.

1.1 The U-marginal $\omega_U := \omega[0,0,0,0,1] \in \mathcal{D}(U)$ equals $\frac{1}{2}|r\rangle + \frac{1}{2}|s\rangle$, because there are equal numbers of read and skip actions in Figure 7.2. Since this is not a point state, we are not done.

1.2 We need to determine the dominant feature in ω. This is where disintegration comes in.

We first extract four channels and states:

$$\begin{aligned}
c_A &:= \omega[0,0,0,0,1 \,|\, 1,0,0,0,0] \,:\, A \rightarrowtail U \\
\omega_A &:= \omega[1,0,0,0,0] \in \mathcal{D}(A) \\
c_T &:= \omega[0,0,0,0,1 \,|\, 0,1,0,0,0] \,:\, T \rightarrowtail U \\
\omega_T &:= \omega[0,1,0,0,0] \in \mathcal{D}(T) \\
c_L &:= \omega[0,0,0,0,1 \,|\, 0,0,1,0,0] \,:\, L \rightarrowtail U \\
\omega_L &:= \omega[0,0,1,0,0] \in \mathcal{D}(L) \\
c_W &:= \omega[0,0,0,0,1 \,|\, 0,0,0,1,0] \,:\, W \rightarrowtail U \\
\omega_W &:= \omega[0,0,0,1,0] \in \mathcal{D}(W).
\end{aligned}$$

Note that these channels go in the opposite direction with respect to the approach for naive Bayesian classification. It is not hard to see that:

$$\begin{aligned}
c_A(k) &= \tfrac{1}{2}|s\rangle + \tfrac{1}{2}|r\rangle & c_A(u) &= \tfrac{1}{2}|s\rangle + \tfrac{1}{2}|r\rangle & \omega_A &= \tfrac{2}{3}|k\rangle + \tfrac{1}{3}|u\rangle \\
c_T(n) &= \tfrac{3}{10}|s\rangle + \tfrac{7}{10}|r\rangle & c_T(f) &= \tfrac{3}{4}|s\rangle + \tfrac{1}{4}|r\rangle & \omega_T &= \tfrac{5}{9}|n\rangle + \tfrac{4}{9}|f\rangle \\
c_L(l) &= 1|s\rangle & c_L(s) &= \tfrac{2}{11}|s\rangle + \tfrac{9}{11}|r\rangle & \omega_L &= \tfrac{7}{18}|l\rangle + \tfrac{11}{18}|s\rangle \\
c_W(h) &= \tfrac{1}{2}|s\rangle + \tfrac{1}{2}|r\rangle & c_W(w) &= \tfrac{1}{2}|s\rangle + \tfrac{1}{2}|r\rangle & \omega_W &= \tfrac{4}{9}|h\rangle + \tfrac{5}{9}|w\rangle.
\end{aligned}$$

In order to determine which of the input features A, T, L, W is most dominant we compute the conditional entropies — sometimes called intrinsice value — for each of these features. Recall the Shannon entropy function $H\colon \mathcal{D}(X) \to \mathbb{R}_{\geq 0}$ from Definition 2.10.7 (2) and its relative form from Remark 7.2.5. Post-composing it with the above channels gives a factor, like $H \circ c_A\colon A \to \mathbb{R}_{\geq 0}$. It computes the entropy $H(c_A(x))$ for each $x \in A$. Hence we can compute its validity in the marginal state $\omega_A \in \mathcal{D}(A)$, giving the relative entropy. Explicitly:

$$\begin{aligned}
\omega_A \models H \circ c_A &= \sum_{x \in A} \omega_A(x) \cdot H(c_A(x)) \\
&= \omega_A(k) \cdot H(c_A(k)) + \omega_A(u) \cdot H(c_A(u)) \\
&= \tfrac{2}{3} \cdot \Big(\tfrac{1}{2} \cdot -\ln\big(\tfrac{1}{2}\big) + \tfrac{1}{2} \cdot -\ln\big(\tfrac{1}{2}\big)\Big) + \tfrac{1}{3} \cdot \Big(\tfrac{1}{2} \cdot -\ln\big(\tfrac{1}{2}\big) + \tfrac{1}{2} \cdot -\ln\big(\tfrac{1}{2}\big)\Big) \;=\; 0.69
\end{aligned}$$

In the same way one computes the other expected entropies as:

$$\omega_T \models H \circ c_T \;=\; 0.59 \qquad \omega_L \models H \circ c_L \;=\; 0.29 \qquad \omega_W \models H \circ c_W \;=\; 0.69.$$

One then picks the lowest relative entropy value, which is 0.29, for feature / component L. Hence L = Length is the dominant feature at this first stage. Therefore it is put on top in the decision tree in Figure 7.3.

1.3 The set L has two elements, l for long and s for short. We update the current state ω with each of these, via suitably weakened point predicates $\mathbf{1}_l$ and $\mathbf{1}_s$, and marginalise out the L component. This gives new states for which we use the following *ad hoc* slash notation.

$$\begin{aligned}
\omega/l &:= \omega\big|_{\mathbf{1}\otimes\mathbf{1}_l\otimes\mathbf{1}\otimes\mathbf{1}}[1,1,0,1,1] \in \mathcal{D}(A \times T \times W \times U) \\
\omega/s &:= \omega\big|_{\mathbf{1}\otimes\mathbf{1}_s\otimes\mathbf{1}\otimes\mathbf{1}}[1,1,0,1,1] \in \mathcal{D}(A \times T \times W \times U).
\end{aligned}$$

We now go into a recursive loop and repeat the previous steps for both these states ω/l and ω/s. Notice that they are 'shorter' than ω, since they only have 4 components instead of 5, since we marginalised out the dominant component L.

2.1.1 In the l-branch we are now done, since the U-marginal of the l-update is a point state:

$$\omega/l\,[0,0,0,1] \;=\; 1|s\rangle.$$

It means that a long article is skipped immedately. This is indicated via the the l-box as (left) child of the Length node in the decision tree in Figure 7.3.

2.1.2 We continue with the s-branch. The U-marginal of ω/s is not a point state.

2.2.2 We will now have to determine the dominant feature in ω/s. We compute the three relative entropies for A, T, W as:

$$\begin{aligned}
\omega/s\,[1,0,0,0] \models H \circ \omega/s\,[0,0,0,1\,\big|\,1,0,0,0] \;&=\; 0.31\\
\omega/s\,[0,1,0,0] \models H \circ \omega/s\,[0,0,0,1\,\big|\,0,1,0,0] \;&=\; 0.25\\
\omega/s\,[0,0,1,0] \models H \circ \omega/s\,[0,0,0,1\,\big|\,0,0,1,0] \;&=\; 0.47.
\end{aligned}$$

The second value is the lowest, so that T = Thread is now the dominant feature. It is added as codomain node of the s-edge out Length in the decision tree in Figure 7.3.

2.3.2 The set T has two elements, n for new and f for follow-up. We take the corresponding updates:

$$\begin{aligned}
\omega/s/n \;&:=\; \omega/s\big|_{\mathbf{1}\otimes\mathbf{1}_n\otimes\mathbf{1}\otimes\mathbf{1}}[1,0,1,1] \;\in\; \mathcal{D}(A\times W\times U)\\
\omega/s/f \;&:=\; \omega/s\big|_{\mathbf{1}\otimes\mathbf{1}_f\otimes\mathbf{1}\otimes\mathbf{1}}[1,0,1,1] \;\in\; \mathcal{D}(A\times W\times U).
\end{aligned}$$

Then we enter new recursions with both these states.

3.1.1 In the n-branch we are done, since we find a point state as U-marginal:

$$\omega/s/n\,[0,0,1] \;=\; 1|r\rangle.$$

This settles the left branch under the Thread node in Figure 7.3.

3.1.2 The f-branch is not done, since the U-marginal of the state $\omega/s/f$ is not a point state.

3.2.2 We thus start computing relative entropies again, in order to find out which of the remaining input features A, W is dominant.

$$\begin{aligned}
\omega/s/f\,[1,0,0] \models H \circ \omega/s/f\,[0,0,1\,\big|\,1,0,0] \;&=\; 0\\
\omega/s/f\,[0,1,0] \models H \circ \omega/s/f\,[0,0,1\,\big|\,0,1,0] \;&=\; 0.69.
\end{aligned}$$

Hence the A feature is dominant, so that the Author node is added to the f-edge out of Thread in Figure 7.3.

3.3.2 The set A has two elements, k for known and and u for unknown. We form the corresponding two updates of the current state $\omega/s/f$.

$$\begin{aligned}
\omega/s/f/k \;&:=\; \omega/s/f\big|_{\mathbf{1}_k\otimes\mathbf{1}\otimes\mathbf{1}}[0,1,1] \;\in\; \mathcal{D}(W\times U)\\
\omega/s/f/u \;&:=\; \omega/s/f\big|_{\mathbf{1}_u\otimes\mathbf{1}\otimes\mathbf{1}}[0,1,1] \;\in\; \mathcal{D}(W\times U).
\end{aligned}$$

The next cycle continues with these two states.

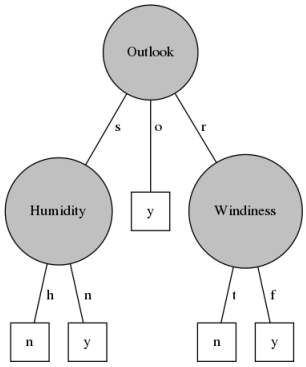


Figure 7.4 Decision tree for playing (y) or not (n) derived from the data in Figure 7.1.

4.1.1 In the k-branch we are done since:

$$\omega/s/f/k\,[0,1] \;=\; 1|r\rangle.$$

4.1.2 Also in the u-branch we are done since:

$$\omega/s/f/u\,[0,1] \;=\; 1|s\rangle.$$

This gives the last two boxes, so that the decision tree in Figure 7.3 is finished.

There are many variations on the above learning algorithm for decision trees. The one that we just described is sometime called the 'classification' version, as in Jensen and Nielsen (2007), since it works with discrete distributions. There is also a 'regression' version, for continuous distributions. The key part of the the above algorithm is deciding which feature is dominant (in step 2). We have described the so-called ID3 version from Quinlan (1986), which uses relative entropies (intrinsic values). One can understand this step in terms of the information value given by successive questions in a riddle, see *e.g.* (Cox, 1961, II,7).

Sometimes the relative entropy is described in terms of 'gains', see Mitchell (1997), where in the above step 1.2 we can define for feature $X \in \{A, T, L, W\}$,

$$gain(X) \;\coloneqq\; H(\omega_X) - (\omega_X \models H \circ c_X).$$

One then looks for the feature with the highest gain. But one may as well look for the lowest expected entropy — given by the valdity expression after the minus sign — as we do above. There are alternatives to using gain, such as what is called 'gini', but that is out of scope.

To conclude, running the decision tree learning algorithm on the distribution associated with the weather and play table from Figure 7.1 — with Play as target feature — yields the decision tree in Figure 7.4, see also (Witten et al., 2011, Fig. 4.4).

Exercises

7.5.1 Write $\sigma \in \mathcal{M}(O \times T \times H \times W \times P)$ for the weather-play table in Figure 7.1, as multidimensional *multiset*.

1 Compute the marginal multiset $\sigma[1,0,0,0,1] \in \mathcal{M}(O \times P)$.

2 Reorganise this marginalised multiset as a 2-dimensional table with only Outlook (horizontal) and Play (vertical) data, as given below, and check how this 'marginalised' table relates to the original one in Figure 7.1.

	Sunny	**Overcast**	**Rainy**
yes	2	4	3
no	3	0	2

3 Deduce a channel $P \multimap O$ from this table, and compare it to the description (7.34) of the channel c_O given in Subsection 7.5.1 — see also Lemma 2.2.5.

4 Do the same for the marginal tables $\sigma[0,1,0,0,1]$, $\sigma[0,0,1,0,1]$, $\sigma[0,0,0,1,1]$, and the corresponding channels c_T, c_H, c_W in Subsection 7.5.1.

7.5.2 Continuing with the weather-play table σ, notice that instead of taking the dagger of the tuple of channels $\langle c_O, c_T, c_H, c_W \rangle \colon P \to O \times T \times H \times W$ in Example 7.5.1 we could have used instead the 'direct' disintegration:

$$O \times T \times H \times W \xrightarrow{\sigma[0,0,0,0,1\,|\,1,1,1,1,0]} P$$

Check that this gives a division-by-zero error.

7.5.3 Naive Bayesian classification, as illustrated in Subsection 7.5.1, is often used for classifying email messages as either 'spam' or 'ham' (not spam). One then looks for words which are typical for spam or ham. This exercise elaborates a standard small example.

Consider the following table of six words, together with the likelihoods of them being spam or ham.

	spam	**ham**
review	$1/4$	1
send	$3/4$	$1/2$
us	$3/4$	$1/2$
your	$3/4$	$1/2$
password	$1/2$	$1/2$
account	$1/4$	0

Lets write $S = \{s, h\}$ for the probability space for spam and ham, and, as usual $\mathbf{2} = \{0, 1\}$.

1 Translate the above table into six channels:

$$\text{review}, \text{send}, \text{us}, \text{your}, \text{password}, \text{account} \;:\; S \multimap \mathbf{2}.$$

Write $c \colon S \multimap \mathbf{2}^6 = \mathbf{2} \times \mathbf{2} \times \mathbf{2} \times \mathbf{2} \times \mathbf{2} \times \mathbf{2}$ for the tuple channel:

$$c \coloneqq \langle \text{review}, \text{send}, \text{us}, \text{your}, \text{password}, \text{account} \rangle.$$

2 We wish to classify the message "review us now". Explain how it gets translated to the point predicate $\mathbf{1}_{(1,0,1,0,0,0)}$ on 2^6.

3 Show that:

$$\begin{aligned} & c^*(\mathbf{1}_{(1,0,1,0,0,0)}) \\ &= \text{review}^*(\mathbf{1}_1) \ \& \ \text{send}^*(\mathbf{1}_0) \ \& \ \text{us}^*(\mathbf{1}_1) \ \& \ \text{your}^*(\mathbf{1}_0) \ \& \ \text{password}^*(\mathbf{1}_0) \ \& \ \text{account}^*(\mathbf{1}_0) \\ &= \tfrac{9}{2048} \cdot \mathbf{1}_s + \tfrac{1}{16} \cdot \mathbf{1}_h \end{aligned}$$

4 Let's assume a prior spam distribution $\omega = \frac{2}{3}|s\rangle + \frac{1}{3}|h\rangle$. Show that the posterior spam distribution for our message is:

$$c^\dagger_\omega(1,0,1,0,0,0) = \omega|_{c^*(\mathbf{1}_{(1,0,1,0,0,0)})} = \tfrac{9}{73}|s\rangle + \tfrac{64}{73}|h\rangle \approx 0.1233|s\rangle + 0.8767|h\rangle.$$

7.5.4 Here is another example in naive Bayesian classification that is commonly used as illustration[3]. The aim is to classify fruit as 'banana', 'orange', or 'other' (neither), for which we use the space $\{b, o, n\}$. Three attributes are considered: 'long', 'sweet', and 'yellow', for which we have three spaces $\{l, l^\perp\}$, $\{s, s^\perp\}$ and $\{y, y^\perp\}$. The training data is summarised in the following table.

	long	**not long**	**sweet**	**not sweet**	**yellow**	**not yellow**
banana	400	100	350	150	450	50
orange	0	300	150	150	300	0
other	100	100	150	50	50	150

1 Turn this table first into three channels:

$$\{b, o, n\} \multimap \{l, l^\perp\} \qquad \{b, o, n\} \multimap \{s, s^\perp\} \qquad \{b, o, n\} \multimap \{y, y^\perp\},$$

and then combine them into a single channel

$$\{b, o, n\} \longrightarrow\!\!\circ\!\!\longrightarrow \{l, l^\perp\} \times \{s, s^\perp\} \times \{y, y^\perp\}.$$

2 Assume a base rate distribution $\omega = \frac{1}{2}|b\rangle + \frac{3}{10}|o\rangle + \frac{1}{5}|n\rangle$. We are confronted with an item that is long and sweet and yellow. Show that it is 93% certain that it is a banana.

7.5.5 Show in detail that we get point states in items [2.1.1], [3.1.1], [4.1.1] and [4.1.2] in Subsection 7.5.2.

7.5.6 In Subsection 7.5.1 we classified the play distribution as $\frac{125}{611}|y\rangle + \frac{486}{611}|n\rangle$, given the input features (s, c, h, t). Check what the play decision is for these inputs in the decision tree in Figure 7.4.

7.5.7 As mentioned, the table in Figure 7.2 is copied from Poole and Mackworth (2017). There the following two questions are asked: what can be said about reading / skipping for the two queries:

$$\begin{aligned} Q_1 &= (\text{unknown, new, long, work}) \\ Q_2 &= (\text{unknown, follow-up, short, home}). \end{aligned}$$

1 Answer Q_1 and Q_2 via the decision tree in Figure 7.3.

2 Compute the distributions on U for both queries Q_1 and Q_2 via naive Bayesian classification.

[3] See *e.g.* the website https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/

(The outcome for Q_2 obtained via Bayesian classification is $\frac{3}{5}|r\rangle + \frac{2}{5}|s\rangle$; it gives reading the highest probability. This does not coincide with the outcome via the decision tree. Thus, one should be careful to rely on such classification methods for important decisions.)

7.6 Daggers of draw channels and Laplace's rule of succession

In this section we are going to review results that can be grouped under the umbrella of 'Laplace's rule of succession'. In general, this rule is about learning the future from past events. This concerns the expected outcome after an update, see *e.g.* Ross (2018) or Blitzstein and Hwang (2019). In the present setting we reformulate the rule informally as: the succession is given by the mean of a dagger. Then we might as well ask about its variance. As we have seen earlier in this chapter, a dagger is a channel that takes an observation as input and produces an updated prior. It is the mean and variance of the updated prior that we are interested in, as indication of what comes next, after the observation. It is thus a pointwise mean / variance of a channel that we will focus on.

As a preparation for the results to come we look at the relatively simple situation of a binomial (draw) channel with a Poisson prior. We recall from Exercise 3.10.1 that binomial after Poisson is still Poisson, as made precise by the equation $bn[-](r)_*(pois[\lambda]) = pois[r \cdot \lambda]$. We extend the relationship between binomial and Poisson to updating. In essence, the next result is in (Jaynes, 2003, §6.11). It applies for instance when one has an imperfect particle counter in physics, where particles appear according to a Poisson distribution (with rate λ) and the imperfection is captured by a subsequent binomial channel (with chance r of being detected). When k particles are counted, the question is: what is the total number of emitted particles that gave rise to these K that were observed? The distribution of these total numbers is obtained by updating the prior Poisson distribution. Interestingly, this yields a new Poisson distribution, but with a shift.

A multivariate version of this result, involving particles of different kinds, is in Exercise 7.6.7.

Proposition 7.6.1 *Let $r \in [0,1]$ and $\lambda \in \mathbb{R}_{\geq 0}$ be given.*

1 *For $k \in \mathbb{N}$ the dagger of the channel $bn[-](r) \colon \mathbb{N} \to \mathcal{D}(\mathbb{N})$ with prior $pois[\lambda] \in \mathcal{D}_\infty(\mathbb{N})$ is of the form:*

$$\begin{aligned} bn[-](r)^\dagger_{pois[\lambda]}(k) &= \sum_{N \geq 0} pois[(1-r)\cdot\lambda](N) \,\big|k+N\rangle \\ &= \mathcal{D}_\infty(k + (-))\Big(pois[(1-r)\cdot\lambda]\Big) \;=\; 1|k\rangle + pois[(1-r)\cdot\lambda]. \end{aligned}$$

The latter formuation involves a shift via convolution, as in Exercise 2.7.8, using the (additive) commutative monoid structure of $\mathbb{N}$.

2 *The mean of this dagger is the number in $\mathbb{R}_{\geq 0}$ given by the formula:*

$$Mean\Big(bn[-](r)^\dagger_{pois[\lambda]}(k)\Big) \;=\; k + (1-r)\cdot\lambda.$$

3 *The variance of this dagger is:*

$$Var\Big(bn[-](r)^\dagger_{pois[\lambda]}(k)\Big) \;=\; (1-r)\cdot\lambda.$$

Proof. 1 By Exercise 3.10.1:

$$\begin{aligned}
bn[-](r)^{\dagger}_{pois[\lambda]}(k) &= \sum_{K\geq k} \frac{pois[\lambda](K)\cdot bn[K](r)(k)}{bn[-](r)_{*}(pois[\lambda])(k)} \,\big|K\rangle \\
&= \sum_{K\geq k} \frac{pois[\lambda](K)\cdot bn[K](r)(k)}{pois[r\cdot\lambda](k)} \,\big|K\rangle \\
&= \sum_{K\geq k} \frac{e^{-\lambda}\cdot \frac{\lambda^K}{K!}\cdot \binom{K}{k}\cdot r^k\cdot (1-r)^{K-k}}{e^{-r\cdot\lambda}\cdot \frac{(r\cdot\lambda)^k}{k!}} \,\big|K\rangle \\
&= \sum_{K\geq k} e^{(1-r)\cdot\lambda}\cdot \frac{\lambda^{K-k}}{(K-k)!}\cdot (1-r)^{K-k}\,\big|K\rangle \\
&= \sum_{N\geq 0} e^{(1-r)\cdot\lambda}\cdot \frac{((1-r)\cdot\lambda)^N}{N!}\,\big|k+N\rangle \;=\; \sum_{N\geq 0} pois[(1-r)\cdot\lambda](N)\,\big|k+N\rangle.
\end{aligned}$$

2 We use that the mean preserves sums, see Lemma 4.2.4, so that:

$$\begin{aligned}
Mean\Big(bn[-](r)^{\dagger}_{pois[\lambda]}(k)\Big) &= Mean\Big(1|k\rangle + pois[(1-r)\cdot\lambda]\Big) \\
&= Mean\Big(1|k\rangle\Big) + Mean\Big(pois[(1-r)\cdot\lambda]\Big) \\
&= k + (1-r)\cdot\lambda, \qquad \text{by Example 4.2.2 (3).}
\end{aligned}$$

3 Similarly, by Exercises 5.1.12 and 5.1.8,

$$\begin{aligned}
Var\Big(bn[-](r)^{\dagger}_{pois[\lambda]}(k)\Big) &= Var\Big(1|k\rangle + pois[(1-r)\cdot\lambda]\Big) \\
&= Var\Big(1|k\rangle\Big) + Var\Big(pois[(1-r)\cdot\lambda]\Big) \;=\; (1-r)\cdot\lambda.
\end{aligned}$$

❑

In order to find the mean of a dagger a multivariate draw channel, we first have to compute the dagger. This section concentrates mostly on the hypergeometric channels, for various priors, and discusses Pólya channels for the uniform prior only. This will lead to a series of results, all in the style of Proposition 7.6.1 above. The same approach can be followed for multinomial (including binomial) channels, but that requires continuous probability theory. This is postponed to **??**. Calculating the dagger of a channel may require some work. We elaborate this for the hypergeometric channel, see Definition 2.6.1 and Section 3.5 for details. We describe its dagger for four different priors: uniform, multinomial, Pólya.

Below we use the mean, like in Proposition 4.6.3 and continued in Section 5.2, as a function $Mean\colon \mathcal{D}(\mathcal{M}(X)) \to \mathcal{M}(X)$, given as $Mean(\sum_i r_i|\varphi_i\rangle) = \sum_i r_i\cdot\varphi_i$.

Proposition 7.6.2 *Let X be a non-empty and finite set, say with $N := |X| > 0$ elements. Consider the hypergeometric channel $hg[K]\colon \mathcal{N}[L](X) \rightsquigarrow \mathcal{N}[K](X)$, for numbers $L \geq K$. We take as prior the uniform distribution on the (finite) set $\mathcal{N}[L](X)$ of multisets of size L.*

1 *The hypergeometric dagger channel $hg[K]^{\dagger}_{unif_{\mathcal{N}[L](X)}}\colon \mathcal{N}[K](X) \rightsquigarrow \mathcal{N}[L](X)$ acts on $\varphi \in \mathcal{N}[K](X)$ and produces a distribution on $\mathcal{N}[L](X)$, of the form:*

$$hg[K]^{\dagger}_{unif_{\mathcal{N}[L](X)}}(\varphi) \;=\; \sum_{\psi\in\mathcal{N}[L-K](X)} \frac{\binom{\varphi+\psi}{\varphi}}{\binom{L+N-1}{K+N-1}}\,\big|\varphi+\psi\rangle.$$

2 *The mean of this dagger can be described in terms of the observed draw φ as:*

$$Mean\Big(hg[K]^{\dagger}_{unif_{\mathcal{N}[L](X)}}(\varphi)\Big) \;=\; \frac{L+N}{K+N}\cdot\varphi \;+\; \frac{L-K}{K+N}\cdot\mathbf{1}.$$

3 *The variance of the dagger is:*

$$Var\Big(hg[K]^{\dagger}_{unif_{\mathcal{N}[L](X)}}(\varphi)\Big) \;=\; \frac{(L-K)\cdot(L+N)}{(K+N+1)\cdot(K+N)^2}\cdot\Big(-\varphi^2+(K+N-2)\cdot\varphi+(K+N-1)\cdot\mathbf{1}\Big).$$

The binary version of the first point, for $N = 2$, occurs in (Jaynes, 2003, §6.3).

Proof. 1 Since we have a uniform distribution as prior, Lemma 7.1.4 (1) applies:

$$\begin{aligned}
hg[K]^{\dagger}_{unif_{\mathcal{N}[L](X)}}(\varphi) \;&=\; \sum_{\upsilon\in\mathcal{N}[L](X),\,\varphi\leq_K\upsilon}\frac{hg[K](\upsilon)(\varphi)}{\sum_{\nu\in\mathcal{N}[L](X),\,\varphi\leq_K\nu}hg[K](\nu)(\varphi)}\,\big|\upsilon\rangle\\
&=\; \sum_{\psi\in\mathcal{N}[L-K](X)}\frac{\binom{\varphi+\psi}{\varphi}/\binom{L}{K}}{\sum_{\nu\in\mathcal{N}[L](X),\,\varphi\leq_K\nu}\binom{\nu}{\varphi}/\binom{L}{K}}\,\big|\varphi+\psi\rangle\\
&=\; \sum_{\psi\in\mathcal{N}[L-K](X)}\frac{\binom{\varphi+\psi}{\varphi}}{\left(\!\binom{K+N}{L-K}\!\right)}\,\big|\varphi+\psi\rangle \quad \text{by Exercise 1.8.9.}
\end{aligned}$$

2 We use the previous point in combination with the first two items of Exercise 1.8.9.

$$\begin{aligned}
Mean\Big(hg[K]^{\dagger}_{unif_{\mathcal{N}[L](X)}}(\varphi)\Big) \;&=\; \sum_{x\in X}\sum_{\psi\in\mathcal{N}[L-K](X)}\frac{\binom{\varphi+\psi}{\varphi}}{\left(\!\binom{K+N}{L-K}\!\right)}\cdot((\varphi+\psi)(x))\,\big|x\rangle\\
&=\; \sum_{x\in X}\varphi(x)\cdot\frac{\sum_{\psi\in\mathcal{N}[L-K](X)}\binom{\varphi+\psi}{\varphi}}{\left(\!\binom{K+N}{L-K}\!\right)} \;+\; \frac{\sum_{\psi\in\mathcal{N}[L-K](X)}\psi(x)\cdot\binom{\varphi+\psi}{\varphi}}{\left(\!\binom{K+N}{L-K}\!\right)}\,\big|x\rangle\\
&=\; \sum_{x\in X}\varphi(x)\cdot 1 \;+\; (\varphi(x)+1)\cdot\frac{\left(\!\binom{K+N+1}{L-K-1}\!\right)}{\left(\!\binom{K+N}{L-K}\!\right)}\,\big|x\rangle\\
&=\; \sum_{x\in X}\varphi(x) \;+\; (\varphi(x)+1)\cdot\frac{L-K}{K+N}\,\big|x\rangle\\
&=\; \sum_{x\in X}\varphi(x)\cdot\frac{L+N}{K+N} \;+\; \frac{L-K}{K+N}\,\big|x\rangle \;=\; \frac{L+N}{K+N}\cdot\varphi \;+\; \frac{L-K}{K+N}\cdot\mathbf{1}.
\end{aligned}$$

3 See Exercise 7.6.5. ❑

We turn to a multinomial prior.

Proposition 7.6.3 *For $\omega \in \mathcal{D}(X)$ we now take the multinomial distribution $mn[L](\omega)$ on $\mathcal{N}[L](X)$ as prior.*

1 *The dagger of the K-sized hypergeometric channel at $\varphi \in \mathcal{N}[K](X)$ is a shifted multinomial:*

$$\begin{aligned}
hg[K]^{\dagger}_{mn[L](\omega)}(\varphi) \;&=\; \sum_{\psi\in\mathcal{N}[L-K](X)} mn[L-K](\omega)(\psi)\,\big|\varphi+\psi\rangle\\
&=\; \mathcal{D}(\varphi+(-))\Big(mn[L-K](\omega)\Big) \;=\; 1\big|\varphi\rangle \;+\; mn[L-K](\omega).
\end{aligned}$$

This last formuation involves a convolution using the commutative monoid structure of $\mathcal{N}(X)$.

2 *The mean of this dagger can be expressed in terms of the distribution ω and observed draw φ as:*

$$Mean\Big(hg[K]^{\dagger}_{mn[L](\omega)}(\varphi)\Big) \;=\; \varphi \;+\; (L-K)\cdot\omega.$$

3 *The variance of the dagger can be expressed similarly as:*

$$Var\Big(hg[K]^{\dagger}_{mn[L](\omega)}(\varphi)\Big) \;=\; (L-K)\cdot\omega\cdot(\mathbf{1}-\omega).$$

Proof. 1 We use that $hg[K] \circ mn[L] = mn[K]$, see Corollary 3.5.2 (6).

$$\begin{aligned}
hg[K]^{\dagger}_{mn[L](\omega)}(\varphi) &= \sum_{\upsilon \in \mathcal{N}[L](X)} \frac{hg[K](\upsilon)(\varphi) \cdot mn[L](\omega)(\upsilon)}{hg[K]_{*}(mn[L](\omega))(\varphi)} \,\big|\upsilon\big\rangle \\
&= \sum_{\upsilon \in \mathcal{N}[L](X)} \frac{hg[K](\upsilon)(\varphi) \cdot mn[L](\omega)(\upsilon)}{mn[K](\omega)(\varphi)} \,\big|\upsilon\big\rangle \\
&= \sum_{\upsilon \in \mathcal{N}[L](X),\, \varphi \leq_K \upsilon} \frac{\binom{\upsilon}{\varphi}}{\binom{L}{K}} \cdot \frac{L!}{\upsilon!} \cdot \frac{\varphi!}{K!} \cdot \frac{\prod_{x \in X} \omega(x)^{\upsilon(x)}}{\prod_{x \in X} \omega(x)^{\varphi(x)}} \,\big|\upsilon\big\rangle \\
&= \sum_{\upsilon \in \mathcal{N}[L](X),\, \varphi \leq_K \upsilon} \frac{(L-K)!}{(\upsilon - \varphi)!} \cdot \prod_{x \in X} \omega(x)^{\upsilon(x) - \varphi(x)} \,\big|\upsilon\big\rangle \\
&= \sum_{\psi \in \mathcal{N}[L-K](X)} mn[L-K](\omega)(\psi) \,\big|\varphi + \psi\big\rangle.
\end{aligned}$$

2 We provide two proofs. The first one uses the formulation of this dagger in combination with Lemma 3.3.3 in:

$$\begin{aligned}
Mean\Big(hg[K]^{\dagger}_{mn[L](\omega)}(\varphi)\Big) &= \sum_{x \in X} \sum_{\psi \in \mathcal{N}[L-K](X)} mn[L-K](\omega)(\psi) \cdot (\varphi + \psi)(x) \,\big|x\big\rangle \\
&= \sum_{x \in X} \varphi(x) \cdot \sum_{\psi \in \mathcal{N}[L-K](X)} mn[L-K](\omega)(\psi) \\
&\qquad + \sum_{\psi \in \mathcal{N}[L-K](X)} mn[L-K](\omega)(\psi) \cdot \psi(x) \,\big|x\big\rangle \\
&= \sum_{x \in X} \varphi(x) \,+\, (L-K) \cdot \omega(x) \,\big|x\big\rangle \;=\; \varphi \,+\, (L-K) \cdot \omega.
\end{aligned}$$

The second, shorter proof uses that the mean function preserves convolution, as in Lemma 4.2.4, so that by Proposition 4.6.3 (1):

$$\begin{aligned}
Mean\Big(hg[K]^{\dagger}_{mn[L](\omega)}(\varphi)\Big) &= Mean\Big(1\big|\varphi\big\rangle \,+\, mn[L-K](\omega)\Big) \\
&= Mean\Big(1\big|\varphi\big\rangle\Big) \,+\, Mean\Big(mn[L-K](\omega)\Big) \;=\; \varphi \,+\, (L-K) \cdot \omega.
\end{aligned}$$

3 We use Proposition 5.2.1 (1).

$$\begin{aligned}
Var\Big(hg[K]^{\dagger}_{mn[L](\omega)}(\varphi)\Big) &= Var\Big(1\big|\varphi\big\rangle \,+\, mn[L-K](\omega)\Big) \\
&= Var\Big(1\big|\varphi\big\rangle\Big) \,+\, Var\Big(mn[L-K](\omega)\Big) \\
&= 0 + Var\Big(mn[L-K](\omega)\Big) \;=\; (L-K) \cdot \omega \cdot (\mathbf{1} - \omega).
\end{aligned}$$

❑

The obvious next step is to consider a Pólya prior.

Proposition 7.6.4 *For $\upsilon \in \mathcal{N}(X)$ be an urn with full support with size $M := \|\upsilon\| > 0$. We take the distribution $pl[L](\upsilon)$ on $\mathcal{N}[L](X)$ as prior.*

1 The dagger hypergeometric channel at $\varphi \in \mathcal{N}[K](X)$ is a shifted Pólya distribution:

$$\begin{aligned}
hg[K]^{\dagger}_{pl[L](\upsilon)}(\varphi) &= \sum_{\psi \in \mathcal{N}[L-K](X)} pl[L-K](\upsilon + \varphi)(\psi) \,\big|\varphi + \psi\big\rangle \\
&= \mathcal{D}(\varphi + (-))\Big(pl[L-K](\upsilon + \varphi)\Big) \;=\; 1\big|\varphi\big\rangle \,+\, pl[L-K](\upsilon + \varphi).
\end{aligned}$$

2 *The mean of this dagger can be expressed in terms of the distribution ω and observed draw φ as:*

$$Mean\Big(hg[K]^{\dagger}_{pl[L](\upsilon)}(\varphi)\Big) \;=\; \varphi \;+\; (L-K)\cdot Flrn(\upsilon+\varphi).$$

3 *The variance of the dagger is:*

$$Var\Big(hg[K]^{\dagger}_{pl[L](\upsilon)}(\varphi)\Big) \;=\; \frac{(L-K)\cdot(M+L)}{M+K+1}\cdot Flrn(\upsilon+\varphi)\cdot\Big(\mathbf{1}-Flrn(\upsilon+\varphi)\Big).$$

Proof. 1 Let urn $\upsilon \in \mathcal{N}(X)$ have size $M := \|\upsilon\| > 0$ with $supp(\upsilon) = X$. We use that $hg[K] \circ pl[L] = pl[K]$, see Proposition 3.6.2 (3). Then, for $\varphi \in \mathcal{N}[K](X)$,

$$\begin{aligned}
hg[K]^{\dagger}_{pl[L](\upsilon)}(\varphi) &= \sum_{\chi\in\mathcal{N}[L](X)} \frac{hg[K](\chi)(\varphi)\cdot pl[L](\upsilon)(\chi)}{hg[K]_{*}(pl[L](\upsilon))(\varphi)}\,\big|\chi\rangle \\
&= \sum_{\chi\in\mathcal{N}[L](X),\,\varphi\leq_K\chi} \frac{\binom{\chi}{\varphi}}{\binom{L}{K}}\cdot\frac{\left(\!\binom{\upsilon}{\chi}\!\right)}{\left(\!\binom{M}{L}\!\right)}\cdot\frac{\left(\!\binom{M}{K}\!\right)}{\left(\!\binom{\upsilon}{\varphi}\!\right)}\,\big|\chi\rangle \\
&= \sum_{\chi\in\mathcal{N}[L](X),\,\varphi\leq_K\chi} \frac{\chi\!\underset{\circ}{!}\cdot K!\cdot(L-K)!}{\varphi\!\underset{\circ}{!}\cdot(\chi-\varphi)\!\underset{\circ}{!}\cdot L!}\cdot\frac{(\upsilon+\chi-\mathbf{1})\!\underset{\circ}{!}\cdot L!\cdot(M-1)!}{\chi\!\underset{\circ}{!}\cdot(\upsilon-\mathbf{1})\!\underset{\circ}{!}\cdot(M+L-1)!} \\
&\qquad\qquad\cdot\frac{(M+K-1)!\cdot\varphi\!\underset{\circ}{!}\cdot(\upsilon-\mathbf{1})\!\underset{\circ}{!}}{K!\cdot(M-1)!\cdot(\upsilon+\varphi-\mathbf{1})\!\underset{\circ}{!}}\,\big|\chi\rangle \\
&= \sum_{\chi\in\mathcal{N}[L](X),\,\varphi\leq_K\chi} \frac{(L-K)!\cdot(\upsilon+\chi-\mathbf{1})\!\underset{\circ}{!}\cdot(M+K-1)!}{(\chi-\varphi)\!\underset{\circ}{!}\cdot(M+L-1)!\cdot(\upsilon+\varphi-\mathbf{1})\!\underset{\circ}{!}}\,\big|\chi\rangle \\
&= \sum_{\psi\in\mathcal{N}[L-K](X)} \frac{(L-K)!\cdot(\upsilon+\varphi+\psi-\mathbf{1})\!\underset{\circ}{!}\cdot(M+K-1)!}{\psi\!\underset{\circ}{!}\cdot(M+L-1)!\cdot(\upsilon+\varphi-\mathbf{1})\!\underset{\circ}{!}}\,\big|\varphi+\psi\rangle \\
&= \sum_{\psi\in\mathcal{N}[L-K](X)} \frac{\left(\!\binom{\upsilon+\varphi}{\psi}\!\right)}{\left(\!\binom{M+K}{L-K}\!\right)}\,\big|\varphi+\psi\rangle \;=\; \sum_{\psi\in\mathcal{N}[L-K](X)} pl[L-K](\upsilon+\varphi)(\psi)\,\big|\varphi+\psi\rangle.
\end{aligned}$$

2 The short proof for Proposition 7.6.3 (2) now works via Proposition 4.6.3 (3).

3 We use Proposition 5.2.1 (3):

$$\begin{aligned}
Var\Big(hg[K]^{\dagger}_{pl[L](\upsilon)}(\varphi)\Big) &= Var\Big(1\big|\varphi\rangle \;+\; pl[L-K](\upsilon+\varphi)\Big) \\
&= Var\Big(pl[L-K](\upsilon+\varphi)\Big) \\
&= (L-K)\cdot\frac{(M+K)+(L-K)}{M+K+1}\cdot Flrn(\upsilon+\varphi)\cdot\Big(\mathbf{1}-Flrn(\upsilon+\varphi)\Big) \\
&= \frac{(L-K)\cdot(M+L)}{M+K+1}\cdot Flrn(\upsilon+\varphi)\cdot\Big(\mathbf{1}-Flrn(\upsilon+\varphi)\Big).
\end{aligned}$$

❑

We finally consider a hypergeometric prior.

Proposition 7.6.5 *1 Let numbers $M \geq L \geq K$ be given. We fix an urn multiset $\upsilon \in \mathcal{N}[M](X)$ and take the hypergeometric distribution $hg[L](\upsilon)$ as prior. In order to have well-defined K-sized draws we have to restrict the domain of the hypergeometric dagger to a subset of $\mathcal{N}[K](X)$, as in:*

$$\{\varphi \in \mathcal{N}[K](X) \;\big|\; \varphi \leq_K \upsilon\} \xrightarrow{hg[K]^{\dagger}_{hg[L](\upsilon)}} \mathcal{N}[L](X)$$

On such a draw $\varphi \leq_K \upsilon$ this dagger is given by:

$$hg[K]^{\dagger}_{hg[L](\upsilon)}(\varphi) \;=\; \sum_{\psi \leq_{L-K} \upsilon-\varphi} hg[L-K](\upsilon-\varphi)(\psi)\,\big|\varphi+\psi\big\rangle \;=\; 1\big|\varphi\big\rangle \;+\; hg[L-K](\upsilon-\varphi).$$

2 *We then get as mean:*

$$Mean\Big(hg[K]^{\dagger}_{hg[L](\upsilon)}(\varphi)\Big) \;=\; \varphi \;+\; (L-K)\cdot Flrn(\upsilon-\varphi).$$

3 *And as variance:*

$$Var\Big(hg[K]^{\dagger}_{hg[L](\upsilon)}(\varphi)\Big) \;=\; \frac{(L-K)\cdot(M-L)}{M-K-1}\cdot Flrn(\upsilon-\varphi)\cdot\Big(\mathbf{1}-Flrn(\upsilon-\varphi)\Big).$$

Proof. 1 Assuming $\varphi \in \mathcal{N}[K](X)$ satisfies $\varphi \leq_K \upsilon$ we use that hypergeometric channels compose, as in Corollary 3.5.2 (5).

$$\begin{aligned}
hg[K]^{\dagger}_{hg[L](\upsilon)}(\varphi) &= \sum_{\chi\in\mathcal{N}[L](X)} \frac{hg[K](\chi)(\varphi)\cdot hg[L](\upsilon)(\chi)}{hg[K]_{*}(hg[L](\upsilon))(\varphi)}\,\big|\chi\big\rangle \\
&= \sum_{\chi\in\mathcal{N}[L](X)} \frac{hg[K](\chi)(\varphi)\cdot hg[L](\upsilon)(\chi)}{hg[K](\upsilon)(\varphi)}\,\big|\chi\big\rangle \\
&= \sum_{\chi\in\mathcal{N}[L](X),\,\varphi\leq_K\chi,\,\chi\leq_L\upsilon} \frac{\binom{\chi}{\varphi}}{\binom{L}{K}}\cdot\frac{\binom{\upsilon}{\chi}}{\binom{M}{L}}\cdot\frac{\binom{M}{K}}{\binom{\upsilon}{\varphi}}\,\big|\chi\big\rangle \\
&= \sum_{\chi\in\mathcal{N}[L](X),\,\varphi\leq_K\chi,\,\chi\leq_L\upsilon} \frac{\chi!\cdot K!\cdot(L-K)!}{\varphi!\cdot(\chi-\varphi)!\cdot L!}\cdot\frac{\upsilon!\cdot L!\cdot(M-L)!}{\chi!\cdot(\upsilon-\chi)!\cdot M!} \\
&\qquad\qquad \cdot\frac{M!\cdot\varphi!\cdot(\upsilon-\varphi)!}{K!\cdot(M-K)!\cdot\upsilon!}\,\big|\chi\big\rangle \\
&= \sum_{\chi\in\mathcal{N}[L](X),\,\varphi\leq_K\chi,\,\chi\leq_L\upsilon} \frac{(L-K)!\cdot(M-L)!\cdot(\upsilon-\varphi)!}{(\chi-\varphi)!\cdot(M-K)!\cdot(\upsilon-\chi)!}\,\big|\chi\big\rangle \\
&= \sum_{\psi\in\mathcal{N}[L-K](X),\,\varphi+\psi\leq_L\upsilon} \frac{(L-K)!\cdot(M-L)!\cdot(\upsilon-\varphi)!}{\psi!\cdot(M-K)!\cdot(\upsilon-\varphi-\psi)!}\,\big|\varphi+\psi\big\rangle \\
&= \sum_{\psi\leq_{L-K}\upsilon-\varphi} \frac{\binom{\upsilon-\varphi}{\psi}}{\binom{M-K}{L-K}}\,\big|\varphi+\psi\big\rangle \;=\; \sum_{\psi\leq_{L-K}\upsilon-\varphi} hg[L-K](\upsilon-\varphi)(\psi)\,\big|\varphi+\psi\big\rangle.
\end{aligned}$$

2 Via the convolution formulation and Proposition 4.6.3 (2).

3 Via Proposition 5.2.1 (2):

$$\begin{aligned}
Var\Big(hg[K]^{\dagger}_{hg[L](\upsilon)}(\varphi)\Big) &= Var\Big(1\big|\varphi\big\rangle \;+\; hg[L-K](\upsilon-\varphi)\Big) \\
&= Var\Big(hg[L-K](\upsilon-\varphi)\Big) \\
&= (L-K)\cdot\frac{(M-K)-(L-K)}{M-K-1}\cdot Flrn(\upsilon-\varphi)\cdot\Big(\mathbf{1}-Flrn(\upsilon-\varphi)\Big) \\
&= \frac{(L-K)\cdot(M-L)}{M-K-1}\cdot Flrn(\upsilon-\varphi)\cdot\Big(\mathbf{1}-Flrn(\upsilon-\varphi)\Big).
\end{aligned}$$

❑

We briefly look daggers of Pólya channels, only with a uniform prior. Other priors, like multinomial are possible, but we do not have a way to simplify them, in essence because we have no formula for the applying a Pólya channel to a multinomial distribution of urns.

Proposition 7.6.6 *Let X be a finite set with $N := |X| > 0$ elements. Consider the Pólya channel $pl[K]\colon \mathcal{N}_{fs}[L](X) \rightsquigarrow \mathcal{N}[K](X)$, for numbers $L > N, K \geq 0$. We take as prior the uniform distribution on the (finite) set $\mathcal{N}_{fs}[L](X)$ of multisets with full support of size L.*

1 The Pólya dagger channel $pl[K]^{\dagger}_{unif_{\mathcal{N}_{fs}[L](X)}} : \mathcal{N}[K](X) \rightsquigarrow \mathcal{N}_{fs}[L](X)$ works as follows.

$$pl[K]^{\dagger}_{unif_{\mathcal{N}_{fs}[L](X)}}(\varphi) = \sum_{\upsilon \in \mathcal{N}_{fs}[L](X)} \frac{\binom{\upsilon}{\varphi}}{\binom{K+N}{L-N}} \big|\upsilon\rangle.$$

2 The mean of this dagger can be described in terms of the observed draw φ as:

$$Mean\Big(pl[K]^{\dagger}_{unif_{\mathcal{N}_{fs}[L](X)}}(\varphi)\Big) = \frac{L-N}{K+N} \cdot \varphi + \frac{L+K}{K+N} \cdot \mathbf{1}.$$

3 The variance of the dagger is:

$$\begin{aligned} &Var\Big(pl[K]^{\dagger}_{unif_{\mathcal{N}_{fs}[L](X)}}(\varphi)\Big) \\ &= \frac{(L-N)\cdot(L+K)}{(K+N+1)\cdot(K+N)^2} \cdot \Big(-\varphi^2 + (K+N-2)\cdot\varphi + (K+N-1)\cdot\mathbf{1}\Big). \end{aligned}$$

Proof. 1 Using Lemma 7.1.4 (1) and Exercise 1.8.10 (1),

$$\begin{aligned} pl[K]^{\dagger}_{unif_{\mathcal{N}_{fs}[L](X)}}(\varphi) &= \sum_{\upsilon \in \mathcal{N}_{fs}[L](X)} \frac{pl[K](\upsilon)(\varphi)}{\sum_{\nu \in \mathcal{N}_{fs}[L](X)} pl[K](\nu)(\varphi)} \big|\upsilon\rangle \\ &= \sum_{\upsilon \in \mathcal{N}_{fs}[L](X)} \frac{\binom{\upsilon}{\varphi}/\binom{L}{K}}{\sum_{\nu \in \mathcal{N}_{fs}[L](X)} \binom{\nu}{\varphi}/\binom{L}{K}} \big|\upsilon\rangle = \sum_{\upsilon \in \mathcal{N}_{fs}[L](X)} \frac{\binom{\upsilon}{\varphi}}{\binom{K+N}{L-N}} \big|\upsilon\rangle. \end{aligned}$$

2 We now use Exercise 1.8.10 (2):

$$\begin{aligned} Mean\Big(pl[K]^{\dagger}_{unif_{\mathcal{N}_{fs}[L](X)}}(\varphi)\Big) &= \sum_{x \in X} \sum_{\upsilon \in \mathcal{N}_{fs}[L](X)} \frac{\binom{\upsilon}{\varphi}}{\binom{K+N}{L-N}} \cdot \upsilon(x) \big|x\rangle \\ &= \sum_{x \in X} \frac{1}{\binom{K+N}{L-N}} \cdot \left[\binom{L-N}{K+N} \cdot \varphi(x) + \binom{L+K}{L-N}\right] \big|x\rangle \\ &= \sum_{x \in X} \frac{(L-N)!\cdot(K+N-1)!}{(L+K-1)!} \\ &\qquad \cdot \left[\frac{(L+K-1)!}{(L-N-1)!\cdot(K+N)!} \cdot \varphi(x) + \frac{(L+K)!}{(L-N)!\cdot(K+N)!}\right] \big|x\rangle \\ &= \sum_{x \in X} \frac{L-N}{K+N} \cdot \varphi(x) + \frac{K+N}{L-N} \big|x\rangle = \frac{L-N}{K+N} \cdot \varphi + \frac{L+K}{K+N} \cdot \mathbf{1}. \end{aligned}$$

3 See Exercise 7.6.6. ❑

Exercises

7.6.1 In Lemma 4.2.4 and Exercise 5.1.12 we have seen that the mean and variance functions $Mean, Var\colon \mathcal{D}(\mathbb{R}) \to \mathbb{R}$ preserves (convolution) sums. Doublecheck that the same preservation property holds for the functions $Mean, Var\colon \mathcal{D}(\mathcal{M}(X)) \to \mathcal{M}(X)$ that we use in this section.

7.6.2 Prove, in the context of Proposition 7.6.1 (1), that pusing its dagger distribution along a binomial channel yields a convolution of the following form, for $s \in [0, 1]$.

$$bn[-](s)_*\Big(bn[-](r)^\dagger_{pois[\lambda]}(k)\Big) \;=\; bn[k](s) + pois[s \cdot (1-r) \cdot \lambda].$$

7.6.3 Check that the sizes of the mean-multisets in the second items of Propositions 7.6.2 – 7.6.6 are all equal to L.

7.6.4 Check, either by calculating the dagger explicitly, or by using Proposition 7.6.2:

$$hg[3]^\dagger_{unif_{\mathcal{N}[5](\{a,b\})}}(1|a\rangle + 2|b\rangle) \;=\; \tfrac{1}{5}\big|3|a\rangle + 2|b\rangle\big\rangle + \tfrac{2}{5}\big|2|a\rangle + 3|b\rangle\big\rangle + \tfrac{2}{5}\big|1|a\rangle + 4|b\rangle\big\rangle.$$

Compute also the posterior predictive distribution:

$$\begin{aligned} &hg[3]_*\Big(hg[3]^\dagger_{unif_{\mathcal{N}[5](\{a,b\})}}(1|a\rangle + 2|b\rangle)\Big) \\ &= \tfrac{1}{50}\big|3|a\rangle\big\rangle + \tfrac{6}{25}\big|2|a\rangle + 1|b\rangle\big\rangle + \tfrac{27}{50}\big|1|a\rangle + 2|b\rangle\big\rangle + \tfrac{1}{5}\big|3|b\rangle\big\rangle. \end{aligned}$$

Notice that the original draw $\varphi = 1|a\rangle + 2|b\rangle$ gets the highest probability.

7.6.5 The aim of this exercise is to prove the (complicated) variance formula in Proposition 7.6.2 (3). So let Let X be a finite set with $N := |X| > 0$ elements and let $\varphi \in \mathcal{N}[K](X)$ be an arbitrary draw of size $K \geq 0$.

1 Prove the following variation on Exercise 1.8.9 (2), using Exercise 1.8.8 (2): for $x \in supp(\psi)$,

$$\sum_{\psi \in \mathcal{N}[L](X)} \psi(x) \cdot (\psi(x) - 1) \cdot \binom{\varphi + \psi}{\psi} \;=\; \left(\!\!\binom{L-1}{K+N+1}\!\!\right) \cdot (\varphi(x) + 1) \cdot (\varphi(x) + 2).$$

2 Write $a^2 = a(a-1) + a$ (like in Exercise 3.3.8) to prove:

$$\begin{aligned} \sum_{\psi \in \mathcal{N}[L](X)} \psi(x)^2 \cdot \binom{\varphi + \psi}{\psi} \;&=\; \left(\!\!\binom{L-1}{K+N+1}\!\!\right) \\ &\cdot \left(\varphi(x)^2 \;+\; \frac{3 \cdot L+K+N-2}{L-1} \cdot \varphi(x) \;+\; \frac{2 \cdot L+K+N-1}{L-1}\right). \end{aligned}$$

3 Now recall the dagger $hg[K]^\dagger_{unif_{\mathcal{N}[L](X)}}(\varphi)$ from Proposition 7.6.2 (1). Check that it can be written as convolution:

$$hg[K]^\dagger_{unif_{\mathcal{N}[L](X)}}(\varphi) \;=\; 1\big|\varphi\big\rangle + \nu \quad \text{for} \quad \nu \;=\; \sum_{\psi \in \mathcal{N}[L-K](X)} \frac{\binom{\varphi+\psi}{\varphi}}{\binom{L+N-1}{K+N-1}} \big|\psi\big\rangle.$$

Show also that:

$$Mean(\nu) \;=\; \frac{L-K}{K+N} \cdot (\varphi + \mathbf{1}) \qquad \text{and} \qquad Var(\nu) \;=\; Var\Big(hg[K]^\dagger_{unif_{\mathcal{N}[L](X)}}(\varphi)\Big).$$

4 Now prove Proposition 7.6.2 (3) by showing that:

$$Var(\nu) \;=\; \frac{(L-K) \cdot (L+N)}{(K+N+1) \cdot (K+N)^2} \cdot \Big(-\varphi^2 + (K+N-2) \cdot \varphi + (K+N-1) \cdot \mathbf{1}\Big).$$

(Be careful to use $L-K$ in place of L in the formula in the earlier item (2).)

7.6.6 This exercise provides the essential ingredients for a proof of Proposition 7.6.6 (3). Let X be a finite set of size $N > 0$ and let $L > N$. We fix a draw $\varphi \in \mathcal{N}[K](X)$ of size $K \geq 0$.

1 Use Exercise 7.6.5 (1) to show that:

$$\sum_{\psi \in \mathcal{N}_{fs}[L](X)} (\psi(x) - 1) \cdot (\psi(x) - 2) \cdot \left(\!\binom{\psi}{\varphi}\!\right) = \left(\!\binom{L-N-1}{K+N+1}\!\right) \cdot (\varphi(x) + 1) \cdot (\varphi(x) + 2).$$

2 Deduce that:

$$\sum_{\psi \in \mathcal{N}_{fs}[L](X)} \psi(x)^2 \cdot \left(\!\binom{\psi}{\varphi}\!\right) = \left(\!\binom{K+N}{L-N}\!\right) \cdot \left(\frac{(L-N) \cdot (L-N-1)}{(K+N) \cdot (K+N+1)} \cdot \varphi(x)^2 + \frac{3 \cdot (L-N) \cdot (L+K)}{(K+N) \cdot (K+N+1)} \cdot \varphi(x) + \frac{(L+K) \cdot (2 \cdot L+K-N+1)}{(K+N) \cdot (K+N+1)} \right).$$

3 Use these results to prove the variance equation formulated in Proposition 7.6.6 (3).

7.6.7 This exercise is about an extension of the particle detection illustration, from the beginning of this section, to particles of different kinds, say from a set X (of kinds). In this situation there is an appearance rate factor $\Lambda \colon X \to \mathbb{R}_{\geq 0}$ and a detection rate predicate $p \colon X \to [0, 1]$, both giving rates per particle. We use the parallel Poisson distribution $ppois[\Lambda]$ and the parallel binomial $pabn[-](p)$, both from Section 4.7.

1 Show that the dagger of the parallel binomial with a parallel Poisson distribution as prior takes the form of a shifted parallel Poisson, for an obsered multiset φ (of particles):

$$pabn[-](p)^{\dagger}_{ppois[\Lambda]}(\varphi) = 1|\varphi\rangle + ppois[p^{\perp} \mathbin{\&} \Lambda].$$

Hint: Recall Lemma 4.7.5.

2 Prove that the mean and variance of this dagger are:

$$\begin{aligned} Mean\Big(pabn[-](p)^{\dagger}_{ppois[\Lambda]}(\varphi)\Big) &= \varphi + (p^{\perp} \mathbin{\&} \Lambda) \\ Var\Big(pabn[-](p)^{\dagger}_{ppois[\Lambda]}(\varphi)\Big) &= p^{\perp} \mathbin{\&} \Lambda. \end{aligned}$$

7.7 Sufficient statistics

It sometimes happens in an update $c(x)|_q$, for a channel c, that the input x drops out, in the sense that the updated distribution $c(x)|_q$ no longer depends on x. This phenomenon lies at the heart of the concept of sufficient statistics, introduced and studied in the 1920s by Ronald Fisher, see Fisher (1922). It is the topic of the current section. We have already encountered several examples of a sufficient statistics situation, without making this fact explicit. Only now we have the tools — updating, daggers and disintegration — to describe this sufficiency at the appropriate level of abstraction, following the string-diagrammatic approach of Fritz (2020). Various examples are given below, based partly on Jacobs (2023b, 2022a). Continuous probability also forms a rich source of examples, see *e.g.* Example **??**.

Let's first look at an example where the input of a channel drops out. We use the iid channel $iid[K] \colon \mathcal{D}(X) \rightsquigarrow X^K$, given by $iid[K] = \omega^K = \omega \otimes \cdots \otimes \omega$. We fix a multiset $\varphi \in \mathcal{N}[K](X)$ and

start calculating the update (dagger):

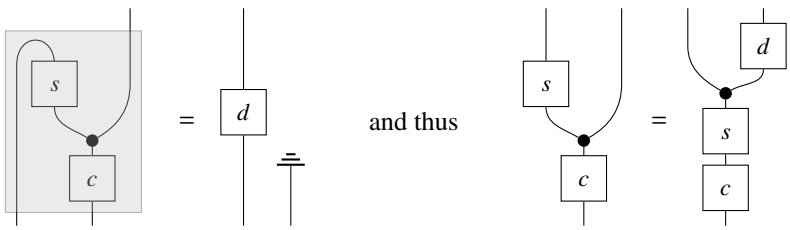
$$
\begin{aligned}
iid[K](\omega)\big|_{acc^*(\mathbf{1}_\varphi)} &= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{\omega^K(\vec{x})}{\mathcal{D}(acc)(\omega^K)(\varphi)}\,\big|\vec{x}\rangle && \text{by Lemma 7.1.4 (2)} \\
&= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{\omega^K(\vec{x})}{mn[K](\omega)(\varphi)}\,\big|\vec{x}\rangle && \text{by Theorem 2.6.7} \\
&= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{\omega^K(\vec{x})}{(\varphi)\cdot\prod_{x\in X}\omega(x)^{\varphi(x)}}\,\big|\vec{x}\rangle \\
&= \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{(\varphi)}\,\big|\vec{x}\rangle && \text{since } acc(\vec{x}) = \varphi \\
&\overset{(2.38)}{=} arr(\varphi).
\end{aligned}
$$

In the latter expression the distribution ω has disappeared. This depends on the fact that sequences $\vec{x}, \vec{y} \in X^K$ with $acc(\vec{x}) = \varphi = acc(\vec{y})$ have the same probability $\omega^K(\vec{x}) = \prod_{z\in X}\omega(z)^{\varphi(z)} = \omega^K(\vec{y})$, see Theorem 3.3.1 (1). Thus, it 'suffices' to know φ: if we condition $iid[K](\omega) = \omega^K$ to those sequences that accumulate to φ, the outcome no longer depends on ω.

The axiomatisation of this phenomenon uses (surjective) functions $s\colon X \to Y$, not channels, to identify elements in X, resulting in abstractions in Y. This Y may give numerical information, *e.g.* when Y is the set of natural or real numbers. Such a map is also called a *statistic*. A statistic s is called *sufficient* for a channel, as statistical model depending on a parameter / input, when the dependence on the parameter disappears via updating through s, as illustrated above. Sufficency refers to being adequate as a summary of essential aspects of the elements in X. Sufficiency involves the existence of a reversal of the *function* $s\colon X \to Y$ to a *channel* $Y \rightsquigarrow X$. This channel d can be used to reconstruct a distribution X, form its summary given by s. We first give a string diagrammatic description, following (Fritz, 2020, Defn. 14.3). Subsequently we explain the relation to disintegration.

Definition 7.7.1 Let $c\colon A \to X$ be a channel, where we think of A as the space of parameters.

1 A *statistic* for the channel c is a function $s\colon X \to Y$.
2 Such a statistic s is *sufficient* if there is a channel $d\colon Y \rightsquigarrow X$ such that:



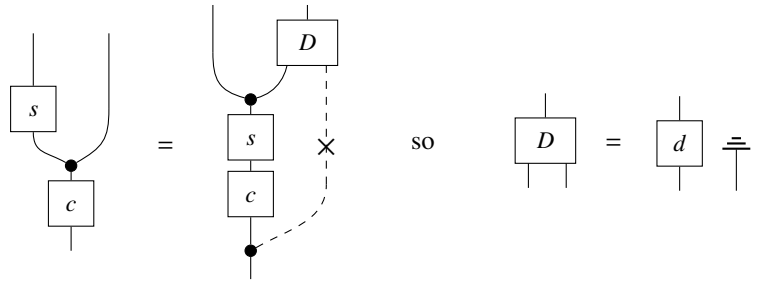
(7.36)

This means that $d = s^{\dagger}_{c(a)}$, for each $a \in A$.

We shall soon look at examples, and non-examples too, but the shape on the right-hand-side of (7.36) may look familiar already since it occurred at several earlier stages in this book.

Remark 7.7.2 1 The channel d on the right-hand-side in (7.36) results in a special way from disintegration, namely disintegration of the composite channel $\langle s, id\rangle \circ c$ on the left-hand-side in (7.36). If we write $D\colon Y\times A \rightsquigarrow X$ for its disintegration channel, as in (7.18), then, we should

get an equation as given on the left below. The distinguishing property of a *sufficient* statistic is that the dashed line is absent.



(7.37)

The situation on the left can be expressed by the equation on the right.

2 The fact that the dashed wire is missing in (7.37) corresponds to the phenomenon that we started with in this section, namely that the dependence on an input may disappear after conditioning. This can now be made more precise. The box / channel D in (7.37) can be calculated — in discrete probability — as a distribution $D(a, y) \in \mathcal{D}(X)$, for $a \in A$ and $y \in Y$, via a dagger (see Subsection 4.4.3). Explicitly, via Lemma 7.1.4 (2), since the statistic s is a function, not a channel:

$$D(a, y) \;=\; s^{\dagger}_{c(a)}(y) \;=\; c(a)|_{s^{*}(\mathbf{1}_y)} \;=\; \sum_{x \in s^{-1}(y)} \frac{c(a)(x)}{\mathcal{D}(s)(c(a))(y)} \,|x\rangle. \tag{7.38}$$

The absence of the dashed arrow in (7.37) corresponds to the non-dependence of the latter expressions on a. This is the essence of the Fisher-Neyman factorisation theorem, see (Fritz, 2020, Thm. 14.5) and (Bernardo and Smith, 2000, Prop 4.10) or (Suhov and Kelbert, 2005, §3.3). Thus, what we can do to find out if a function s is a sufficient statistic for a channel c is to compute the right-hand-side of 7.38 and check if the parameter a drops out.

Examples 7.7.3 1 We have already seen several sufficient statistics situations:

- The bitsum function $sum \colon \mathbf{2}^K \to \{0, \ldots, K\}$ is a sufficient statistic for the K-ary tuple channel $\langle flip, \ldots, flip\rangle \colon [0, 1] \rightsquigarrow \mathbf{2}^K$, see Theorem 2.5.2. A related situation occurs in Exercise 7.7.1 below.
- The function $size \colon \mathcal{N}(X) \to \mathbb{N}$ is a sufficient statistic for the multiset Poisson channel $mpois$ from Example 2.1.8 (2), see Exercise 4.5.4.
- The accumulation function $acc \colon X^K \to \mathcal{N}[K](X)$ is a sufficient statistic for the K-fold tensor channel $iid[K] \colon \mathcal{D}(X) \rightsquigarrow X^K$, see Theorem 3.3.1. Accumulation is also a sufficient statistic for the Poisson-iid channel, see Theorem 3.10.7.
- The sum of two multisets $sum \colon \mathcal{N}[K](X) \times \mathcal{N}[L](X) \to \mathcal{N}[K+L](X)$ is a sufficient statistic for the tuple channel $\langle mn[K], mn[L]\rangle \colon \mathcal{D}(X) \rightsquigarrow \mathcal{N}[K](X) \times \mathcal{N}[L](X)$ of two multinomials, see Theorem 3.5.4.
- Decoupling $dcpl = \langle \mathcal{N}(\pi_1), \mathcal{N}(\pi_2)\rangle \colon \mathcal{N}[K](X \times Y) \to \mathcal{N}[K](X) \times \mathcal{N}[K](Y)$ is a sufficient statistic for the channel $mn[K] \circ \otimes \colon \mathcal{D}(X) \times \mathcal{D}(Y) \rightsquigarrow \mathcal{N}[K](X \times Y)$, see Theorem 3.3.5.

2 As listed above, accumulation $acc \colon X^K \to \mathcal{N}[K](X)$ is a sufficient statistic for $iid[K]$. What about matching $mat \colon X^K \to SP(K)$, see Definition 1.5.8? It turns out *not* to be sufficient. We

compute the relevant dagger (7.38), for a distribution $\omega \in \mathcal{D}(X)$ and $P \in SP(K)$,

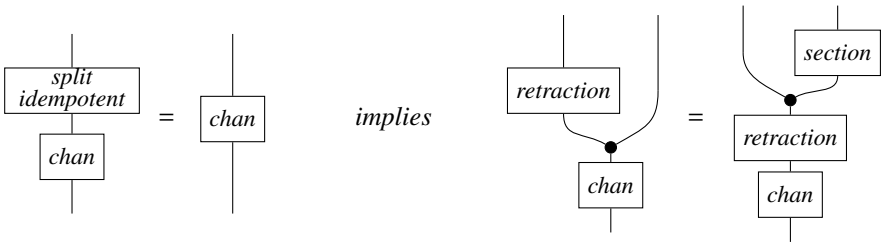
$$mat^{\dagger}_{iid[K](\omega)}(P) = \sum_{\vec{x} \in mat^{-1}(P)} \frac{\omega^K(\vec{x})}{\mathcal{D}(mat)(\omega^K)(P)} \left|\vec{x}\right\rangle = \sum_{\vec{x} \in mat^{-1}(P)} \frac{\omega^K(\vec{x})}{\sum_{\vec{y} \in mat(P)} \omega^K(\vec{y})} \left|\vec{x}\right\rangle.$$

What we would now need is a result saying: $mat(\vec{x}) = mat(\vec{y})$ implies $\omega^K(\vec{x}) = \omega^K(\vec{y})$. But this does not hold. Take for instance $\omega = \frac{1}{2}|a\rangle + \frac{1}{3}|b\rangle + \frac{1}{6}|c\rangle$ and $K = 2$. Then sequence (a, b) and (a, c) satisfy $mat(a, b) = \{\{1\}, \{2\}\} = mat(a, c)$, but $\omega^2(a, b) = \frac{1}{2} \cdot \frac{1}{3} \neq \frac{1}{2} \cdot \frac{1}{6} = \omega^2(a, c)$.

In Remark 3.3.2 we saw a similar argument showing that the sum of the elements of a multisets (*e.g.* over $\mathbb{N}$) is not a sufficient statistic for multinomial distributions.

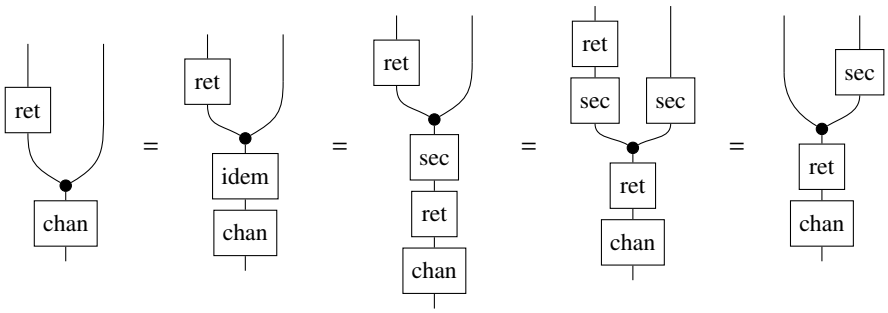
We include a formulation of the Fisher-Neyman factorisation theorem (see *e.g.* (Bernardo and Smith, 2000, Prop 4.10) or (Suhov and Kelbert, 2005, §3.3)) in diagrammatic form in (Fritz, 2020, Thm. 14.5) and in (Jacobs, 2023b, Lem. 3.4). It uses that the category **Chan** of probabilistic channels is a 'positive' Markov category, justifying the last equation below, see (Fritz, 2020, Rem.11.23).

Lemma 7.7.4 *Consider a channel together with a split idempotent in the category* **Chan**, *consisting of a deterministic section (a function) followed by a retraction. This retraction can then be a sufficient statistic for the channel:*



A typical example of a retraction is a composite $f^{\dagger}_{\omega} \circ f$, for a function f, see Lemma 7.1.4 (2).

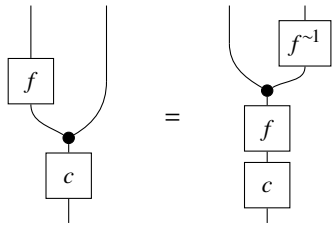
Proof. When we abbreviate 'retraction' to 'ret', 'section' to 'sec' and 'split idempotent' to 'idem'. The result now follows from the following sequence of equations between string diagrams.



Notice that we use that the section is a function, and thus commutes with the copier, see Remark 2.5.4. ❑

The following result can be seen as a special case of Proposition 7.7.4, involving the probabilistic inverse $f^{\sim 1}$ from Definition 2.4.7, given by $f^{\sim 1}(y) = \sum_{x \in f^{-1}(y)} \frac{1}{|f^{-1}(y)|}|x\rangle$. We recall that we write $\ker(f) = \{(x, x') \mid f(x) = f(x')\} \subseteq X \times X$ for the kernel relation, for a function f.

Proposition 7.7.5 *Let $c\colon A \to X$ be a channel and $f\colon X \to Y$ a surjective function with finite inverse images, satisfying $\ker(f) \subseteq \ker(c(a))$, for each $a \in A$, then f is a sufficient statistic, with its probabilistic inverse $f^{\sim 1}\colon Y \rightsquigarrow X$ as associated channel, in a situation:*



Proof. For each $a \in A$ we have:

$$\begin{aligned}
\langle f, id\rangle_*(c(a)) &= \sum_{x\in X}\sum_{y\in Y} \langle f\rangle(x)(y)\cdot c(a)(x)\,\big|y,x\rangle \\
&= \sum_{y\in Y}\sum_{x\in f^{-1}(y)} c(a)(x)\,\big|y,x\rangle \\
&= \sum_{y\in Y}\sum_{x\in f^{-1}(y)}\sum_{z\in f^{-1}(y)} \frac{c(a)(z)}{|f^{-1}(y)|}\,\big|y,x\rangle \qquad \text{since } \ker(f) \subseteq \ker(c(a)) \\
&= \sum_{x\in X}\sum_{y\in Y} f^{\sim 1}(y)(x)\cdot \mathcal{D}(f)(c(a))(y)\,\big|y,x\rangle \\
&= \langle id, f^{\sim 1}\rangle_*((f\circ c)(a)).
\end{aligned}$$

❑

Theorem 3.3.1 is an instance of this result, but more instance follow below.

7.7.1 Ewens distributions

There is a rich theory of distributions on multiset and set partitions $MP(K)$ and $SP(K)$, as introduced in Section 1.9. This theory involves mutations, via a mutation parameter $\theta \in \mathbb{R}_{\geq 0}$, and are used for instance in population biology, see Ewens (1972); Joyce (1998); Kingman (1978a,b); Crane (2016). A central role is played by so-called Ewens distributions on multiset partitions, see Jacobs (2022b); Blanchi and Paquet (2024) for recent references. These distributions give rise to various sufficient statistic situations, especially using size. We present several examples below and refer to Jacobs (2022a) for a more complete account.

We start by recalling the draw-add channel $DA\colon \mathcal{N}[K](X) \to \mathcal{D}(\mathcal{N}[K{+}1](X))$ from Section 3.2 and introduce analogous operations on multiset partitions and set partitions in a (commuting) diagram, for $K \geq 1$. They involve a mutation rate $\theta \in \mathbb{R}_{\geq 0}$.

$$\begin{tikzcd}
MP(K) \arrow[r, "MDA(\theta)"] & \mathcal{D}(MP(K+1)) \\
SP(K) \arrow[u, "sc"] \arrow[r, "SDA(\theta)"] & \mathcal{D}(SP(K+1)) \arrow[u, "\mathcal{D}(sc)"']
\end{tikzcd} \tag{7.39}$$

We define draw-add channels $MDA(\theta)$ and $SDA(\theta)$ on a multiset partitions $\alpha \in MP(K)$ and a set

partition $P \in SP(K)$ as:

$$\begin{aligned} MDA(\theta)(\alpha) &:= \frac{\theta}{\theta+K}\Big|\alpha + 1|1\rangle\Big\rangle + \sum_{1\leq i\leq K} \frac{\alpha(i)\cdot i}{\theta+K}\Big|\alpha - 1|i\rangle + 1|i+1\rangle\Big\rangle \\ SDA(\theta)(P) &:= \frac{\theta}{\theta+K}\Big|P \cup \{\{K+1\}\}\Big\rangle + \sum_{B\in P} \frac{|B|}{\theta+K}\Big|(P\backslash B) \cup \{B \cup \{K+1\}\}\Big\rangle. \end{aligned} \tag{7.40}$$

It is easy to see that these definitions yield appropriately typed distributions, on $MP(K+1)$ and on $SP(K+1)$. There are also corresponding draw-delete channels, see Exercise 7.7.6 below.

Illustrations may help to see what happens: for the multiset partition $\alpha = 1|1\rangle + 2|3\rangle + 1|6\rangle \in MP(13)$ we get:

$$\begin{aligned} MDA[1](\alpha) &= \tfrac{1}{14}\Big|2|1\rangle + 2|3\rangle + 1|6\rangle\Big\rangle + \tfrac{1}{14}\Big|1|2\rangle + 2|3\rangle + 1|6\rangle\Big\rangle \\ &\quad + \tfrac{3}{7}\Big|1|1\rangle + 1|3\rangle + 1|4\rangle + 1|6\rangle\Big\rangle + \tfrac{3}{7}\Big|1|1\rangle + 2|3\rangle + 1|7\rangle\Big\rangle. \end{aligned}$$

And for the set partition $P = \Big\{\{1\}, \{2,3\}, \{4,5,6\}\Big\} \in SP(7)$ one gets:

$$\begin{aligned} SDA[1](P) &= \tfrac{1}{7}\Big|\Big\{\{1\}, \{2,3\}, \{4,5,6\}, \{7\}\Big\}\Big\rangle + \tfrac{1}{7}\Big|\Big\{\{1,7\}, \{2,3\}, \{4,5,6\}\Big\}\Big\rangle \\ &\quad + \tfrac{2}{7}\Big|\Big\{\{1\}, \{2,3,7\}, \{4,5,6\}\Big\}\Big\rangle + \tfrac{3}{7}\Big|\Big\{\{1\}, \{2,3\}, \{4,5,6,7\}\Big\}\Big\rangle. \end{aligned}$$

Lemma 7.7.6 *Diagram (7.39) commutes, that is, the draw-add channels MDA for multiset partitions and SDA for set partitions commute with size-count.*

Proof. By unpacking the definitions from (7.39):

$$\begin{aligned} (\mathcal{D}(sc) \circ SDA(\theta))(P) &= \frac{\theta}{\theta+K}\Big|sc\Big(P \cup \{\{K+1\}\}\Big)\Big\rangle + \sum_{B\in P} \frac{|B|}{\theta+K}\Big|sc\Big((P\backslash B) \cup \{B \cup \{K+1\}\}\Big)\Big\rangle \\ &= \frac{\theta}{\theta+K}\Big|sc(P) + 1|1\rangle\Big\rangle + \sum_{B\in P} \frac{|B|}{\theta+K}\Big|sc(P) - 1\Big||B|\rangle + 1\Big||B|+1\rangle\Big\rangle \\ &= \frac{\theta}{\theta+K}\Big|sc(P) + 1|1\rangle\Big\rangle + \sum_{1\leq i\leq K} \frac{sc(P)(i)\cdot i}{\theta+K}\Big|sc(P) - 1\Big|i\rangle + 1\Big|i+1\rangle\Big\rangle \\ &= MDA(\theta)(sc(P)) = (MDA(\theta) \circ sc)(P). \end{aligned}$$

❑

Warren Ewens originally Ewens (1972) defined the distributions that are named after him on multiset partitions. As we shall see, there are similar distributions on set partitions, which we call Bell distributions, after Eric Temple Bell, whose name is used for the Bell numbers, that describe the sizes of the sets $SP(K)$ of set partitions, of the set $\{1, \ldots, K\}$.

Definition 7.7.7 Fix $K \in \mathbb{N}_{>0}$ and $\theta \in \mathbb{R}_{\geq 0}$.

1 The *Ewens distribution* on multiset partitions $ew[K](\theta) \in \mathcal{D}(MP(K))$ is defined inductively as:

$$ew[1](\theta) := 1\Big|1|1\rangle\Big\rangle \qquad ew[K+1](\theta) := MDA(\theta)_*(ew[K](\theta)).$$

2 Similarly one defines the *Bell distribution* on set partitions $bell[K](\theta) \in \mathcal{D}(SP(K))$ as:

$$bell[1](\theta) := 1\Big|\Big\{\{1\}\Big\}\Big\rangle \qquad bell[K+1](\theta) := SDA(\theta)_*(bell[K](\theta)).$$

Since size-count commutes with the draw-adds in (7.39) and preserves the base case, we immediately get:

$$\mathcal{D}(sc)\Big(bell[K](\theta)\Big) = ew[K](\theta). \tag{7.41}$$

There are single formulas for these inductively defined distributions. The proofs are a bit of work.

Proposition 7.7.8 *Let $K \in \mathbb{N}_{>0}$ and $\theta \in \mathbb{R}_{>0}$*

1 The Ewens distribution $\mathrm{ew}[K](\theta) \in \mathcal{D}(MP(K))$ is given by:

$$\mathrm{ew}[K](\theta) = \frac{1}{(\theta)^K} \cdot \sum_{\alpha \in MP(K)} \frac{\theta^{\|\alpha\|} \cdot K!}{\alpha\!\!\!\!\circ\!\!\!\!\circ \cdot prod(\alpha)} \big|\alpha\rangle, \tag{7.42}$$

where $(\theta)^K = \prod_{i<K} \theta + i$ is the rising factorial, $\alpha\!\!\!\!\circ\!\!\!\!\circ = \prod_i \alpha(i)!$ and $\mathrm{prod}(\alpha) = \prod_i i^{\alpha(i)}$. This is Ewens' sampling formula, expressed as convex formal sum.

2 The Bell distribution $\mathrm{bell}[K](\theta) \in \mathcal{D}(SP(K))$ is:

$$bell[K](\theta) = \frac{1}{(\theta)^K} \cdot \sum_{P \in SP(K)} \theta^{|P|} \cdot \prod_{B \in P} (|B|-1)! \,\big|P\rangle. \tag{7.43}$$

For these formulations we require that the mutation rate θ is non-zero. See Exercise 7.7.2 for the case $\theta = 0$.

Proof. 1 The induction case $K = 1$ is simple, since:

$$\frac{1}{\prod_{i<1} \theta + i} \cdot \sum_{\alpha \in MP(1)} \frac{\theta^{\|\alpha\|} \cdot 1!}{\alpha\!\!\!\!\circ\!\!\!\!\circ \cdot prod(\alpha)} \big|\alpha\rangle = \frac{1}{\theta} \cdot \frac{\theta^{\|1|1\rangle\|}}{1|1\rangle\!\!\!\!\circ\!\!\!\!\circ \cdot prod(1|1\rangle)} \Big|1|1\rangle\Big\rangle = 1\Big|1|1\rangle\Big\rangle$$
$$= ew[1](\theta).$$

The induction step is more work:

$$
\begin{aligned}
ew[K+1](\theta) &= MDA(\theta)_*(ew[K](\theta)) \\
&= \sum_{\beta\in MP(K+1)} \sum_{\alpha\in MP(K)} MDA(\theta)(\alpha)(\beta)\cdot ew[K](\theta)(\alpha)\,\big|\beta\rangle \\
&\overset{(7.40)}{=} \sum_{\alpha\in MP(K)} \frac{\theta}{\theta+K}\cdot ew[K](\theta)(\alpha)\,\Big|\alpha+1|1\rangle\Big\rangle \\
&\quad + \sum_{\alpha\in MP(K)} \sum_{1\le i\le K} \frac{\alpha(i)\cdot i}{\theta+K}\cdot ew[K](\theta)(\alpha)\,\Big|\alpha-1|i\rangle+1|i+1\rangle\Big\rangle \\
&\overset{(IH)}{=} \frac{1}{\prod_{i<K+1}\theta+i}\cdot\Bigg(\sum_{\alpha\in MP(K)} \theta\cdot\frac{\theta^{\|\alpha\|}\cdot K!}{\alpha!\cdot prod(\alpha)}\,\Big|\alpha+1|1\rangle\Big\rangle \\
&\quad + \sum_{\alpha\in MP(K)} \sum_{1\le i\le K} \alpha(i)\cdot i\cdot\frac{\theta^{\|\alpha\|}\cdot K!}{\alpha!\cdot prod(\alpha)}\,\Big|\alpha-1|i\rangle+1|i+1\rangle\Big\rangle\Bigg) \\
&= \frac{1}{(\theta)^{K+1}}\cdot\Bigg(\sum_{\alpha\in MP(K)} \frac{(\alpha(1)+1)\cdot\theta^{\|\alpha+1|1\rangle\|}\cdot K!}{(\alpha+1|1\rangle)!\cdot prod(\alpha+1|1\rangle)}\,\Big|\alpha+1|1\rangle\Big\rangle \\
&\quad + \sum_{\alpha\in MP(K)} \sum_{1\le i\le K} (\alpha(i+1)+1)\cdot(i+1)\cdot \\
&\quad \frac{\theta^{\|\alpha-1|i\rangle+1|i+1\rangle\|}\cdot K!}{(\alpha-1|i\rangle+1|i+1\rangle)!\cdot prod(\alpha-1|i\rangle+1|i+1\rangle)}\,\Big|\alpha-1|i\rangle+1|i+1\rangle\Big\rangle\Bigg) \\
&= \frac{1}{(\theta)^{K+1}}\cdot\sum_{\beta\in MP(K+1)}\left(\beta(1)+\sum_{1\le i\le K}\beta(i+1)\cdot(i+1)\right)\cdot\frac{\theta^{\|\beta\|}\cdot K!}{\beta!\cdot prod(\beta)}\,|\beta\rangle \\
&= \frac{1}{(\theta)^{K+1}}\cdot\sum_{\beta\in MP(K+1)}\frac{\theta^{\|\beta\|}\cdot(K+1)!}{\beta!\cdot prod(\beta)}\,|\beta\rangle.
\end{aligned}
$$

2 For $K=1$ we have:

$$
\begin{aligned}
\frac{1}{\prod_{i<1}\theta+i}\cdot\sum_{P\in SP(1)}\theta^{|P|}\cdot\prod_{B\in P}(|B|-1)!\,\big|P\rangle &= \frac{1}{\theta}\cdot\theta^{|\{1\}|}\cdot\prod_{B\in\{\{1\}\}}(|B|-1)!\,\Big|\big\{\{1\}\big\}\Big\rangle = 1\,\Big|\big\{\{1\}\big\}\Big\rangle \\
&= bell[1](\theta).
\end{aligned}
$$

The induction step works as follows.

$$
\begin{aligned}
bell[K+1](\theta) &= SDA(\theta)_*(bell[K](\theta)) \\
&= \sum_{Q\in SP(K+1)} \sum_{P\in SP(K)} SDA(\theta)(P)(Q)\cdot bell[K](\theta)(P)\,\big|Q\rangle \\
&\overset{(7.40)}{=} \sum_{P\in SP(K)} \frac{\theta}{\theta+K}\cdot bell[K](\theta)(P)\,\Big|P\cup\{\{K+1\}\}\Big\rangle \\
&\qquad + \sum_{P\in SP(K)} \sum_{A\in P} \frac{|A|}{\theta+K}\cdot bell[K](\theta)(P)\,\Big|(P\backslash A)\cup\{A\cup\{K+1\}\}\Big\rangle \\
&\overset{(\mathrm{IH})}{=} \frac{1}{\prod_{i<K+1}\theta+i}\cdot\Bigg(\sum_{P\in SP(K)} \theta\cdot\theta^{|P|}\cdot\prod_{B\in P}(|B|-1)!\,\Big|P\cup\{\{K+1\}\}\Big\rangle \\
&\qquad + \sum_{P\in SP(K)} \sum_{A\in P} |A|\cdot\theta^{|P|}\cdot\prod_{B\in P}(|B|-1)!\,\Big|(P\backslash A)\cup\{A\cup\{K+1\}\}\Big\rangle\Bigg) \\
&= \frac{1}{(\theta)^{K+1}}\cdot\Bigg(\sum_{P\in SP(K)} \theta^{|P\cup\{\{K+1\}\}|}\cdot\prod_{B\in P\cup\{\{K+1\}\}}(|B|-1)!\,\Big|P\cup\{\{K+1\}\}\Big\rangle \\
&\qquad + \sum_{P\in SP(K)} \sum_{A\in P} \theta^{|(P\backslash A)\cup\{A\cup\{K+1\}\}|}\cdot \\
&\qquad\qquad \prod_{B\in(P\backslash A)\cup\{A\cup\{K+1\}\}}(|B|-1)!\,\Big|(P\backslash A)\cup\{A\cup\{K+1\}\}\Big\rangle\Bigg) \\
&= \frac{1}{(\theta)^{K+1}}\cdot\sum_{Q\in SP(K+1)} \theta^{|Q|}\cdot\prod_{B\in Q}(|B|-1)!\,\big|Q\rangle. \qquad \square
\end{aligned}
$$

We illustrate these two Ewens distributions:

$$
\begin{aligned}
ew[4](\tfrac{2}{3}) &= \tfrac{1}{55}\Big|4|1\rangle\Big\rangle + \tfrac{9}{55}\Big|2|1\rangle + 1|2\rangle\Big\rangle + \tfrac{27}{220}\Big|2|2\rangle\Big\rangle \\
&\quad + \tfrac{18}{55}\Big|1|1\rangle + 1|3\rangle\Big\rangle + \tfrac{81}{220}\Big|1|4\rangle\Big\rangle \\
bell[3](\tfrac{1}{2}) &= \tfrac{1}{15}\Big|\big\{\{1\},\{2\},\{3\}\big\}\Big\rangle + \tfrac{2}{15}\Big|\big\{\{1,2\},\{3\}\big\}\Big\rangle + \tfrac{2}{15}\Big|\big\{\{1,3\},\{2\}\big\}\Big\rangle \\
&\quad + \tfrac{2}{15}\Big|\big\{\{1\},\{2,3\}\big\}\Big\rangle + \tfrac{8}{15}\Big|\big\{\{1,2,3\}\big\}\Big\rangle.
\end{aligned}
$$

Our aim is to show that both size functions, of type $MP(K) \to (K]$ and $SP(K) \to (K]$, for $(K] = \{1,\ldots,K\}$, are a sufficient statistic, for the Ewens and Bell distributions. We first establish how these size functions act on these two distributions (see also (Tavaré, 2021, §2.2)).

Lemma 7.7.9 *We use the following ad hoc notations to distinguish the multiset-partition-size and set-partition-size functions.*

$$msize := \big\|-\big\| : MP(K) \longrightarrow (K] \qquad ssize := \big|-\big| : SP(K) \longrightarrow (K].$$

These size functions send both the Ewens and the Bell distribution to a new distribution that we call the Stiring distribution:

$$\mathcal{D}(msize)\big(ew[K](\theta)\big) = stir[K](\theta) = \mathcal{D}(ssize)\big(bell[K](\theta)\big),$$

where $stir[K](\theta) \in \mathcal{D}((K])$ *is the 'Stirling' distribution that is defined in terms of Stirling numbers*

of the first kind:

$$stir[K](\theta) \coloneqq \frac{1}{(\theta)^K} \cdot \sum_{1 \leq n \leq K} \begin{bmatrix} K \\ n \end{bmatrix} \cdot \theta^n \left| n \right\rangle. \tag{7.44}$$

The probabilities add up to one by Lemma 1.2.4 (6).

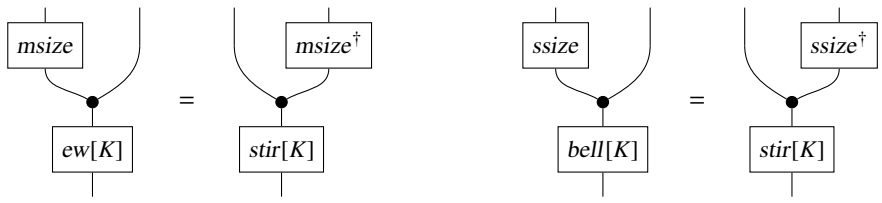
Proof. We prove the first equation:

$$\begin{aligned}
\mathcal{D}(msize)\Big(ew[K](\theta)\Big) &= \frac{1}{(\theta)^K} \cdot \sum_{\alpha \in MP(K)} \frac{\theta^{\|\alpha\|} \cdot K!}{\alpha\llbracket \cdot prod(\alpha)} \left| msize(\alpha) \right\rangle \\
&= \frac{1}{(\theta)^K} \cdot \sum_{1 \leq n \leq K} \sum_{\alpha \in MP(K),\, \|\alpha\| = n} \frac{\theta^n \cdot K!}{\alpha\llbracket \cdot prod(\alpha)} \left| n \right\rangle \\
&= \frac{1}{(\theta)^K} \cdot \sum_{1 \leq n \leq K} \begin{bmatrix} K \\ n \end{bmatrix} \cdot \theta^n \left| n \right\rangle, \qquad \text{by Exercise 1.9.5.}
\end{aligned}$$

The second equation is obtained similarly via Exercise 1.5.8. ❑

We now come to the main result, stating that size is a sufficient statistic for Ewens distributions. The multiset partition version goes back to Ewens (1972); the set partition version is an analogue, that may be found in Jacobs (2022a).

Theorem 7.7.10 *For a number $K \in \mathbb{N}_{>0}$ and mutatation rate $\theta \in \mathbb{R}_{>0}$ there are sufficient statistic situations:*



The dagger channels $msize^\dagger : (K] \rightsquigarrow MP(K)$ and $ssize^\dagger : (K] \rightsquigarrow SP(K)$ involved are:

$$\begin{aligned}
msize^\dagger(n) &\coloneqq \frac{1}{\left[{K \atop n}\right]} \cdot \sum_{\alpha \in msize^{-1}(n)} \frac{K!}{\alpha\llbracket \cdot prod(\alpha)} \left| \alpha \right\rangle \\
ssize^\dagger(n) &\coloneqq \frac{1}{\left[{K \atop n}\right]} \cdot \sum_{P \in ssize^{-1}(n)} \prod_{B \in P} (|B| - 1)! \left| P \right\rangle.
\end{aligned}$$

These channels are well-defined by Exercises 1.9.5 and 1.5.8.

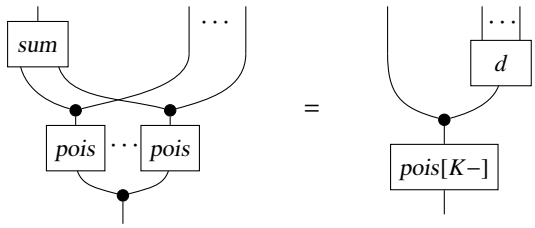
Proof. We do the first equation and leave the second one to the interested reader.

$$\begin{aligned}
\langle id, msize \rangle_*(ew[K](\theta)) &= \sum_{\alpha \in MP(K),\, n \in (K]} \langle msize \rangle(\alpha)(n) \cdot ew[K](\theta)(\alpha) \left| n, \alpha \right\rangle \\
&\overset{(7.42)}{=} \sum_{n \in (K]} \sum_{\alpha \in msize^{-1}(n)} \frac{1}{(\theta)^K} \cdot \frac{\theta^{\|\alpha\|} \cdot K!}{\alpha\llbracket \cdot prod(\alpha)} \left| n, \alpha \right\rangle \\
&= \sum_{n \in (K]} \sum_{\alpha \in msize^{-1}(n)} msize^\dagger(n)(\alpha) \cdot \frac{1}{(\theta)^K} \cdot \begin{bmatrix} K \\ n \end{bmatrix} \cdot \theta^n \left| n, \alpha \right\rangle \\
&\overset{(7.44)}{=} \sum_{\alpha \in MP(K),\, n \in (K]} msize^\dagger(n)(\alpha) \cdot stir[K](\theta)(n) \left| n, \alpha \right\rangle \\
&= \langle msize^\dagger, id \rangle_*(stir[K](\theta)).
\end{aligned}$$

❑

Exercises

7.7.1 Prove that the K-fold sum $sum\colon \mathbb{N}^K \to \mathbb{N}$ is a sufficient statistic for the K-tuple of Poisson channels $\langle pois, \ldots, pois\rangle\colon \mathbb{R}_{>0} \rightsquigarrow \mathbb{N}^K$ in a situation:



Define yourself what this channel $d\colon \mathbb{N} \rightsquigarrow \mathbb{N}^K$ does.
Hint: Use Exercise 1.7.9.

7.7.2 In Proposition 7.7.8 we require that the mutation rate θ is non-zero. One can allow $\theta = 0$ but then one gets different descriptions for the Ewens and Bell distributions.

1 Prove the two equations below by induction on K, using the formulations of Definition 7.7.7.

$$ew[K](0) = 1\Big|\, 1|K\rangle\Big\rangle \qquad bell[K](0) = 1\Big|\big\{\{1, \ldots, K\}\big\}\Big\rangle.$$

2 Let's define the Stirling distribution with zero mutation rate as: $stir[K](0) = 1|1\rangle \in \mathcal{D}(\{1, \ldots, K\})$. Check that the sufficient statistic equations between string diagrams in Theorem 7.7.10 also hold for input $\theta = 0$.

7.7.3 Check that the summation results in Theorem 1.9.9 for multiset partitions, and also the special case (for $r = 1$) in Proposition 1.5.10 for set partitions, are consequences of Proposition 7.7.8.

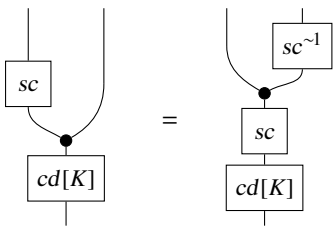
7.7.4 Recall the size count function $sc\colon SP(K) \to MP(K)$, see Definition 1.9.2 (2), given by $sc(P) = \sum_{B\in P} 1\big||B|\rangle$.

1 Show that for a distribution $\omega \in \mathcal{D}(X)$,

$$\ker(sc) \subseteq \ker(cd[K](\omega)).$$

where $cd[K](\omega) = \mathcal{D}(mat)(\omega^K)$ is the coincidence distribution from Definition 2.3.7 (1).

2 Show also that this size count function $sc\colon SP(K) \to MP(K)$ is a sufficient statistic for the coincidence channel $cd[K]\colon \mathcal{D}(X) \rightsquigarrow SP(K)$, in a situation:



where the probabilistic inverse channel $sc^{\sim 1}\colon MP(K) \rightsquigarrow SP(K)$ is, see Lemma 1.9.7,

$$sc^{\sim 1}(\alpha) = \sum_{P\in sc^{-1}(\alpha)} \frac{\alpha\,\underset{\circ}{!}}{(\alpha)_p}\,|P\rangle.$$

7.7.5 Check that the (analogue of the) first point in the previous exercise does not hold for multiplicity count *mc* from natural multisets to multiset partitions, see Definition 1.9.2 (1), and the multinomial distribution, that is, in general:

$$mc(\varphi) = mc(\psi) \not\Longrightarrow mn[K](\omega)(\varphi) = mn[K](\omega)(\psi).$$

7.7.6 In (7.40) we have seen draw-add analogues for set / multiset partitions. There are also draw-delete channels $MDD\colon MP(K{+}1) \rightsquigarrow MP(K)$ and $SDD\colon SP(K{+}1) \rightsquigarrow SP(K)$, without mutation rate parameter. They are defined in Jacobs (2022b,a) as:

$$\begin{aligned} MDD(\theta)(\beta) &:= \frac{\beta(1)}{K+1}\Big|\beta - 1|1\rangle\Big\rangle + \sum_{2\leq i\leq K+1} \frac{\beta(i)\cdot i}{K+1}\Big|\beta + 1|i{-}1\rangle - 1|i\rangle\Big\rangle. \\ SDD(\theta)(Q) &:= \sum_{1\leq i\leq K+1} \frac{1}{K+1}\Big|\{\{j \mid j\in B, j<i\}\cup\{j{-}1 \mid j\in B, j>i\} \;\Big|\; B\in Q, B\neq\{i\}\}\Big\rangle. \end{aligned}$$

1 Compute the outcomes:

$$MDD\Big(1|1\rangle + 2|3\rangle + 1|6\rangle\Big) \qquad\qquad SDD\Big(\Big\{\{1\},\{2,3\},\{4,5,6\}\Big\}\Big).$$

2 Show that the following diagram commutes.

$$\begin{tikzcd} \mathcal{D}(MP(K)) & MP(K+1) \arrow[l, "MDD"'] \\ \mathcal{D}(SP(K)) \arrow[u, "\mathcal{D}(sc)"] & SP(K+1) \arrow[l, "SDD"'] \arrow[u, "sc"'] \end{tikzcd}$$

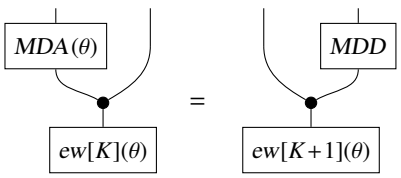
3 Prove that the Ewens and Bell distributions form draw-delete cones, like in Proposition 3.2.9:

$$SDD_*(bell[K{+}1](\theta)) = bell[K](\theta) \qquad\qquad MDD_*(ew[K{+}1](\theta)) = ew[K](\theta).$$

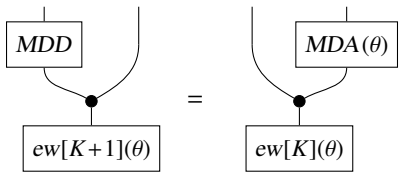
The corresponding 'inverse' properties, for draw-add's, hold by construction, see Definition 7.7.7. This cone property is an essential ingredient of a De Finetti result for Ewens distributions, see Blanchi and Paquet (2024).

4 (See (Jacobs, 2022b, Thm. 12)) Show that — unlike the ordinary draw-add and draw-delete channels, see Exercise 7.1.6 — the multiset partition versions are each other's daggers, as described below, as equations between terms on the left and between string diagrams on the right.

$$MDA(\theta)^{\dagger}_{ew[K](\theta)} = MDD$$



$$MDD^{\dagger}_{ew[K+1](\theta)} = MDA(\theta)$$



These dagger equations do not hold for the set partition versions *SDA* and *SDD*. A

counterexample is obtained easily, for instance with set partition $P = \big\{\{1\}, \{2\}, \big\}$, $K = 2$ and $\theta = 1$.

7.8 Factorisation of joint states

Earlier, in Subsection 2.3.2, we have called a binary joint distribution non-entwined when it is the product of its marginals. This can be seen as an intrinsic property of the distribution, which we will now express in terms of a string diagram, called its *shape*. This topic is often described in terms of d-separation, see *e.g.* Pearl (2009); Fritz and Klingler (2023).

For instance, the joint distribution:

$$\omega = \tfrac{1}{4}|a, b\rangle + \tfrac{1}{2}|a, b^\perp\rangle + \tfrac{1}{12}|a^\perp, b\rangle + \tfrac{1}{6}|a^\perp, b^\perp\rangle$$

is non-entwined: it is the product of its marginals $\omega[1, 0] = \frac{3}{4}|a\rangle + \frac{1}{4}|a^\perp\rangle$ and $\omega[0, 1] = \frac{1}{3}|b\rangle + \frac{2}{3}|b^\perp\rangle$. We will formulate this as:

ω has *shape*
or as: and write this as: $\omega \models\!\approx$.
ω *factorises* as

Definition 7.8.1 Let ω be distribution and S a string diagram. We write $\sigma \models\!\approx S$ when there is an interpretation of the boxes in S as channels that produces ω as composite. We then say that ω has shape S or also: ω factorises as S.

This same formulation $c \models\!\approx S$ can be used for a channel c.

In the above case of $\omega \models\!\approx$ we obtain $\omega = \omega_1 \otimes \omega_2$, for some state ω_1 that interpretes the box on the left, and some ω_2 interpreting the box on the right. But then:

$$\omega[1, 0] = (\omega_1 \otimes \omega_2)[1, 0] = \omega_1.$$

Similarly, $\omega[0, 1] = \omega_2$. Thus, in this case the interpretations of the boxes are uniquely determined, namely as first and second marginal of ω.

We conclude that non-entwinedness of an arbitrary binary joint state ω can be expressed as: $\omega \models\!\approx$. Here we are interested in similar intrinsic 'shape' properties of states and channels that can be expressed via string diagrams. These matters are often discussed in the literature in terms of (conditional) independencies. Here we prefer to use shapes instead of independencies since they are (visually) more expressive. For instance, a channel $c\colon X \twoheadrightarrow Y \times Z$ is called *conditionally independent* when:

$c \models\!\approx$

In general there may be several interpretations of a string diagram (as shape). Consider for instance the statement:

$\frac{14}{25}|H\rangle + \frac{11}{25}|T\rangle \models\!\approx$

The distribution on the left-hand-side matches the shape on the right-hand-side in multiple ways, for instance as:

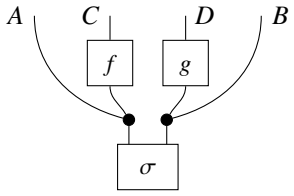
$$(c_1)_*(\sigma_1) = \tfrac{14}{25}|H\rangle + \tfrac{11}{25}|T\rangle = (c_2)_*(\sigma_2)$$

for:

$$\begin{array}{rcl} \sigma_1 &=& \frac{1}{5}|1\rangle + \frac{4}{5}|0\rangle \\ c_1(1) &=& \frac{4}{5}|H\rangle + \frac{1}{5}|T\rangle \\ c_1(0) &=& \frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle \end{array} \qquad \text{and} \qquad \begin{array}{rcl} \sigma_2 &=& \frac{2}{5}|1\rangle + \frac{3}{5}|0\rangle \\ c_2(1) &=& \frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle \\ c_2(0) &=& \frac{3}{5}|H\rangle + \frac{2}{5}|T\rangle. \end{array}$$

We note that is not an 'accessible' string diagram: the wire inbetween the two boxes cannot be accessed from the outside. If these wires are accessible, then we can access the individual boxes of a string diagram and use disintegration to compute them. We illustrate how this works.

Example 7.8.2 Consider two-element sets $A = \{a, a^\perp\}$, $B = \{b, b^\perp\}$, $C = \{c, c^\perp\}$ and $D = \{d, d^\perp\}$ and an (accessible) string diagram S of the form:

$S =$ 

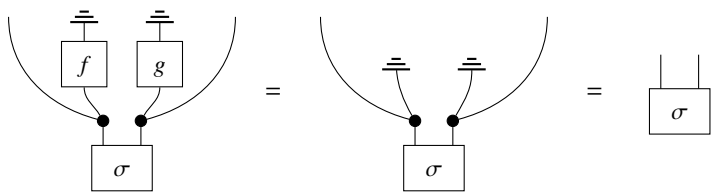
(7.45)

Now suppose we have a joint distribution $\omega \in \mathcal{D}(A \times C \times D \times B)$ given by:

$$\begin{array}{rl} \omega = & \frac{1}{25}|a,c,d,b\rangle + \frac{9}{50}|a,c,d,b^\perp\rangle + \frac{3}{50}|a,c,d^\perp,b\rangle + \frac{1}{50}|a,c,d^\perp,b^\perp\rangle + \frac{1}{25}|a,c^\perp,d,b\rangle + \\ & + \frac{9}{50}|a,c^\perp,d,b^\perp\rangle + \frac{3}{50}|a,c^\perp,d^\perp,b\rangle + \frac{1}{50}|a,c^\perp,d^\perp,b^\perp\rangle + \frac{3}{125}|a^\perp,c,d,b\rangle \\ & + \frac{9}{500}|a^\perp,c,d,b^\perp\rangle + \frac{9}{250}|a^\perp,c,d^\perp,b\rangle + \frac{1}{500}|a^\perp,c,d^\perp,b^\perp\rangle + \frac{12}{125}|a^\perp,c^\perp,d,b\rangle \\ & + \frac{9}{125}|a^\perp,c^\perp,d,b^\perp\rangle + \frac{18}{125}|a^\perp,c^\perp,d^\perp,b\rangle + \frac{1}{125}|a^\perp,c^\perp,d^\perp,b^\perp\rangle \end{array}$$

We ask ourselves: does $\omega \models\!\approx S$ hold? More specifically, can we somehow obtain interpretations of the boxes σ, f and g in S so that ω emerges? We shall show that by appropriately using marginalisation and disintegration we can 'factorise' this joint state according to the above string diagram.

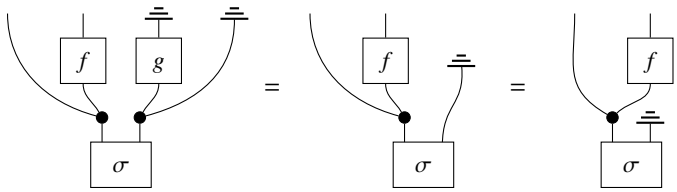
First we can obtain the state $\sigma \in \mathcal{D}(A \times B)$ by discarding the C, D outputs in the middle, as in:



We can thus compute σ as:

$$\begin{array}{rcl} \sigma = \omega[1,0,0,1] &=& \left(\sum_{u\in C,\, v\in D} \omega(a,u,v,b)\right)|a,b\rangle + \left(\sum_{u\in C,\, v\in D} \omega(a,u,v,b^\perp)\right)|a,b^\perp\rangle \\ && + \left(\sum_{u\in C,\, v\in D} \omega(a^\perp,u,v,b)\right)|a^\perp,b\rangle + \left(\sum_{u\in C,\, v\in D} \omega(a^\perp,u,v,b^\perp)\right)|a^\perp,b^\perp\rangle \\ &=& (\frac{1}{25} + \frac{3}{50} + \frac{1}{25} + \frac{3}{50})|a,b\rangle + (\frac{9}{50} + \frac{1}{50} + \frac{9}{50} + \frac{1}{50})|a,b^\perp\rangle \\ && + (\frac{3}{125} + \frac{9}{250} + \frac{12}{125} + \frac{18}{125})|a^\perp,b\rangle + (\frac{9}{500} + \frac{1}{500} + \frac{9}{125} + \frac{1}{125})|a^\perp,b^\perp\rangle \\ &=& \frac{1}{5}|a,b\rangle + \frac{2}{5}|a,b^\perp\rangle + \frac{3}{10}|a^\perp,b\rangle + \frac{1}{10}|a^\perp,b^\perp\rangle. \end{array}$$

Next we concentrate on the channels f and g, from A to C and from B to D. We first illustrate how to restrict the string diagram to the relevant part via marginalisation. For f we concentrate on:



The string diagram on the right tells us that we can obtain f via disintegration from the marginal $\omega[1,1,0,0]$, using that extracted channels are unique, in diagrams of this form, see (7.11). In the same way one obtains g from the marginal $\omega[0,0,1,1]$. Thus:

$$
\begin{aligned}
f \;=\; \omega[0,1,0,0 \,|\, 1,0,0,0] &= \begin{cases} a \mapsto \dfrac{\sum_{v,y} \omega(a,c,v,y)}{\sum_{u,v,y} \omega(a,u,v,y)}|c\rangle + \dfrac{\sum_{v,y} \omega(a,c^\perp,v,y)}{\sum_{u,v,y} \omega(a,u,v,y)}|c^\perp\rangle \\ a^\perp \mapsto \dfrac{\sum_{v,y} \omega(a^\perp,c,v,y)}{\sum_{u,v,y} \omega(a^\perp,u,v,y)}|c\rangle + \dfrac{\sum_{v,y} \omega(a^\perp,c^\perp,v,y)}{\sum_{u,v,y} \omega(a^\perp,u,v,y)}|c^\perp\rangle \end{cases} \\
&= \begin{cases} a \mapsto \frac{1}{2}|c\rangle + \frac{1}{2}|c^\perp\rangle \\ a^\perp \mapsto \frac{1}{5}|c\rangle + \frac{4}{5}|c^\perp\rangle \end{cases} \\
g \;=\; \omega[0,0,1,0 \,|\, 0,0,0,1] &= \begin{cases} b \mapsto \dfrac{\sum_{x,u} \omega(x,u,d,b)}{\sum_{x,u,v} \omega(x,u,v,b)}|d\rangle + \dfrac{\sum_{x,u} \omega(x,u,d^\perp,b)}{\sum_{x,u,v} \omega(x,u,v,b)}|d^\perp\rangle \\ b^\perp \mapsto \dfrac{\sum_{x,u} \omega(x,u,d,b^\perp)}{\sum_{x,u,v} \omega(x,u,v,b^\perp)}|d\rangle + \dfrac{\sum_{x,v} \omega(x,u,d^\perp,b^\perp)}{\sum_{x,u,v} \omega(x,u,v,b^\perp)}|d^\perp\rangle \end{cases} \\
&= \begin{cases} b \mapsto \frac{2}{5}|d\rangle + \frac{3}{5}|d^\perp\rangle \\ b^\perp \mapsto \frac{9}{10}|d\rangle + \frac{1}{10}|d^\perp\rangle. \end{cases}
\end{aligned}
$$

At this stage one can check that the joint distribution ω can be reconstructed from these extracted state and channels, namely as:

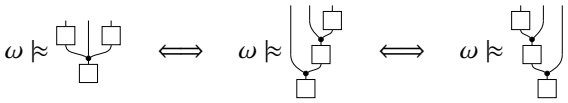
$$\omega \;=\; (id \otimes f \otimes g \otimes id)_*\big((\Delta \otimes \Delta)_*(\sigma)\big) \;=\; \big(\langle id, f\rangle \otimes \langle g, id\rangle\big)_*(\sigma).$$

This proves $\omega \models\!\approx S$.

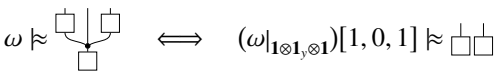
The following illustration is a classical one, showing how seemingly different shapes are related. It is often used to describe conditional independence — in this case of A, C, given B.

Theorem 7.8.3 *Let $\omega \in \mathcal{D}(X \times Y \times Z)$ be a distribution with full support.*

1 There are equivalences:



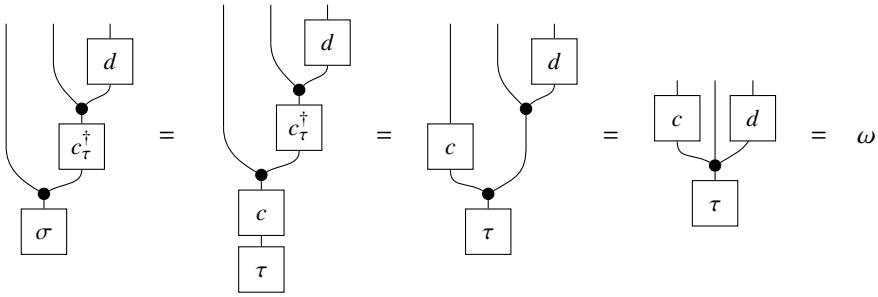
2 These equivalences can be extended as:



for all $y \in Y$.

The string diagrams on the left in item (1) is often called a fork; the other two are called a chain. In item (2) this shape is related to non-entwinedness, pointwise.

Proof. 1 Since ω has full support, so have all its marginals, see Exercise 2.3.1. This allows us to perform all disintegrations below. We start on the left-hand side, and assume an interpretation $\omega = \langle c, id, d\rangle_*(\tau)$, consisting of a state $\tau = \omega[0,1,0] \in \mathcal{D}(Y)$ and channels $c\colon Y \to X$ and $d\colon Y \to Z$. We write $\sigma = c_*(\tau) = \omega[1,0,0]$ and take the Bayesian inversion $c^\dagger_\tau\colon X \to Y$. We now have an interpretation of the string diagram in the middle, which is equal to ω, since by (7.3):



Similarly one obtains an interpretation of the string diagram on the right via $\rho = d_*(\tau) = \omega[0,0,1]$ and the inversion $d^\dagger_\tau\colon Z \to Y$.

In the direction ($\Leftarrow$) one uses Bayesian inversion in a similar manner to transform one interpretation into another one.

2 The direction ($\Rightarrow$) is easy and is left to the reader. In fact, as we shall see below, Proposition 7.8.4 (1) gives a slightly stronger result.

We concentrate on ($\Leftarrow$). By assumption, for $y \in Y$ we can write:

$$(\omega|_{\mathbf{1}\otimes\mathbf{1}_y\otimes\mathbf{1}})[1,0,1] = \sigma_y \otimes \tau_y \quad \text{for } \sigma_y \in \mathcal{D}(X),\ \tau_y \in \mathcal{D}(Z).$$

In a next step we define channels $f\colon Y \to X$ and $g\colon Y \to Z$ and a distribution $\rho \in \mathcal{D}(Y)$ as:

$$f(y) := \sigma_y \qquad g(y) := \tau_y \qquad \rho := \omega[0,1,0].$$

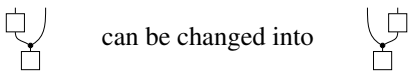
Then, for $x \in X$ and $z \in Z$,

$$\begin{aligned} f(y)(x)\cdot g(y)(z) &= (f(y)\otimes g(y))(x,z) = (\sigma_y\otimes\tau_y)(x,z) = (\omega|_{\mathbf{1}\otimes\mathbf{1}_y\otimes\mathbf{1}})[1,0,1](x,z) \\ &= \sum\nolimits_v \frac{\omega(x,v,z)\cdot \mathbf{1}_y(v)}{\omega \models \mathbf{1}\otimes\mathbf{1}_y\otimes\mathbf{1}} \\ &= \frac{\omega(x,y,z)}{\omega[0,1,0]\models \mathbf{1}_y} = \frac{\omega(x,y,z)}{\rho(y)}. \end{aligned}$$

Now we see that ω has a fork shape:

$$\omega(x,y,z) = f(y)(x)\cdot g(y)(z)\cdot\rho(y) = \langle f, id, g\rangle_*(\rho)(x,y,z). \qquad \square$$

What the first item of this result shows is that (sub)shapes of the form:



and vice-versa. We knew this already from Theorem 7.2.1. By applying these transformations directly to the shapes in Theorem 7.8.3 the equivalences $\Leftrightarrow$ can be obtained.

We have seen that a distribution can have a certain shape. An interesting question that arises is: what happens to such a shape when the distribution is updated? The result below answers this

question, much like in Jacobs and Zanasi (2017), for three basic shapes, called fork, chain and collider,

Proposition 7.8.4 *Let $\omega \in \mathcal{D}(X \times Y \times Z)$ be an arbitrary distribution and let $q \in \mathrm{Fact}(Y)$ be a factor on its middle component Y. We write $a \in Y$ for an arbitrary element with associated point predicate $\mathbf{1}_a$.*

1 *Let ω have* fork *shape:*

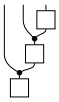
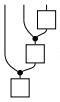
$\omega \approx$ *then also* $\omega|_{\mathbf{1}\otimes q\otimes \mathbf{1}} \approx$.

In the special case of conditioning with a point predicate we get:

$\omega|_{\mathbf{1}\otimes \mathbf{1}_a\otimes \mathbf{1}} \approx$.

where the middle box is $1|a\rangle$.

2 *If ω has a* chain *shape:*

$\omega \approx$  *then also* $\omega|_{\mathbf{1}\otimes q\otimes \mathbf{1}} \approx$  .

3 *Let ω have* collider *shape:*

$\omega \approx$ *then again* $\omega|_{\mathbf{1}\otimes q\otimes \mathbf{1}} \approx$

For this shape it does not matter if q is a point predicate or not.

Proof. 1 Let's assume we have an interpretation $\omega = \langle c, id, d\rangle_*(\sigma)$, for $c\colon Y \twoheadrightarrow X$, $d\colon Y \twoheadrightarrow Z$ and $\sigma \in \mathcal{D}(Y)$. Then:

$$\begin{aligned}
\omega|_{\mathbf{1}\otimes q\otimes \mathbf{1}} &= \langle c, id, d\rangle_*(\sigma)|_{\mathbf{1}\otimes q\otimes \mathbf{1}} \\
&= (\langle c, id, d\rangle|_{\mathbf{1}\otimes q\otimes \mathbf{1}})_*(\sigma|_{\langle c,id,d\rangle^*(\mathbf{1}\otimes q\otimes \mathbf{1})}) && \text{by Corollary 6.3.5 (2)} \\
&= \langle c, id, d\rangle_*(\sigma|_q) && \text{by Exercises 6.1.17 and Lemma 4.5.2 (7).}
\end{aligned}$$

Hence we see the same shape that ω has.

In the special case when $q = \mathbf{1}_a$ we can extend the above calculation and obtain a parallel product of distributions:

$$\begin{aligned}
\omega|_{\mathbf{1}\otimes \mathbf{1}_a\otimes \mathbf{1}} &= \langle c, id, d\rangle_*(\sigma|_{\mathbf{1}_a}) && \text{as just shown} \\
&= ((c \otimes id \otimes d) \circ \Delta_3)_*(1|a\rangle) && \text{by Lemma 6.1.7 (2)} \\
&= (c \otimes id \otimes d)_*((\Delta_3)_*(1|a\rangle)) \\
&= (c \otimes id \otimes d)_*(1|a\rangle \otimes 1|a\rangle \otimes 1|a\rangle) \\
&= c_*(1|a\rangle) \otimes 1|a\rangle \otimes d_*(1|a\rangle) \;=\; c(a) \otimes 1|a\rangle \otimes d(a).
\end{aligned}$$

2 By the previous point and Theorem 7.8.3.

3 Let's assume as interpretation of the collider shape:

$$\omega = (id \otimes c \otimes id)_*((\Delta \otimes \Delta)_*(\sigma \otimes \tau)),$$

for states $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Z)$ and channel $c\colon X \times Z \twoheadrightarrow Y$. Then:

$$\begin{aligned}
\omega|_{\mathbf{1}\otimes q\otimes \mathbf{1}} &= (id \otimes c \otimes id)_*((\Delta \otimes \Delta)_*(\sigma \otimes \tau))\big|_{\mathbf{1}\otimes q\otimes \mathbf{1}} \\
&= (id \otimes c|_q \otimes id)_*\Big((\Delta \otimes \Delta)_*(\sigma \otimes \tau)\big|_{\mathbf{1}\otimes c^*(q)\otimes \mathbf{1}}\Big) && \text{by Corollary 6.3.5 (3)} \\
&= (id \otimes c|_q \otimes id)_*\Big(((\Delta \otimes \Delta)|_{\mathbf{1}\otimes c^*(q)\otimes \mathbf{1}})_*\Big((\sigma \otimes \tau)\big|_{(\Delta\otimes\Delta^*(\mathbf{1}\otimes c^*(q)\otimes \mathbf{1})}\Big)\Big) && \text{by Theorem 6.3.4} \\
&= (id \otimes c|_q \otimes id)_*\Big((\Delta \otimes \Delta)_*\Big((\sigma \otimes \tau)\big|_{(\Delta\otimes\Delta)^*(\mathbf{1}\otimes c^*(q)\otimes \mathbf{1})}\Big)\Big) && \text{by Exercise 6.1.16.}
\end{aligned}$$

The updated state $(\sigma \otimes \tau)\big|_{(\Delta\otimes\Delta)^*(\mathbf{1}\otimes c^*(q)\otimes \mathbf{1})}$ is typically entwined, even if q is a point predicate, see also Exercise 6.1.14. ❑

The fact that conditioning with a point predicate destroys the shape is an important phenomenon since it allows us to break entwinedness / correlations. This is relevant in statistical analysis, esp. w.r.t. causality Pearl (2009); Pearl and Mackenzie (2019), see the causal surgery procedure in Section 9.1. In such a context, conditioning on a point predicate is often expressed in terms of 'controlling for'. For instance, if there is a gender component $G = \{m, f\}$ with elements m for male and f for female, then conditioning with a point predicate $\mathbf{1}_m$ or $\mathbf{1}_f$, suitably weakenend via tensoring with truth $\mathbf{1}$, amounts to controlling for gender. Via restriction to one gender value one fragments the shape and thus controls the influence of gender in the situation at hand.

We conclude with an observation about entropy and KL-divergence in this setting. Earlier, in Lemma 2.10.8 (3) we have seen that the entropy of a joint (binary) distribution is less than the entropy of the product of its marginals. A similar results holds when we force a ternary distribution is forced to have a conditional indepedence shape. The shape that we use below is slightly different from the one in Theorem 7.8.3. The difference is inessential, the form used below makes the proof a bit easier. This leads to an extension of Proposition 2.10.9 (2).

Proposition 7.8.5 *Let $\omega \in \mathcal{D}(X\times Y\times Z)$ be a distribution. We use it construct a new distribution $\omega' \in \mathcal{D}(X \times Y \times Z)$ of the form $\omega' = \langle id, c, d\rangle_*(\sigma)$, with shape:*

$\omega' \models$ *via*

$$\begin{cases} \sigma := \omega[1,0,0] & \in \mathcal{D}(X) \\ c := \omega[0,1,0 \,|\, 1,0,0] & : X \twoheadrightarrow Y \\ d := \omega[0,0,1 \,|\, 1,0,0] & : X \twoheadrightarrow Z. \end{cases}$$

Then:

1 *$H(\omega) \leq H(\omega')$;*
2 *for all $\rho \in \mathcal{D}(X \times Y \times Z)$,*

$\rho \models$ $\implies D_{KL}(\omega, \rho) = D_{KL}(\omega, \omega') + D_{KL}(\omega', \rho)$.

This second point yields an inequality $D_{KL}(\omega, \omega') \leq D_{KL}(\omega, \rho)$ for ρ with the above conditional independence shape. This is an 'I-projection' property, see Csiszár (1975); Diaconis and Zabell (1982); Murphy (2023): it shows that ω' diverges least from ω among all distributions ρ with the above shape. That is, ω' is the projection of ω onto the subset of distributions in $\mathcal{D}(X \times Y \times Z)$ with this shape.

Proof. 1 We also extract a channel $e\colon \omega[0,1,1 \,|\, 1,0,0]\colon X \twoheadrightarrow Y \times Z$ so that $\omega = \langle id, e\rangle_*(\sigma)$.

We then have (pointwise) marginals $c = e[1,0]$ and $d = e[0,1]$. We then obtain the inequality $H(\omega) \leq H(\omega')$ via (pointwise) use of Lemma 2.10.8 (3).

$$\begin{aligned}
H(\omega) - H(\omega') &= -\sum_{x,y,z} \omega(x,y,z) \cdot \ln\big(\omega(x,y,z)\big) + \sum_{x,y,z} \omega'(x,y,z) \cdot \ln\big(\omega'(x,y,z)\big) \\
&= \sum_{x,y,z} \sigma(x) \cdot c(x)(y) \cdot d(x)(z) \cdot \ln\big(\sigma(x) \cdot c(x)(y) \cdot d(x)(z)\big) \\
&\qquad - \sum_{x,y,z} \sigma(x) \cdot e(y,z) \cdot \ln\big(\sigma(x) \cdot e(y,z)\big) \\
&= \sum_{x} \sigma(x) \cdot \Bigg[\sum_{y} c(x)(y) \cdot \ln\big(c(x)(y)\big) + \sum_{z} d(x)(z) \cdot \ln\big(d(x)(z)\big) \\
&\qquad - \sum_{y,z} e(x)(y,z) \cdot \ln\big(e(x)(y,z)\big)\Bigg] \\
&= \sum_{x} \sigma(x) \cdot \Big[H(e(x)) - H(c(x)) - H(d(x))\Big] \\
&= \sum_{x} \sigma(x) \cdot \Big[H(e(x)) - H(e(x)[1,0]) - H(e(x)[0,1])\Big] \\
&\leq 0, \qquad \text{since } H(e(x)) \leq H(e(x)[1,0]) + H(e(x)[0,1]).
\end{aligned}$$

2 Let $\rho = \langle id, f, g\rangle_*(\tau)$. Then:

$$\begin{aligned}
&D_{KL}(\omega,\rho) - D_{KL}(\omega,\omega') \\
&= \sum_{x,y,z} \omega(x,y,z) \cdot \ln\left(\frac{\omega(x,y,z)}{\tau(x) \cdot f(x)(y) \cdot g(x)(z)}\right) - \\
&\qquad \sum_{x,y,z} \omega(x,y,z) \cdot \ln\left(\frac{\omega(x,y,z)}{\sigma(x) \cdot c(x)(y) \cdot d(x)(z)}\right) \\
&= \sum_{x,y,z} \omega(x,y,z) \cdot \ln\left(\frac{\sigma(x) \cdot c(x)(y) \cdot d(x)(z)}{\tau(x) \cdot f(x)(y) \cdot g(x)(z)}\right) \\
&= \sum_{x} \left(\sum_{y,z} \omega(x,y,z)\right) \cdot \ln\left(\frac{\sigma(x)}{\tau(x)}\right) + \\
&\qquad \sum_{x,y} \left(\sum_{z} \omega(x,y,z)\right) \cdot \ln\left(\frac{c(x)(y)}{f(x)(y)}\right) + \sum_{x,z} \left(\sum_{y} \omega(x,y,z)\right) \cdot \ln\left(\frac{d(x)(z)}{g(x)(z)}\right) \\
&= \sum_{x} \sigma(x) \cdot \ln\left(\frac{\sigma(x)}{\tau(x)}\right) + \\
&\qquad \sum_{x,y} \sigma(x) \cdot c(x)(y) \cdot \ln\left(\frac{c(x)(y)}{f(x)(y)}\right) + \sum_{x,z} \sigma(x) \cdot d(x)(z) \cdot \ln\left(\frac{d(x)(z)}{g(x)(z)}\right) \\
&= \sum_{x,y,z} \sigma(x) \cdot c(x)(y) \cdot d(x)(z) \cdot \ln\left(\frac{\sigma(x) \cdot c(x)(y) \cdot d(x)(z)}{\tau(x) \cdot f(x)(y) \cdot g(x)(z)}\right) \\
&= D_{KL}(\omega',\rho).
\end{aligned}$$

❑

There is a general mechanism behind this result where we force a distribution to be of a certain shape, in a minimal manner. We include two more such results, of a simpler nature. Somewhat mysteriously, the minimality now works not in the first but in the second argument of KL-divergence.

Proposition 7.8.6 *Let $\omega \in \mathcal{D}(X \times Y)$ be a joint distribution, written in graph form as $\omega = \langle id, c\rangle_*(\omega_1)$, for first marginal $\omega_1 \in \mathcal{D}(X)$ and channel $c\colon X \rightsquigarrow Y$.*

Further, let distributions $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$ be given, with associated shapes, including the distributions σ and τ.

$S =$ $\qquad T =$

1 Write $\omega_S := \langle id, c\rangle_*(\sigma)$ so that $\omega_S \models\!\approx S$. Then for all $\rho \in \mathcal{D}(X \times Y)$,

$$\rho \models\!\approx S \implies D_{KL}(\rho, \omega) = D_{KL}(\rho, \omega_S) + D_{KL}(\omega_S, \omega).$$

2 We now write $\omega_T := \langle d, id\rangle_*(\tau)$ where $d := c^\dagger_{\omega_1}\colon Y \rightsquigarrow X$, so that $\omega_T \models\!\approx T$. Then for all $\rho \in \mathcal{D}(X \times Y)$,

$$\rho \models\!\approx T \implies D_{KL}(\rho, \omega) = D_{KL}(\rho, \omega_T) + D_{KL}(\omega_T, \omega).$$

Proof. 1 Since $\rho \models\!\approx S$ we can write $\rho = \langle id, e\rangle_*(\sigma)$ for an extracted channel $e = \rho[1, 0 \,|\, 0, 1]\colon Y \rightsquigarrow X$. Then:

$$\begin{aligned}
D_{KL}(\rho, \omega) - D_{KL}(\rho, \omega_S) &= \sum_{x,y} \rho(x, y) \cdot \ln\left(\frac{\omega_S(x, y)}{\omega(x, y)}\right) \\
&= \sum_{x,y} \sigma(x) \cdot e(x)(y) \cdot \ln\left(\frac{\sigma(x) \cdot c(x)(y)}{\omega_1(x) \cdot c(x)(y)}\right) \\
&= D_{KL}(\sigma, \omega_1) \\
&= \sum_{x,y} \sigma(x) \cdot c(x)(y) \cdot \ln\left(\frac{\sigma(x) \cdot c(x)(y)}{\omega_1(x) \cdot c(x)(y)}\right) = D_{KL}(\omega_S, \omega).
\end{aligned}$$

2 Since $\rho \models\!\approx T$ we can write $\rho = \langle e, id\rangle_*(\tau)$ for a channel $e = \rho[0, 1 \,|\, 1, 0]\colon X \rightsquigarrow Y$. Then:

$$\begin{aligned}
D_{KL}(\rho, \omega) - D_{KL}(\rho, \omega_T) &= \sum_{x,y} \rho(x, y) \cdot \ln\left(\frac{\omega_T(x, y)}{\omega(x, y)}\right) \\
&= \sum_{x,y} e(y)(x) \cdot \tau(y) \cdot \ln\left(\frac{\tau(y) \cdot c^\dagger_{\omega_1}(y)(x)}{\omega_1(x) \cdot c(x)(y)}\right) \\
&= \sum_{x,y} e(y)(x) \cdot \tau(y) \cdot \ln\left(\frac{\tau(y)}{c_*(\omega_1)}\right) \quad \text{by (7.1)} \\
&= D_{KL}(\tau, c_*(\omega_1)) \\
&= \sum_{x,y} c^\dagger_{\omega_1}(y)(x) \cdot \tau(y) \cdot \ln\left(\frac{\tau(y) \cdot c^\dagger_{\omega_1}(y)(x)}{\omega_1(x) \cdot c(x)(y)}\right) = D_{KL}(\omega_T, \omega). \quad \square
\end{aligned}$$

Exercises

7.8.1 The updates in Theorem 7.8.3 (2) involve point predicates $\mathbf{1}_b$. The result does not hold for arbitrary predicates. The aim of this exercise is to elaborate a counterexample. Consider the sets:

$$A = \{1, 2, 3\} \qquad B = \{b, b^\perp\} \qquad C = \{4, 5\}.$$

We use the distribution $\rho = \frac{1}{3}|b\rangle + \frac{2}{3}|b^\perp\rangle$ with channels $c\colon B \rightsquigarrow A$ and $d\colon B \rightsquigarrow C$ given by:

$$\begin{cases} c(b) = \frac{1}{3}|1\rangle + \frac{1}{3}|2\rangle + \frac{1}{3}|3\rangle \\ c(b^\perp) = \frac{1}{6}|1\rangle + \frac{1}{3}|2\rangle + \frac{1}{2}|3\rangle \end{cases} \qquad \begin{cases} d(b) = \frac{1}{4}|4\rangle + \frac{3}{4}|5\rangle \\ d(b^\perp) = \frac{1}{6}|4\rangle + \frac{5}{6}|5\rangle. \end{cases}$$

We then form the distribution ω of the shape used in Theorem 7.8.3 (2), where:

$$\omega = \langle c, id, d\rangle_*(\rho) \in \mathcal{D}(A \times B \times C).$$

1 Let $\omega' = (\omega|_{\mathbf{1}\otimes\mathbf{1}_b\otimes\mathbf{1}})[1,0,1] \in \mathcal{D}(B \times C)$ be obtained via an update with a point predicate $\mathbf{1}_b$, followed by a marginalisation. Check that the distribution ω' is non-entwined, following Theorem 7.8.3 (2), of the form:

$$\omega' = \omega'[1,0] \otimes \omega'[0,1] = \tfrac{1}{12}\big|2,5\big\rangle + \tfrac{1}{4}\big|2,6\big\rangle + \tfrac{1}{12}\big|3,5\big\rangle + \tfrac{1}{4}\big|3,6\big\rangle + \tfrac{1}{12}\big|4,5\big\rangle + \tfrac{1}{4}\big|4,6\big\rangle.$$

2 Now we take the predicate $p = \frac{5}{6} \cdot \mathbf{1}_b + \frac{1}{2} \cdot \mathbf{1}_{b^\perp}$ and write $\omega'' = (\omega|_{\mathbf{1}\otimes p\otimes\mathbf{1}})[1,0,1]$. Show that ω'' is entwined (*i.e.* not non-entwined). Explicitly:

$$\omega'' = \tfrac{7}{132}\big|2,5\big\rangle + \tfrac{25}{132}\big|2,6\big\rangle + \tfrac{3}{44}\big|3,5\big\rangle + \tfrac{35}{132}\big|3,6\big\rangle + \tfrac{1}{12}\big|4,5\big\rangle + \tfrac{15}{44}\big|4,6\big\rangle.$$

But:

$$\begin{aligned} \omega''[1,0] \otimes \omega''[0,1] = {} & \tfrac{6}{121}\big|2,5\big\rangle + \tfrac{70}{363}\big|2,6\big\rangle + \tfrac{3}{44}\big|3,5\big\rangle \\ & + \tfrac{35}{132}\big|3,6\big\rangle + \tfrac{21}{242}\big|4,5\big\rangle + \tfrac{245}{726}\big|4,6\big\rangle. \end{aligned}$$

7.8.2 Check the aim of Exercise 2.3.2 really is to prove the shape statement

$\omega \models\!\approx$

for the (ternary) state ω defined there.

7.8.3 Consider the joint Asia distribution $\alpha \in \mathcal{D}(S \times B \times L \times D \times E \times X \times T \times A)$ from Figure 6.10. Show that the B and A parts, for bronchitis and Asia are non-entwined, via purely diagrammatic reasoning with the rules for $\overset{=}{\top}$, as in:

$\alpha[0,1,0,0,0,0,0,1] \models\!\approx$

7.8.4 Prove that a collider shape correponds to non-entwinedness of the outer marginals:

$\omega \models\!\approx$ $\iff$ $\omega[1,0,1] \models\!\approx$

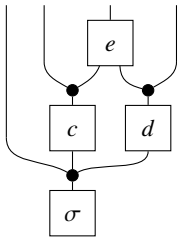
7.8.5 Reconsider the joint 'tooth decay' distribution $\tau \in \mathcal{D}(F \times S \times D)$ from Exercise 2.3.3. Extract a channel $e\colon F \times S \rightsquigarrow D$ from τ and use it to prove:

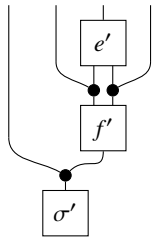
$\tau \models\!\approx$

Relate this shape statement to the previous Exercise and to the non-entwinedness in Exercise 2.3.3.

7.8.6 Recall Exercise 6.3.6 and notice that it describes a case where updating maintains the shape of a distribution. This exercise considers a case where the shape is distorted by

updating. Consider a joint distribution $\omega \in \mathcal{D}(X_1 \times X_2 \times X_3 \times X_4)$ with shape as on the left below.

$\omega \models\!\approx$ 

$\omega\big|_{p_1 \otimes p_2 \otimes p_3 \otimes p_4} \models\!\approx$ 

Use Theorem 6.3.4 and Corollary 6.3.5 to show specifically that:

$$\begin{aligned} \sigma' &= \sigma\big|_{p_1 \,\&\, \langle c,d\rangle^*(e^*(p_3) \,\&\, (p_2 \otimes p_4))} \\ f &= \langle c, d\rangle\big|_{e^*(p_3) \,\&\, (p_2 \otimes p_4)} \end{aligned} \qquad\qquad e' = e\big|_{p_3}.$$

7.8.7 This exercise extends the conditional independence characterisation from Theorem 7.8.3 with disintegration. Let $\omega \in \mathcal{D}(A \times B \times C)$ be a distribution with full support.

$\omega \models\!\approx$ $\iff \omega[1,0,1 \,|\, 0,1,0] \models\!\approx$ $\iff \omega[1,0,0 \,|\, 0,1,1] \models\!\approx$

Notice that the last shape statement says that conditional independence of A, C given B means that the extracted channel $B \times C \twoheadrightarrow A$ does not need the C input to compute the A output.

7.8.8 Prove the following items, which are known as the 'semigraphoid' properties, see Verma and Pearl (1988); Geiger et al. (1990); they are seen as the basic rules of conditional independence.

1 Symmetry: $\omega \models\!\approx$ implies $\langle \pi_3, \pi_2, \pi_1 \rangle_*(\omega) \models\!\approx$.

2 Decomposition: $\omega \models\!\approx$ implies $\omega[0,1,1,1] \models\!\approx$

3 Weak union: $\omega \models\!\approx$ implies $\omega \models\!\approx$

Hint: Apply disintegration to the upper-left box.

4 Contraction: if $\omega \models\!\approx$ and also $\omega[1,1,1,0] \models\!\approx$ then $\omega \models\!\approx$.

Hint: Form a suitable combination of the two upper-right boxes in the assumptions.

7.9 Pearl's and Jeffrey's update rules

In Chapter 6, and especially in Section 6.2, we have seen many demonstrations showing that backward inference is a useful reasoning technique. The general situation involves a prior distribution $\sigma \in \mathcal{D}(X)$, on the domain X of a channel $c\colon X \twoheadrightarrow Y$, that we wish to update in the light of evidence on the codomain Y of the channel. The evidence is given in the form of a predicate (or factor) on Y. This backward inference is also called Pearl's update rule.

It turns out that there is an alternative update mechanism in this situation: "Richard Jeffrey (references to Jeffrey (1983, 1968), adapted by the author) has argued pervasuasively that Bayes' rule

is not the only reasonable way to update", quoted from Diaconis and Zabell (1982), see also Chan and Darwiche (2005); Pan et al. (2006); Darwiche (2009); Valtorta et al. (2002); Buscemi and Scarani (2021). The evidence for this update rule of Jeffrey is not given by a predicate on the codomain of the channel, but by a distribution. Jeffrey's rule can be formulated conveniently (see Jacobs (2019)) in terms of the Bayesian inverse (dagger) $c^\dagger$ of the channel — as introduced in the beginning of this chapter.

As we shall see below, the update rules of Pearl and Jeffrey can give quite different outcomes. Hence the question arises: when to use which rule? What is the difference? There is a clear difference, which can be summarised as follows: Pearl's rule increases validity and Jeffrey's rule decreases divergence. This will be made technically precise, in Theorem 7.9.3 below. The proof in Pearl's case is easy, but the proof in Jeffrey's case is remarkably difficult. We postpone it to **??**, where it arises in a broader analysis of (statistical) learning.

In general terms, one can learn by reinforcing what goes well, or by steering away from what goes wrong. In the first case one improves a positive evaluation and in the second case one reduces a negative outcome, as a form of correction. Thus, Jeffrey's update rule is a correction mechanism, for correcting errors. As such it is used in 'predictive coding', a theory in cognitive science that views the human brain as a Bayesian prediction and correction engine, see Rao and Ballard (1999); Friston (2010); Allen and Friston (2018); Hohwy (2013); Clark (2016); Smith et al. (2022).

This section introduces and illustrates Jeffrey's update rule, especially in contrast to Pearl's rule, that is, to backward inference. We repeat Pearl's rule from Definition 6.2.1 (2), for convenience and clarity.

Definition 7.9.1 Let $c\colon X \to\!\!\!\!\bullet\, Y$ be channel with a prior state $\sigma \in \mathcal{D}(X)$.

1 *Pearl's rule* is backward inference: it involves updating of the prior σ with evidence in the form of a *factor* $q \in Fact(Y)$ to the posterior:

$$\sigma|_{c^*(q)} \;\in\; \mathcal{D}(X).$$

2 *Jeffrey's rule* involves update of the prior σ with evidence in form of a *distribution* $\tau \in \mathcal{D}(Y)$ to the posterior:

$$(c^\dagger_\sigma)_*(\tau) \;\in\; \mathcal{D}(X).$$

Both rules occur in the literature, for instance in Pan et al. (2006). There, the distribution τ is called 'soft evidence' and the factor q is called a 'likelihood ratio'; it occurs in the form of a predicate.

We illustrate the difference between Pearl's and Jeffreys's rules in a standard medical test situation.

Example 7.9.2 Consider the situation from Exercise 6.2.1 with set $D = \{d, d^\perp\}$ for disease (or not) and $T = \{p, n\}$ for a positive or negative test outcome, with a test channel $c\colon D \to\!\!\!\!\bullet\, T$ given by:

$$c(d) \;=\; \tfrac{9}{10}|p\rangle + \tfrac{1}{10}|n\rangle \qquad\qquad c(d^\perp) \;=\; \tfrac{1}{20}|p\rangle + \tfrac{19}{20}|n\rangle.$$

The test thus has a sensitivity of 90% and a specificity of 95%. We assume a prevalence of 10% via a prior state $\omega = \frac{1}{10}|d\rangle + \frac{9}{10}|d^\perp\rangle$.

The test is performed, under unfavourable circumstances, like bad light, and so we are only

80% sure that the test is positive. With Pearl's update rule we thus use as evidence predicate $q = \frac{4}{5} \cdot \mathbf{1}_p + \frac{1}{5} \cdot \mathbf{1}_n$. It gives as posterior disease probability:

$$\omega|_{c^*(q)} = \tfrac{74}{281}|d\rangle + \tfrac{207}{281}|d^\perp\rangle \approx 0.26|d\rangle + 0.73|d^\perp\rangle.$$

This gives a disease likelihood of 26%.

When we decide to use Jeffrey's rule we translate the 80% certainty of a positive test into a distribution $\tau = \frac{4}{5}|p\rangle + \frac{1}{5}|n\rangle$. Then we compute:

$$\begin{aligned}(c^\dagger_\omega)_*(\tau) &= \tfrac{4}{5} \cdot c^\dagger_\omega(p) + \tfrac{1}{5} \cdot c^\dagger_\omega(n) \overset{(7.1)}{=} \tfrac{4}{5} \cdot \omega|_{c^*(\mathbf{1}_p)} + \tfrac{1}{5} \cdot \omega|_{c^*(\mathbf{1}_n)} \\ &= \tfrac{4}{5} \cdot \Big(\tfrac{2}{3}|d\rangle + \tfrac{1}{3}|d^\perp\rangle\Big) + \tfrac{4}{5} \cdot \Big(\tfrac{2}{173}|d\rangle + \tfrac{171}{173}|d^\perp\rangle\Big) \\ &= \tfrac{278}{519}|d\rangle + \tfrac{241}{519}|d^\perp\rangle \\ &\approx 0.54|d\rangle + 0.46|d^\perp\rangle.\end{aligned} \tag{7.46}$$

The disease likelihood is now 54%, more than twice as high as with Pearl's update rule. This is a serious difference, which may have serious consequences. Should we start asking our doctors: does your computer use Pearl's or Jeffrey's rule to calculate probabilities, in order to produce an advice about medical treatment (or not)?

Interpreting the difference is a challenge. Here is one way to view the situation[4]. Suppose that two people are involved, say Alice and Bob, where Alice does the updating and Bob the tests. In the Pearl case, one could say that Bob himself adds the uncertainty when he reports the test outcome, saying something like: I'm 80% certain that I saw a positive test. This gets incorporated into the predicate $q = \frac{4}{5} \cdot \mathbf{1}_p + \frac{1}{5} \cdot \mathbf{1}_n$ in the beginning of the example. In the Jeffrey case Bob is also reporting test outcomes, but now with certainty, and it is Alice who adds uncertainty, corresponding to how much she trusts Bob's work. Hence when Bob reports a postive test outcome and Alice has only 80% trust in Bob, she takes a weighted (convex) sum of $(\frac{4}{5}, \frac{1}{5})$ of the two point updates, with positive and negative tests. This is incorporated in the evidence distribution $\tau = \frac{4}{5}|p\rangle + \frac{1}{5}|n\rangle$ in the example. Consistent with this 80% trust, when Bob reports a *negative* test, Alice would compute, as in (7.46):

$$\begin{aligned}\tfrac{4}{5} \cdot c^\dagger_\omega(n) + \tfrac{1}{5} \cdot c^\dagger_\omega(p) &= \tfrac{4}{5} \cdot \Big(\tfrac{2}{173}|d\rangle + \tfrac{171}{173}|d^\perp\rangle\Big) + \tfrac{1}{5} \cdot \Big(\tfrac{2}{3}|d\rangle + \tfrac{1}{3}|d^\perp\rangle\Big) \\ &= \tfrac{74}{519}|d\rangle + \tfrac{445}{519}|d^\perp\rangle \approx 0.143|d\rangle + 0.857|d^\perp\rangle.\end{aligned}$$

The disease probability is higher than the prevalence of 0.1. In contrast, in the Pearl case, when Bob reports a negative test with 80% certainty we get as posterior, for $q' = \frac{1}{5} \cdot \mathbf{1}_p + \frac{4}{5} \cdot \mathbf{1}_n$. It gives as posterior disease probability:

$$\omega|_{c^*(q')} = \tfrac{26}{719}|d\rangle + \tfrac{693}{719}|d^\perp\rangle \approx 0.036|d\rangle + 0.964|d^\perp\rangle.$$

The disease probability is now below the prevalence.

We conclude with brief descriptions of this medical test example in the probabilistic programming languages ProbLog and WebPPL, with both Pearl's and Jeffrey's updating. In ProbLog Pearl's approach can be handled in a standard manner with an additional predicate / channel, see the code fragment at the top in Figure 7.5. Jeffrey's approach is not directly supported in ProbLog but can be incorporated by exploiting ProbLog's parametrisation mechanism, see Figure 7.5 at the bottom. The different instantations (for numbers 1, 2) give us fresh copies of the prior disease

[4] Suggested by Tom Heskes.

```
0.10 :: pearl_disease.

0.9  :: positive_test :- pearl_disease.
0.05 :: positive_test :- not pearl_disease.

0.8 :: pearl_evidence :- positive_test.
0.2 :: pearl_evidence :- not positive_test.

evidence(pearl_evidence).
query(pearl_disease).
```

```
0.10 :: disease(V).

0.9  :: positive_test(V) :- disease(V).
0.05 :: positive_test(V) :- not disease(V).

evidence(positive_test(1)).
evidence(not positive_test(2)).

0.8 :: jeffrey_evidence.

jeffrey_disease :- jeffrey_evidence, disease(1).
jeffrey_disease :- not jeffrey_evidence, disease(2).

query(jeffrey_disease).
```

Figure 7.5 ProbLog code for updating, via Pearl's rule on top, and via Jeffrey's rule at the bottom, see Example 7.9.2.

distribution, which are separately updated with different point evididence. The posterior disease distribution for Jeffrey is then computed as a convex sum, like in (7.46), following the general approach described in (2.66). Running these programs produces the probabilities for Pearl and Jeffrey updating that we calculated above:

```
pearl_disease: 0.2633452          jeffrey_disease: 0.53564547
```

We turn to update rules of Pearl and Jeffrey in WebPPL, in this medical test example, see Figure 7.6. It contains the Pearl update in two forms. The function `pearl_update1` is the most direct implementation, using a fuzzy predicate. The function `pearl_update2` does not use this predicate, but incorporates the 80% certainty about a positive test in a sample `ev` that is required to be equal to the test outcome. Both distributions written to the console are equal to `0.737|false> + 0.263|true>`.

The Jeffrey update in WebPPL at the bottom of Figure 7.6 also uses such an evidence sample, but then goes into a nested inference to compute the posterior via sampling from the inner `Infer`. In this way one forms the pushforward along a dagger channel, as in (7.46), see also Ex-

```
var test_channel = function(d) {
    return d ? Categorical({ ps : [0.9, 0.1], vs :['pos', 'neg'] })
             : Categorical({ ps : [0.05, 0.95], vs : ['pos', 'neg'] }) }

var pearl_evidence_predicate = function(t) {
    return (t == 'pos') ? flip(0.8) : flip(0.2); };

var pearl_update1 = Infer(function() {
   var d = flip(0.1);
   var t = sample(test_channel(d));
   var ev = pearl_evidence_predicate(t);
   condition( ev == true );
   return disease; })

var pearl_update2 = Infer(function() {
    var ev = flip(0.8);
    var d = flip(0.1);
    var t = sample(test_channel(d));
    condition( (t == 'pos') == ev )
    return d})

console.log( as_kets( pearl_update1 ) )
console.log( as_kets( pearl_update2 ) )
```

```
var jeffrey_update = Infer(function() {
    var ev = flip(0.8);
    return sample(Infer(function() {
              var d = flip(0.1);
              var t = sample(test_channel(d));
              condition( (t == 'pos') == ev );
              return d})) })

console.log( as_kets( jeffrey_update ) )
```

Figure 7.6 WebPPL code for updating, on top via Pearl's rule and at the bottom via Jeffrey's rule, see Example 7.9.2.

ercise 7.9.11. This function `jeffrey_update` is subtly different from the second implementation `pearl_update2`. It produces the distribution `0.536|true> + 0.464|false>`.

The following theorem (from Jacobs (2021b)) captures the essences of the update rules of Pearl and Jeffrey: validity increase or divergence decrease. Informally, we could say that Pearl is about *improvement* learning and Jeffrey about *correction* learning. For convenience we repeat the Pearl case.

Theorem 7.9.3 *Consider the situation in Definition 7.9.1, with a prior distribution $\sigma \in \mathcal{D}(X)$ and a channel $c \colon X \rightsquigarrow Y$.*

1 *For a factor q on Y, Pearl's update rule gives an increase of validity:*

$$c_*(\sigma_P) \models q \;\geq\; c_*(\sigma) \models q \quad \textit{for the updated state} \quad \sigma_P \coloneqq \sigma|_{c^*(q)}.$$

2 *For an evidence distribution $\tau \in \mathcal{D}(Y)$, Jeffrey's update rule gives a decrease of Kullback-Leibler divergence:*

$$D_{KL}\big(\tau, c_*(\sigma_J)\big) \;\leq\; D_{KL}\big(\tau, c_*(\sigma)\big) \qquad \textit{for} \qquad \sigma_J \coloneqq (c_\sigma^\dagger)_*(\tau). \tag{7.47}$$

In fact, this inequality can be strengthened to:

$$D_{KL}\big(\tau,\, c_*(\sigma_J)\big) \,+\, D_{KL}\big(\sigma_J, \sigma\big) \;\leq\; D_{KL}\big(\tau, c_*(\sigma)\big). \tag{7.48}$$

However, we mostly use the above weaker form (7.47). In both these cases we assume that the predicted distribution $c_*(\sigma)$ has fulll support, so that the dagger $c_\sigma^\dagger$ is well-defined.

In this sitation $c_*(\sigma)$ is the prediction, which is used to evaluate the evidence.

- In Pearl's case we look at the validity $c_*(\sigma) \models q$ of the evidence q in this predicted state. The above theorem tells that by switching to the Pearl-update σ_P that we get a higher validity $c_*(\sigma)_P \models q$ of the evidence, in the newly predicted state $c_*(\sigma_P)$.
- In Jeffrey's case we look at the divergence / mismatch $D_{KL}(\tau, c_*(\sigma))$ between the evidence distribution τ and the prediction $c_*(\sigma)$. By changing to the Jeffrey-update σ_J we get a lower divergence $D_{KL}(\tau, c_*(\sigma_J))$. Thus, via Jeffrey's rule one reduces 'prediction errors', in the terminology of predictive coding theory.

The above theorem is formulated for a 'generative model' in the form of a distribution $\sigma \in \mathcal{D}(X)$ and a channel $c\colon X \rightsquigarrow Y$. It may equivalently be formulated in terms of a joint distribution $\omega \in \mathcal{D}(X \times Y)$, see Exercise 7.9.10. The latter formulation leads to a unique characterisation of the updates.

A first proof of the Jeffrey inequality (7.47) appeared in Jacobs (2021b), using rather heavy machinery (the Gelfand spectral radius theorem). A more elementary proof will be given below; it emerged from discussions with Dario Stein. Another proof, for the special case where the evidence distribution τ is fractional, will appear later, in Remark 8.2.6.

Proof. Item (1) is a repetition of Theorem 6.2.2. The proof of item (7.47) will be based on Proposition 7.8.6 (1), instantiated with $\sigma_J = (c_\sigma^\dagger)_*(\tau)$ as base distribution in the shape $S =$ and with distributions $\rho = \langle c_\sigma^\dagger, id\rangle_*(\tau)$, $\omega = \langle id, c\rangle_*(\sigma)$, and thus $\omega_S = \langle id, c\rangle_*(\sigma_J)$. Further, we use basic properties of KL-divergence, as in Proposition 2.10.6 (3) and in the 'special case' of Proposition 2.10.2 (2).

$$\begin{aligned}
&D_{KL}\big(\tau,\, c_*(\sigma_J)\big) \,+\, D_{KL}\big(\sigma_J,\, \sigma\big) \\
&= D_{KL}\Big((\pi_2)_*(\langle c_\sigma^\dagger, id\rangle_*(\tau)),\, (\pi_2)_*(\langle id, c\rangle_*(\sigma_J))\Big) \,+\, D_{KL}\big(\sigma_J,\, \sigma\big) \\
&\leq D_{KL}\Big(\langle c_\sigma^\dagger, id\rangle_*(\tau),\, \langle id, c\rangle_*(\sigma_J)\Big) \,+\, D_{KL}\Big(\langle id, c\rangle_*(\sigma_J),\, \langle id, c\rangle_*(\sigma)\Big) \\
&= D_{KL}\Big(\langle c_\sigma^\dagger, id\rangle_*(\tau),\, \langle id, c\rangle_*(\sigma)\Big) && \text{by Proposition 7.8.6 (1)} \\
&= D_{KL}\Big(\langle c_\sigma^\dagger, id\rangle_*(\tau),\, \langle c_\sigma^\dagger, id\rangle_*(c_*(\sigma))\Big) && \text{by definition of the dagger} \\
&= D_{KL}\big(\tau,\, c_*(\sigma)\big).
\end{aligned}$$

❑

We review an earlier illustration, now with Jeffrey's approach emphasising the decrease of divergence between the predication and the evidence distribution.

Example 7.9.4 Recall the situation of Example 2.4.3, with a teacher predicting the performance of pupils, depending on the teacher's mood. The evidence q from Example 6.2.5 is used for Pearl's update. It can be translated into a distribution τ on the set G of grades:

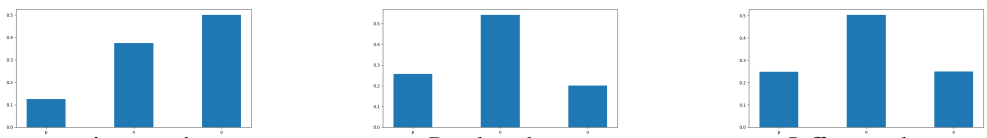
$$\tau \;=\; \tfrac{1}{10}|1\rangle + \tfrac{3}{10}|2\rangle + \tfrac{3}{10}|3\rangle + \tfrac{2}{10}|4\rangle + \tfrac{1}{10}|5\rangle.$$

There is an a priori divergence $D_{KL}(\tau, c_*(\sigma)) \approx 1.336$. With some effort one can prove that the Jeffrey-update of σ is:

$$\sigma' \;=\; (c_\sigma^\dagger)_*(\tau) \;=\; \tfrac{972795}{3913520}|p\rangle + \tfrac{1966737}{3913520}|n\rangle + \tfrac{973988}{3913520}|o\rangle \;\approx\; 0.2486|p\rangle + 0.5025|n\rangle + 0.2489|o\rangle.$$

The divergence has now dropped to: $D_{KL}(\tau, c_*(\sigma')) \approx 1.087$.

In the end it is interesting to compare the orinal (prior) mood with its Pearl- and Jeffrey-updates. In both cases we see that after the bad marks of the pupils the teacher has become less optimistic — as expressed via the third bar.



prior mood Pearl-update Jeffrey-update

The prior mood is reproduced from Example 2.4.3 for easy comparison. The Pearl and Jeffrey updates differ only slightly in this case.

The fact that Jeffrey's rule involves correction of prediction errors, as stated in Theorem 7.9.3 (2), supports the view that Jeffrey's update rule should be used in situations where one is confronted with 'surprises' Dietrich et al. (2016) or with 'unanticipated knowledge' Diaconis and Zabell (1983) that one has to adapt to, as a form of correction. Here we include an example from Dietrich et al. (2016) (also used in Jacobs (2019)) that involves such error correction after a 'surprise'.

Example 7.9.5 Ann must decide about hiring Bob, whose characteristics are described in terms of a combination of competence (c or $c^\perp$) and experience (e or $e^\perp$). The prior, based on dealing with many earlier candidates, is a joint distribution on the product space $C \times E$, for $C = \{c, c^\perp\}$ and $E = \{e, e^\perp\}$, given as:

$$\omega \;=\; \tfrac{4}{10}|c, e\rangle + \tfrac{1}{10}|c, e^\perp\rangle + \tfrac{1}{10}|c^\perp, e\rangle + \tfrac{4}{10}|c^\perp, e^\perp\rangle.$$

The first marginal of ω is the uniform distribution $\frac{1}{2}|c\rangle + \frac{1}{2}|c^\perp\rangle$. It is the neutral base rate for Bob's competence.

We use the two projection functions $C \xleftarrow{\pi_1} C \times E \xrightarrow{\pi_2} E$ as deterministic channels along which we reason, using both Pearl's and Jeffrey's rules.

When Ann would learn that Bob has relevant work experience, given by point evidence $\mathbf{1}_e$, her strategy is to factor this in via Pearl's rule / backward inference: this gives as posterior $\omega|_{(\pi_2)^*(\mathbf{1}_e)} = \omega|_{\mathbf{1}\otimes\mathbf{1}_e}$, whose first marginal is $\frac{4}{5}|c\rangle + \frac{1}{5}|c^\perp\rangle$. It is then more likely that Bob is competent.

Ann reads Bob's letter to find out if he actually has relevant experience. We quote from Dietrich et al. (2016):

Bob's answer reveals right from the beginning that his written English is poor. Ann notices this even before figuring out what Bob says about his work experience. In response to this unforeseen learnt input, Ann lowers her probability that Bob is competent from $\frac{1}{2}$ to $\frac{1}{8}$. It is natural to model this as an instance of Jeffrey revision.

Bob's poor English is a new state of affairs: a surprise that causes Ann to switch to error reduction mode, via Jeffrey's rule. Bob's poor command of the English language translates into a competence distribution $\rho = \frac{1}{8}|c\rangle + \frac{7}{8}|c^{\perp}\rangle$. Ann wants to adapt to this new surprising situation, so she uses Jeffrey's rule, giving a new joint state:

$$\omega' = ((\pi_1)^{\dagger}_{\omega})_*(\rho) = \tfrac{1}{10}|c, e\rangle + \tfrac{1}{40}|c, e^{\perp}\rangle + \tfrac{7}{40}|c^{\perp}, e\rangle + \tfrac{7}{10}|c^{\perp}, e^{\perp}\rangle.$$

If the letter now tells that Bob has work experience, Ann will factor this in, in this new situation ω', giving, like above, via Pearl's rule followed by marginalisation,

$$\omega'|_{(\pi_2)^*(\mathbf{1}_e)}[1, 0] = \tfrac{4}{11}|c\rangle + \tfrac{7}{11}|c^{\perp}\rangle.$$

The likelihood of Bob being competent is now lower than in the prior state, since $\frac{4}{11} < \frac{1}{2}$. This example reconstructs the illustration from Dietrich et al. (2016) in channel-based form, with the associated formulations of Pearl's and Jeffrey's rules from Definition 7.9.1, and produces exactly the same outcomes as in *loc. cit.*

Jeffrey's update role plays an important role in computational neuroscience, as illustrated next.

Remark 7.9.6 Recall the two forms of surprise from active inference described in Example 7.1.9. In this cognitive setting the notion of *variational free energy* plays an important role. It assumes a 'generative model' in the form of a distribution $\sigma \in \mathcal{D}(X)$ and a channel $c\colon X \rightsquigarrow Y$. This free energy is defined as a function $\mathcal{F}\colon \mathcal{D}(X) \times Y \to \mathbb{R}_{\geq 0}$. It can be formulated in several equivalent ways, following (Parr et al., 2022, Eqn. (2.5)):

$$\begin{aligned}
\mathcal{F}(\omega, y) &:= -\Big(\omega \models \ln\big(\sigma(-) \cdot c(-)(y)\big)\Big) - H(\omega) \\
&= \sum_{x \in X} \omega(x) \cdot \ln\left(\frac{\omega(x)}{\sigma(x) \cdot c(x)(y)}\right) \\
&= D_{KL}\Big(\omega,\, c^{\dagger}_{\sigma}(y)\Big) - \ln\Big(c_*(\sigma)(y)\Big) \\
&= D_{KL}\Big(\omega,\, (c^{\dagger}_{\sigma})_*(1|y\rangle)\Big) + D_{KL}\Big(1|y\rangle,\, c_*(\sigma)\Big).
\end{aligned} \tag{7.49}$$

As in Example 7.1.9, it makes sense to replace the point distribution $1|y\rangle$ in the last line by an arbitrary distribution on Y. We thus extend variational free energy to a function $\mathcal{F}\colon \mathcal{D}(X) \times \mathcal{D}(Y) \to \mathbb{R}_{\geq 0}$, defined as:

$$\mathcal{F}(\omega, \tau) := D_{KL}\Big(\omega,\, (c^{\dagger}_{\sigma})_*(\tau)\Big) + D_{KL}\Big(\tau,\, c_*(\sigma)\Big). \tag{7.50}$$

This is a key concept: "Active Inference is a theory of how living organisms maintain their existence by minimizing surprise — or a tractable proxy to surprise, *variational free energy* — via perception and action." (Parr et al., 2022, p. 42). There is a duality of perception and action. When the external state $\tau \in \mathcal{D}(Y)$ is fixed, the free energy in (7.50) can be reduced by moving to a new internal state ω that is close to the Jeffrey update $(c^{\dagger}_{\sigma})_*(\tau)$, via perception. Dually, when one's internal state $\omega \in \mathcal{D}(X)$ is fixed one can reduce free energy by bringing the external state τ close to the prediction $c_*(\sigma)$, via action.

We briefly discuss some further commonalities and differences between the rules of Pearl and Jeffrey. We have seen some of these results before, but now they are put in the context of Pearl's and Jeffrey's update rule.

Proposition 7.9.7 *Let $c\colon X \rightsquigarrow Y$ be a channel with a prior state $\sigma \in \mathcal{D}(X)$.*

1 *Pearl's rule and Jeffrey's rule agree on point predicate / states: for $y \in Y$, with associated point predicate $\mathbf{1}_y$ and point state $1|y\rangle$, one has:*

$$\sigma|_{c^*(\mathbf{1}_y)} = c^\dagger_\sigma(y) = (c^\dagger_\sigma)_*(1|y\rangle).$$

2 *Pearl's updating with a constant predicate (no information) does not change the prior state ω:*

$$\sigma|_{c^*(r\cdot\mathbf{1})} = \sigma, \qquad \text{for } r > 0.$$

Jeffrey's update does not change anything when we update with what we already predict:

$$(c^\dagger_\sigma)_*(c_*(\sigma)) = \sigma.$$

3 *For a factor $q \in \mathsf{Fact}(Y)$ we can both update the prior distribution σ, according to Pearl, and also the channel c, so that the updated predicted state is predicted:*

$$(c|_q)_*(\sigma|_{c^*(q)}) = (c_*(\sigma))|_q.$$

In Jeffrey's case, with an evidence state $\tau \in \mathcal{D}(Y)$, we can update both the state and the channel, via a double-dagger, so that the evidence state τ is predicted:

$$\Big((c^\dagger_\sigma)^\dagger_\tau\Big)_*((c^\dagger_\sigma)_*(\tau)) = \tau.$$

4 *Multiple updates with Pearl's rule commute, but multiple updates with Jeffrey's rule do not commute.*

Proof. These items follow from earlier results.

1 Directly by Definition 7.9.1.
2 The first equation follows from Lemma 6.1.7 (1) and (4); the second one is Equation 7.4.
3 The first claim is Theorem 6.3.4 and the second one is Lemma 7.1.8 (4).
4 By Lemma 6.1.7 (3) we have:

$$\sigma|_{c^*(q)_1}|_{c^*(q)_2} = \sigma|_{c^*(q_1)\,\&\,c^*(q_2)} = \sigma|_{c^*(q_2)}|_{c^*(q_1)}.$$

However, in general, for evidence states $\tau_1, \tau_2 \in \mathcal{D}(Y)$,

$$\Big(c^\dagger_{(c^\dagger_\sigma)_*(\tau_1)}\Big)_*(\tau_2) \neq \Big(c^\dagger_{c^\dagger_{\sigma*}(\tau_2)}\Big)_*(\tau_1).$$

Exercise 7.9.5 contains a concrete example. ❑

Item (2) presents different views on 'learning', where learning is now used in an informal sense: according to Pearl's rule you learning nothing when you get no information; but according to Jeffrey you learn nothing when you are presented with what you already know. Both interpretations make sense.

Probabilistic updating may be used as a model for what is called *priming* in cognitive science, see *e.g.* Griffiths et al. (2008); Hohwy (2013). It is well-known that the human mind is sensitive to the order in which information is processed, that is, to the order of priming / updating. Thus, the last item (4) above suggests that Jeffrey's rule might be more appropriate in such a setting.

This strengthens the view in predictive coding that the human mind learns from error correction, as in Jeffrey's update rule.

Remark 7.9.8 In the context of the proof of Proposition 7.9.7 (4) there is a way to a Jeffrey update with both distributions τ_1 and τ_2, namely via the tuple channel, as:

$$\Big(\langle c, c\rangle^{\dagger}_{\sigma}\Big)_{*}(\tau_1 \otimes \tau_2). \tag{$*$}$$

In this formula the order of the τ_i does not matter. Hence this tuple-channel approach makes sense for multiple Jeffrey updates. It occurs in the last item of Exercise 7.9.9. The characterisation of Jeffrey updating in terms of I-projection from Exercise 7.9.10 (3) extends to this multiple update form: the joint distribution $\omega = \langle id, c, c\rangle_{*}(\omega) \in \mathcal{D}(X \times Y \times Y)$ has an I-projection into the set $\Pi_{\tau_1,\tau_2} = \{\rho \in \mathcal{D}(X \times Y \times Y) \mid \rho[0,1,0] = \tau_1 \text{ and } \rho[0,0,1] = \tau_2\}$. This I-projection is of the form $\langle\langle c, c\rangle^{\dagger}_{\sigma}, id\rangle_{*}(\tau_1 \otimes \tau_2)$ and has the above Jeffrey update $(*)$ as first marginal.

This tuple-channel approach can also be used for Pearl updating and corresponds to updating with a conjunction, and thus to iterated updating in any order:

$$\omega\big|_{\langle c,c\rangle^{*}(q_1 \otimes q_2)} = \omega\big|_{c^{*}(q_1)\,\&\,c^{*}(q_2)} = \omega\big|_{c^{*}(q_1)}\big|_{c^{*}(q_2)} = \omega\big|_{c^{*}(q_2)}\big|_{c^{*}(q_1)}.$$

This approach will be elaborated in Section 8.3.

There are translations back-and-forth between Pearl's and Jeffrey's rules, due to Chan and Darwiche (2005); they are translated here to the current setting.

Proposition 7.9.9 *Let $c\colon X \rightsquigarrow Y$ be a channel with a prior state $\sigma \in \mathcal{D}(X)$ on its domain, such that the predicted state $c_{*}(\sigma)$ has full support.*

1 *Pearl's updating can be expressed as Jeffrey's update, by turning a factor $q\colon Y \to \mathbb{R}_{\geq 0}$ into a state $c_{*}(\sigma)|_{q} \in \mathcal{D}(Y)$, so that it can be used as evidence in:*

$$(c^{\dagger}_{\sigma})_{*}\Big(c_{*}(\sigma)\big|_{q}\Big) = \sigma|_{c^{*}(q)}.$$

2 *Jeffrey's updating can also be expressed as Pearl's updating: for an evidence state $\tau \in \mathcal{D}(Y)$ write $\tau/c_{*}(\sigma)$ for the factor $y \mapsto \frac{\tau(y)}{c_{*}(\sigma)(y)}$; then:*

$$(c^{\dagger}_{\sigma})_{*}(\tau) = \sigma|_{c^{*}(\tau/c_{*}(\sigma))}.$$

Proof. The first item is exactly Theorem 7.1.7 (1). For the second item we first note that the fraction $\tau/c_{*}(\sigma)$ is a well-defined factor, since $c_{*}(\sigma)$ has full support. Further, its validity $c_{*}(\sigma) \models \tau/c_{*}(\sigma)$ is one, since τ is a distribution:

$$c_{*}(\sigma) \models \tau/c_{*}(\sigma) = \sum_{y \in Y} c_{*}(\sigma)(y) \cdot \frac{\tau(y)}{c_{*}(\sigma)(y)} = \sum_{y \in Y} \tau(y) = 1. \tag{7.51}$$

But then, for $x \in X$,

$$\begin{aligned}
(c^\dagger_\sigma)_*(\tau)(x) &= \sum_{y \in Y} \tau(y) \cdot c^\dagger_\sigma(y)(x) \overset{(7.1)}{=} \sum_{y \in Y} \tau(y) \cdot \frac{\sigma(x) \cdot c(x)(y)}{c_*(\sigma)(y)} \\
&= \sigma(x) \cdot \sum_{y \in Y} c(x)(y) \cdot \frac{\tau(y)}{c_*(\sigma)(y)} \\
&= \frac{\sigma(x) \cdot c^*(\tau/c_*(\sigma))(x)}{c_*(\sigma) \models \tau/c_*(\sigma)} \qquad \text{as just shown} \\
&= \frac{\sigma(x) \cdot c^*(\tau/c_*(\sigma))(x)}{\sigma \models c^*(\tau/c_*(\sigma))} \\
&= \sigma|_{c^*(\tau/c_*(\sigma))}(x).
\end{aligned}$$

❑

Combination of Theorem 7.9.3 and Proposition 7.9.9 tells us how Pearl's rule can also lead to a decrease of divergence. The situation is quite subtle and will be discussed further in the subsequent remarks.

Corollary 7.9.10 *Let $c \colon X \rightsquigarrow Y$ be a channel with a prior state $\sigma \in \mathcal{D}(X)$ and a factor q on Y. Then:*

$$D_{KL}\Big(c_*(\sigma)|_q,\ c_*(\sigma|_{c^*(q)})\Big) \leq D_{KL}\Big(c_*(\sigma)|_q,\ c_*(\sigma)\Big).$$

This result can be interpreted as follows. With my prior state σ I can predict $c_*(\sigma)$. I can update this prediction with evidence q to $c_*(\sigma)|_q$. The divergence between this update and my prediction is bigger than the divergence between the update and the prediction from the Pearl / Bayes posterior $\sigma|_{c^*(q)}$. Thus, the posterior gives a correction.

Proof. Take $\tau = c_*(\sigma)|_q$. Theorem 7.9.3 (2) says that:

$$D_{KL}\Big(\tau, c_*((c^\dagger_\sigma)_*(\tau))\Big) \leq D_{KL}\Big(\tau, c_*(\sigma)\Big).$$

But Proposition 7.9.9 (1) tells that $(c^\dagger_\sigma)_*(\tau) = \sigma|_{c^*(q)}$. ❑

Lemma 7.9.11 *Consider a channel $c \colon X \rightsquigarrow Y$ with states $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ on its domain and codomain, such that $c_*(\omega)$ has full support. Then there are inequalities:*

$$D_{KL}\Big((c^\dagger_\omega)_*(\rho),\ \omega\Big) \leq D_{KL}\Big(\rho,\ c_*(\omega)\Big) \qquad\qquad D_{KL}\Big(\omega,\ (c^\dagger_\omega)_*(\rho)\Big) \leq D_{KL}\Big(c_*(\omega),\ \rho\Big).$$

Proof. Both inequalities are obtained via Equation 7.4 and Proposition 2.10.6 (3):

$$\begin{aligned}
D_{KL}\Big((c^\dagger_\omega)_*(\rho),\ \omega\Big) &= D_{KL}\Big((c^\dagger_\omega)_*(\rho),\ (c^\dagger_\omega)_*(c_*(\omega))\Big) \leq D_{KL}\Big(\rho,\ c_*(\omega)\Big) \\
D_{KL}\Big(\omega,\ (c^\dagger_\omega)_*(\rho)\Big) &= D_{KL}\Big((c^\dagger_\omega)_*(c_*(\omega)),\ (c^\dagger_\omega)_*(\rho)\Big) \leq D_{KL}\Big(c_*(\omega),\ \rho\Big).
\end{aligned}$$

❑

Remarks 7.9.12 1 In Corollary 7.9.10 we use that Pearl's rule can be expressed as Jeffrey's, see point (1) in Proposition 7.9.9. The other way around, Jeffrey's rule is expressed via Pearl's in point (2). This leads, in principle, to a validity increase property for Jeffrey's rule: let $c \colon X \rightsquigarrow Y$ be a channel with states $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$. Assuming that $c_*(\sigma)$ has full support we can form the factor $q = \tau/c_*(\sigma)$. We then get an inequality:

$$c_*((c^\dagger_\sigma)_*(\tau)) \models q \ = \ c_*(\sigma|_{c^*(q)}) \models q \ \geq \ c_*(\sigma) \models q \overset{(7.51)}{=} 1.$$

This is not very useful.

2 We have emphasised the message "Pearl increases validity" and "Jeffrey decreases divergence". But Corollary 7.9.10 seems to nuance this message, since it describes a divergence decrease for Pearl too, and the previous item gives a validity increase for Jeffrey. What is going on? We try to clarify the situation via an example, demonstrating that the validity increase of Pearl's rule fails for Jeffrey's rule and that the divergence decrease of Jeffrey's rule fails for Pearl's rule.

Take sets $X = \{0, 1\}$ and $Y = \{a, b, c\}$ with uniform prior $\sigma = \frac{1}{2}|0\rangle + \frac{1}{2}|1\rangle \in \mathcal{D}(X)$. We use the channel $c\colon X \rightsquigarrow Y$ given by:

$$c(0) = \tfrac{1}{9}|a\rangle + \tfrac{2}{3}|b\rangle + \tfrac{2}{9}|c\rangle \qquad \text{and} \qquad c(1) = \tfrac{7}{25}|a\rangle + \tfrac{7}{25}|b\rangle + \tfrac{11}{25}|c\rangle.$$

The predicted state is then $c_*(\sigma) = \frac{44}{225}|a\rangle + \frac{71}{150}|b\rangle + \frac{149}{450}|c\rangle$. We use as 'equal' evidence predicate and state:

$$q = \tfrac{1}{2} \cdot \mathbf{1}_a + \tfrac{1}{3} \cdot \mathbf{1}_b + \tfrac{1}{6} \cdot \mathbf{1}_c \qquad \text{and} \qquad \tau = \tfrac{1}{2}|a\rangle + \tfrac{1}{3}|b\rangle + \tfrac{1}{6}|c\rangle.$$

We then get the following updates, according to Pearl and Jeffrey, respectively:

$$\begin{aligned} \sigma_P &= \sigma|_{c^*(q)} \\ &= \tfrac{425}{839}|0\rangle + \tfrac{414}{839}|1\rangle \\ &\approx 0.5066|0\rangle + 0.4934|1\rangle \end{aligned} \qquad \text{and} \qquad \begin{aligned} \sigma_J &= (c_\sigma^\dagger)_*(\tau) \\ &= \tfrac{805675}{1861904}|0\rangle + \tfrac{1056229}{1861904}|1\rangle \\ &\approx 0.4327|0\rangle + 0.5673|1\rangle. \end{aligned}$$

The validities are summarised in the following table.

description	**formula**	**value**
prior validity	$c_*(\sigma) \models q$	0.31074
after Pearl	$c_*(\sigma)_P \models q$	0.31079
after Jeffrey	$c_*(\sigma)_J \models q$	0.31019

The differences are small, but relevant. Pearl's updating increases validity, as Theorem 7.9.3 (1) dictates, but Jeffrey's updating does not, in this example.

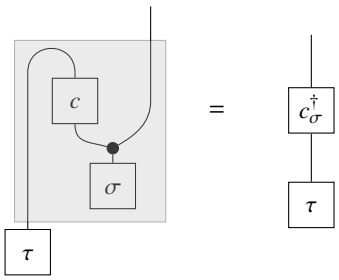
The divergences in this example are as follows.

description	**formula**	**value**
prior divergence	$D_{KL}(\tau, c_*(\sigma))$	0.238
after Pearl	$D_{KL}(\tau, c_*(\sigma)_P)$	0.240
after Jeffrey	$D_{KL}(\tau, c_*(\sigma)_J)$	0.221

Jeffrey's rule decreases divergence, in line with Theorem 7.9.3 (2), but Pearl's updating does not. Is there a contradiction with Corollary 7.9.10, which does involve a divergence decrease? No, since there the state τ has a very particular shape, namely $c_*(\sigma)|_q$. We conclude that for that particular state Pearl's rule gives a divergence decrease, but there is no such decrease in general.

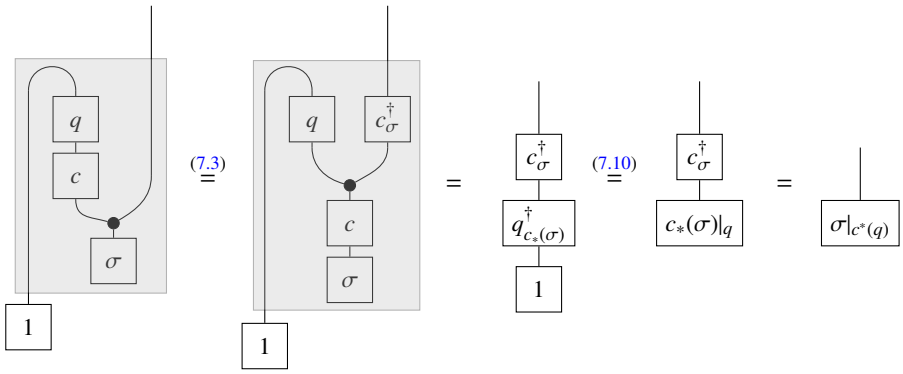
Thus, in general, validity increase works exclusively for Pearl's rule, and divergence decrease works exclusively for Jeffrey's rule.

We conclude this section with a graphical description of the update rules of Pearl and Jeffrey, using string diagrams extended with the shaded box notation, see Remark 7.3.4. So let a distribution $\sigma \in \mathcal{D}(X)$ and a channel $c\colon X \rightsquigarrow Y$ be given. According to Definition 7.9.1 (2), the Jeffrey update of σ with an evididence distribution $\tau \in \mathcal{D}(Y)$ is given below, following (7.21).



(7.52)

The Pearl update of σ with a factor $q\colon Y \to [0,1]$, considered as channel $Y \rightsquigarrow 2 = \{0,1\}$, is given below, see also (Lavore and Roman, 2023, Defn. 3.28).



(7.53)

The last equation comes from Proposition 7.9.9 (1). We thus see that if we view the predicate q as a channel itself, the mechanisms of Jeffrey and Pearl are similar. This usage of shaded boxes is developed further in Section 7.10.

Exercises

7.9.1 Check for yourself the claimed outcomes in Example 7.9.5:

1 $\omega|_{(\pi_2)^*(\mathbf{1}_e)}[1,0] = \frac{4}{5}|c\rangle + \frac{1}{5}|c^\perp\rangle$;
2 $\omega' := ((\pi_1)^\dagger_\omega)_*(\rho) = \frac{1}{10}|c,e\rangle + \frac{1}{40}|c,e^\perp\rangle + \frac{7}{40}|c^\perp,e\rangle + \frac{7}{10}|c^\perp,e^\perp\rangle$;
3 $\omega'|_{(\pi_2)^*(\mathbf{1}_e)}[1,0] = \frac{4}{11}|c\rangle + \frac{7}{11}|c^\perp\rangle$.

7.9.2 This example is taken from Diaconis and Zabell (1982), where it is attributed to Whitworth: there are three contenders A, B, C for winning a race with a prori distribution $\omega = \frac{2}{11}|A\rangle + \frac{4}{11}|B\rangle + \frac{5}{11}|C\rangle$. Surprising information comes in that A's chances have become $\frac{1}{2}$. What are the adapted chances of B and C?

1 Split up the sample space $X = \{A, B, C\}$ into a suitable two-element partition via a function $f\colon X \to \mathbf{2}$.

2 Use the uniform distribution $unif = \frac{1}{2}|1\rangle + \frac{1}{2}|0\rangle$ on $\mathbf{2}$ and show that Jeffrey's rule gives as adapted distribution:

$$(f_\omega^\dagger)_*(unif) = \tfrac{1}{2} \cdot \omega|_{\mathbf{1}_{\{A\}}} + \tfrac{1}{2} \cdot \omega|_{\mathbf{1}_{\{B,C\}}} = \tfrac{1}{2}|A\rangle + \tfrac{2}{9}|B\rangle + \tfrac{5}{18}|C\rangle.$$

7.9.3 The next illustration is attributed to Jeffrey, and is reproduced for instance in Chan and Darwiche (2005); Darwiche (2009). We consider three colors: green (g), blue (b) and violet (v), which are combined in a space $C = \{g, b, v\}$. These colors apply to cloths, which can additionally be sold or not, as represented by the space $S = \{s, s^\perp\}$. There is a prior joint distribution τ on $C \times S$, namely:

$$\omega = \tfrac{3}{25}|g, s\rangle + \tfrac{9}{50}|g, s^\perp\rangle + \tfrac{3}{25}|b, s\rangle + \tfrac{9}{50}|b, s^\perp\rangle + \tfrac{8}{25}|v, s\rangle + \tfrac{2}{25}|v, s^\perp\rangle.$$

A cloth is inspected by candlelight and the following likelihoods are reported per color: 70% certainty that it is green, 25% that it is blue, and 5% that it is violet.

1 Compute the two marginals $\omega_1 := \omega[1, 0] \in \mathcal{D}(C)$ and $\omega_2 := \omega[0, 1] \in \mathcal{D}(S)$ and show that we can write the joint state ω in two ways as:

$$\langle c,\, id\rangle_*(\omega_2) = \omega = \langle id,\, d\rangle_*(\omega_1)$$

for channels $c := \omega[1, 0 \,|\, 0, 1] \colon S \rightsquigarrow C$ and $d := \omega[0, 1 \,|\, 1, 0] \colon C \rightsquigarrow S$, given by:

$$\begin{cases} c(s) = \frac{3}{14}|b\rangle + \frac{3}{14}|b\rangle + \frac{8}{14}|v\rangle \\ c(s^\perp) = \frac{9}{22}|b\rangle + \frac{9}{22}|b\rangle + \frac{2}{11}|v\rangle \end{cases} \qquad \begin{cases} d(g) = \frac{2}{5}|s\rangle + \frac{3}{5}|s^\perp\rangle \\ d(b) = \frac{2}{5}|s\rangle + \frac{3}{5}|s^\perp\rangle \\ d(v) = \frac{4}{5}|s\rangle + \frac{1}{5}|s^\perp\rangle. \end{cases}$$

These channels c, d are each other's daggers, see Theorem 7.2.1 (2).

2 Capture the above inspection evidence as a predicate q on C and show that Pearl's rule gives:

$$\omega_2|_{c^*(q)} = (\omega|_{q\otimes\mathbf{1}})[0, 1] = d_*(\omega_1|_q) = \tfrac{26}{61}|s\rangle + \tfrac{35}{61}|s^\perp\rangle.$$

3 Describe the evidence now as a state τ on C and show that Jeffrey's rule gives:

$$d_*(\tau) = ((\pi_1)_\omega^\dagger)_*(\tau)[0, 1] = \tfrac{21}{50}|s\rangle + \tfrac{29}{50}|s^\perp\rangle.$$

4 Check also that:

$$\begin{aligned}\langle id, c\rangle_*(\tau) = ((\pi_1)_\omega^\dagger)_*(\tau) = \tfrac{7}{25}|g, s\rangle + \tfrac{21}{50}|g, s^\perp\rangle) + \tfrac{1}{10}|b, s\rangle \\ + \tfrac{3}{20}|b, s^\perp\rangle) + \tfrac{1}{25}|v, s\rangle + \tfrac{1}{100}|v, s^\perp\rangle.\end{aligned}$$

The latter outcome is given in Darwiche (2009).

7.9.4 The following alarm example is in essence due to Pearl Pearl (1988), see Example 6.1.2 (2) and also (Darwiche, 2009, §3.6); we have adapted the numbers in order to make the calculations a little bit easier. There is an 'alarm' set $A = \{a, a^\perp\}$ and a 'burglary' set $B = \{b, b^\perp\}$, with the following a priori joint distribution on $A \times B$.

$$\omega = \tfrac{1}{200}|a, b\rangle + \tfrac{7}{500}|a, b^\perp\rangle + \tfrac{1}{1000}|a^\perp, b\rangle + \tfrac{98}{100}|a^\perp, b^\perp\rangle.$$

Someone reports that the alarm went off, but with only 80% certainty because of deafness.

1 Translate the alarm information into a predicate $p \colon A \to [0, 1]$ and show that crossover updating leads to a burglary distribution:

$$\omega|_{p\otimes\mathbf{1}}[0, 1] = \tfrac{3}{151}|b\rangle + \tfrac{148}{151}|b^\perp\rangle \approx 0.02|b\rangle + 0.98|b^\perp\rangle.$$

2 Compute the extracted channel $c := \omega[1,0\,|\,0,1]\colon B \multimap A$ as in Theorem 7.2.1, and express the answer in the previous item in terms of Pearl's update rule / backward inference using c.

3 Use the Bayesian inversion / dagger $d := c^{\dagger}_{\omega[0,1]}\colon A \multimap B$ of this channel c to calculate the outcome of Jeffrey's update rule as:

$$d_*(\tfrac{4}{5}|a\rangle + \tfrac{1}{5}|a^{\perp}\rangle) \;=\; \tfrac{19639}{93195}|b\rangle + \tfrac{73556}{93195}|b^{\perp}\rangle \;\approx\; 0.21|b\rangle + 0.79|b^{\perp}\rangle.$$

(Notice again the considerable difference in outcomes between Pearl and Jeffrey.)

7.9.5 Recall from Example 7.9.2 the Jeffrey update:

$$\sigma_1 \;:=\; (c^{\dagger}_{\sigma})_*(\rho) \;=\; \tfrac{278}{519}|d\rangle + \tfrac{278}{519}|d^{\perp}\rangle.$$

Suppose we have another evidence state $\rho = \frac{3}{5}|p\rangle + \frac{2}{5}|n\rangle$. Compute:

1 $(c^{\dagger}_{\sigma_1})_*(\rho)$;
2 $\sigma_2 := (c^{\dagger}_{\sigma})_*(\rho)$;
3 $(c^{\dagger}_{\sigma_2})_*(\tau)$.

The first and third items produce different results, which proves Proposition 7.9.7 (4).

7.9.6 Consider a channel $c\colon X \multimap Y$ with prior $\sigma \in \mathcal{D}(X)$ and evidence distribution $\tau \in \mathcal{D}(Y)$. Assume that τ is a point state $1|z\rangle$, for some $z \in Y$. Write $\omega := \langle id, c\rangle_*(\sigma) \in \mathcal{D}(X \times Y)$.

1 Show that the Jeffrey-updated joint state $\omega' := ((\pi_2)^{\dagger}_{\omega})_*(\tau) \in \mathcal{D}(X \times Y)$ is of the form:

$$\omega' \;=\; \sigma|_{c^*(\mathbf{1}_z)} \otimes 1|z\rangle.$$

2 Verify that the double-dagger channel $c' := (c^{\dagger}_{\sigma})^{\dagger}_{\tau}\colon X \multimap Y$ is extremely trivial, namely a 'constant-point' channel, that is, of the form $c'(x) = 1|z\rangle$ for all $x \in X$.

7.9.7 Jeffrey's rule is frequently formulated (notably in Halpern (2003), to which we refer for details) and used (like in Exercise 7.9.2), in situations where the channel involved is deterministic, see Lemma 7.1.4. Consider an arbitrary function $f\colon X \to I$, giving a partition of the set X via subsets $U_i := f^{-1}(i) = \{x \in X \mid f(x) = i\}$. Let $\omega \in \mathcal{D}(X)$ be a prior.

1 Show that applying Jeffrey's rule to a new state of affairs $\rho \in \mathcal{D}(I)$ gives as posterior:

$$(f^{\dagger}_{\omega})_*(\rho) \;=\; \sum_{i \in I} \rho(i) \cdot \omega|_{\mathbf{1}_{U_i}} \qquad \text{satisfying} \qquad f_*((f^{\dagger}_{\omega})_*(\rho)) \;=\; \rho.$$

2 Prove the following minimal-distance result:

$$d((f^{\dagger}_{\omega})_*(\rho), \omega) \;=\; \bigwedge\{d(\omega, \omega') \mid \omega' \in \mathcal{D}(X) \text{ with } f_*(\omega') = \rho\},$$

where d is the total variation distance from Section 4.8.

7.9.8 Prove that for a general, not necessarily deterministic channel $c\colon X \multimap Y$ with prior state $\omega \in \mathcal{D}(X)$ and state $\rho \in \mathcal{D}(Y)$, there is an inequality:

$$d((c^{\dagger}_{\omega})_*(\rho), \omega) \;\leq\; \bigwedge_{\omega' \in \mathcal{D}(X)} d(\omega, \omega') + d(c_*(\omega'), \rho).$$

7.9.9 The following setting is copied from (Diaconis and Zabell, 1982, Ex. 3.2). Consider the set $D = \{a, b, c, d\}$ of four defendants in a criminal case with information about their handedness and gender given by the functions h, g in:

$$\{L, R\} \xleftarrow{\ h\ } D \xrightarrow{\ g\ } \{F, M\}$$

with:

$$\begin{cases} h(a) = h(b) = L \\ h(c) = h(d) = R \end{cases} \qquad \begin{cases} g(a) = g(c) = F \\ g(b) = g(d) = M. \end{cases}$$

There is evidence in the form of distributions about handedness and gender, namely:

$$\sigma = \tfrac{4}{5}|L\rangle + \tfrac{1}{5}|R\rangle \qquad \tau = \tfrac{7}{10}|F\rangle + \tfrac{3}{10}|M\rangle.$$

A uniform prior $\upsilon = \frac{1}{4}|a\rangle + \frac{1}{4}|b\rangle + \frac{1}{4}|c\rangle + \frac{1}{4}|d\rangle$ is used for the defendants.

1 Show that in this special case the order of doing Jeffrey updates does not matter, that is:

$$\upsilon_{HG} = \tfrac{14}{25}|a\rangle + \tfrac{6}{25}|b\rangle + \tfrac{7}{50}|c\rangle + \tfrac{3}{50}|d\rangle = \upsilon_{GH},$$

where:

$$\begin{cases} \upsilon_{HG} := (h^{\dagger}_{\upsilon_G})_*(\sigma) \\ \upsilon_{GH} := (g^{\dagger}_{\upsilon_H})_*(\tau). \end{cases} \qquad \text{with} \qquad \begin{cases} \upsilon_G := (g^{\dagger}_{\upsilon})_*(\tau) \\ \upsilon_H := (h^{\dagger}_{\upsilon})_*(\sigma). \end{cases}$$

2 Show also that in this situation:

$$g_*(\upsilon_{HG}) = \tau \qquad \text{and} \qquad h_*(\upsilon_{GH}) = \sigma.$$

According to (Diaconis and Zabell, 1982, Thm. 3.1) these two equations charachterise commutation of two Jeffrey updates (along deterministic channels).

3 Check that the distribution $\upsilon_{HG} = \upsilon_{GH} \in \mathcal{D}(D)$ from the first point can also be obtained as a single Jeffrey update via the (deterministic) tuple channel $\langle h, g\rangle \colon D \to \{L, R\} \times \{F, M\}$, namely as:

$$\Big(\langle h, g\rangle^{\dagger}_{\upsilon}\Big)_*(\sigma \otimes \tau).$$

7.9.10 In Theorem 7.9.3 we assumed a distribution $\sigma \in \mathcal{D}(X)$ and a channel $c \colon X \rightsquigarrow Y$. Equivalently, we can start from a joint distribution $\omega \in \mathcal{D}(X \times Y)$, via the correspondence (7.11). In this Exercise we rephrase and reconsider Theorem 7.9.3 in terms such a joint distribution ω — instead of σ, c.

So let $\omega \in \mathcal{D}(X \times Y)$ be an arbitrary distribution. We write $d \colon Y \rightsquigarrow X$ as shorthand for the extracted channel $d = \omega[1, 0 \,|\, 0, 1]$.

1 Show that the Pearl update with a factor $q \in \mathit{Fact}(Y)$ corresponds to the cross-over update and pushforward of the updated marginal:

$$\omega|_{\mathbf{1}\otimes q}[1, 0] = d_*\Big(\omega[0, 1]\big|_q\Big).$$

2 Show that the Jeffrey update of the joint distribution $\omega \in \mathcal{D}(X \times Y)$ with a distribution $\tau \in \mathcal{D}(Y)$ corresponds to the pushforward:

$$d_*(\tau).$$

3 Let's write $\Pi_\tau := \{\rho \in \mathcal{D}(X \times Y) \mid \rho[0,1] = \tau\}$, for $\tau \in \mathcal{D}(Y)$. Write:

$$\omega' \;:=\; \langle d, id\rangle_*(\tau) \;\in\; \mathcal{D}(X \times Y).$$

Clearly, $\omega' \in \Pi_\tau$. Prove that $\omega' \in \Pi_\tau$ is least divergent from ω, that is:

$$\omega' \;\in\; \operatorname*{argmin}_{\rho \in \Pi_\tau} D_{KL}(\rho, \omega).$$

The name 'information-projection' or just 'I-projection' is used for ω', in the area of information geometry (see *e.g.* Csiszár (1975)).
Since $\omega'[1,0] = \sigma_J$ by the previous point, where $\sigma_J \in \mathcal{D}(X)$ is the Jeffrey update from Theorem 7.9.3 (2), this gives a unique characterisation of the Jeffrey update, namely as the first marginal of the distribution $\omega' \in \Pi_\tau$ with minimal divergence from $\omega = \langle id, c\rangle_*(\sigma)$. This observation occurs in Pan et al. (2006), without a proof. The characterisation can be adapted to the Pearl update, using the distribution $\omega[0,1]\big|_q$ instead of τ, see the first point of this exercise — or Proposition 7.9.9 (1).

7.9.11 In Figure 7.6 we have implicitly used a dagger in WebPPL. Check for yourself that it can be defined abstractly as the following (higher-order) function:

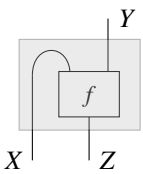
```
var dagger = function(chan, prior) {
    var dag_chan = function(z) { Infer(function() {
       var x = sample(prior);
       var y = sample(chan(x));
       condition( y == z );
       return x })}
    return dag_chan }
```

Refactor the code of the function `jeffrey_update` in Figure 7.6 in terms of this dagger.

7.10 Disintegration of Bayesian Networks

At the end of Section 7.3, in Remark 7.3.4, we have introduced the shaded box notation for disintegration. The shaded area describes the block that is being normalised, in order to produce a new channel. It turns out that there are rules for moving channels out of such normalisation boxes, leading to a more compositional approach to disintegration. In this section we shall first formulate some of these properties for shaded boxes. Then we shall apply them to conditioning queries for Bayesian networks in a compositional manner, building on Jacobs et al. (2025).

First we try to get a better grip on what precisely happens in the bending of arrows and shading in (7.20), as expressed in a disintegration diagram:



(7.54)

This involves normalisation, from a subdistribution to a proper distribution. Such normalisation may fail if the subdistribution involved is the zero distribution. We shall be fairly relaxed about

such side-conditions in this section and simply assume that the relevant normalisations are well-defined.

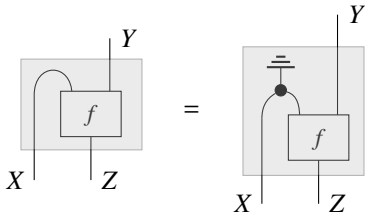
We can make more precise what happens in (7.54) in terms of comparators, as in Lavore and Roman (2023).

Definition 7.10.1 For a set X we define the comparator subdistribution channel *cpr* in:

$$\begin{array}{rcl} X \times X & \xrightarrow{\quad cpr \quad} & \mathcal{D}_{\leq 1}(X) \\ (x, x') & \longmapsto & \begin{cases} 1|x\rangle & \text{if } x = x' \\ \mathbf{0} & \text{otherwise.} \end{cases} \end{array}$$

In a diagram we shall write such a comparator as , dually to copier . In fact, the copier and comparator are closely connected, as Frobenius structures, see Exercise 7.10.2.

Writing this comparator as suggests that it is a string diagram, but strictly speaking it is not, since string diagrams as introduced in Section 2.5 involve only probabilistic $\mathcal{D}$-channels, and not $\mathcal{D}_{\leq 1}$-channels. We allow ourselves a bit of freedom here, but immediately normalise from $\mathcal{D}_{\leq 1}$ and $\mathcal{D}$ via the shaded box. Using these copiers we can now reformulate the bending in (7.54) as:



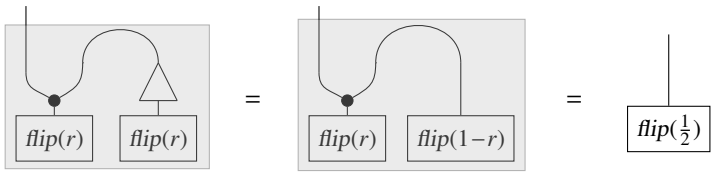
The shading on the right-hand-side can now be understood as normalisation. We have reduced disintegration to normalisation, via the comparator subdistribution channel. Indeed, if we first compute what is inside the box we get:

$$\begin{aligned} \Big(\big(\,\overline{\top} \otimes id\big) \circ (cpr \otimes id) \circ (id \otimes f)\Big)(x, z) &= \sum_{y \in Y} \sum_{x', x'' \in X} cpr(x, x')(x'') \cdot f(z)(x', y) \,\big|y\big\rangle \\ &= \sum_{y \in Y} f(z)(x, y) \,\big|y\big\rangle. \end{aligned}$$

Normalisation, via frequentist learning *Flrn*, thus yields the disintegration of f in (7.54):

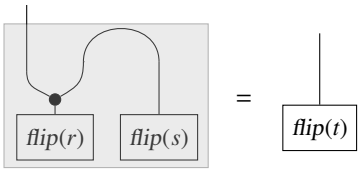
$$Flrn\Big(\big(\,\overline{\top} \otimes id\big) \circ (cpr \otimes id) \circ (id \otimes f)\Big)(x, z) = \sum_{y \in Y} \frac{f(z)(x, y)}{\sum_{y'} f(z)(x, y')} \,\big|y\big\rangle.$$

Example 7.10.2 John Von Neumann invented a method to obain a fair coin from a coin with arbitrary bias. The description of this trick from Piedeleu et al. (2025) can be reformulated in terms of a cap inside a shaded normalisation box as:



This involves for an arbitrary bias $r \in (0, 1)$ the coin distribution $flip(r) = r|1\rangle + (1 - r)|0\rangle$.

The triangle is the (deterministic) negation channel *nc* from Exercise 2.5.9, sending $i \in \{0, 1\}$ to $1|1-i\rangle$. The last equality is an instance of the following more general equation, for $r, s \in (0, 1)$.



where $t = \dfrac{r \cdot s}{r \cdot s + (1-r) \cdot (1-s)}$.

The parallel (tensor) product of flips takes the form:

$$flip(r) \otimes flip(s) \;=\; r \cdot s\big|11\rangle + r \cdot (1-s)\big|10\rangle + (1-r) \cdot s\big|01\rangle + (1-r) \cdot (1-s)\big|00\rangle.$$

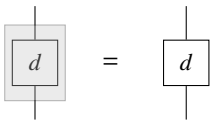
The comparator selects the entries with equal elements and turns $|ii\rangle$ into $|i\rangle$. This gives the subdistribution $r \cdot s|1\rangle + (1-r) \cdot (1-s)|0\rangle$, which normalises to $flip(t)$ as desired.

Now that we understand shaded areas as normalisation scopes, we can study them as such. Normalisation is a peculiar operation that is typically highly non-compositional. Hence these shaded boxes do not seem to match well with the formalism of string diagrams. However, gradually it became clear that there are useful compositional rules for these shaded (normalisation) boxes, see Lemma 7.10.3 below. This was realised by several authors, see for instance Lorenz and Tull (2023); Fritz et al. (2024); Tull et al. (2024) (and also more implicitly in Fritz (2020); Lavore and Roman (2023)). The removal rules for shaded boxes in Lemma 7.10.4 comes from Jacobs et al. (2025).

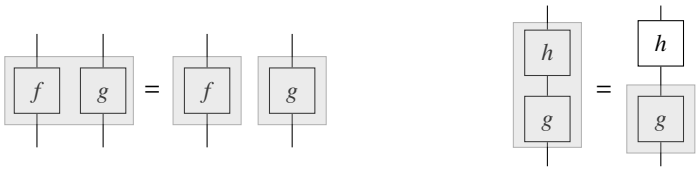
We start with the properties about sequential and parallel composition. Our aim is to restrict the shading / normalisation to the relevant areas, basically by making the shaded area as small as possible.

Lemma 7.10.3 *Let c be an $\mathcal{M}_*$-channel that can be normalised pointwise to a $\mathcal{D}$-channel $\mathsf{Flrn} \circ c$ via frequentist learning Flrn. Below, d and f are properly typed probabilistic $\mathcal{D}$-channels.*

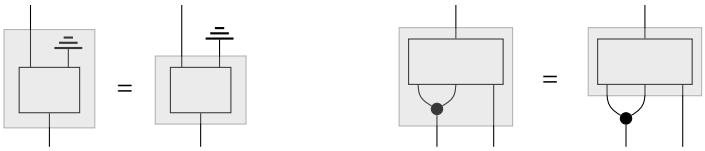
1 *The normalisation of a $\mathcal{D}$-channel d is d itself:*



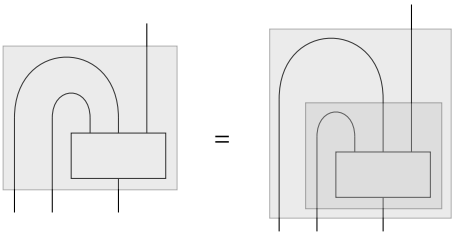
2 *Normalisation of parallel composition can be done separately, and a (proper) channel h can be pulled out of sequential composition, as in:*



3 *Discarders and copiers can be pulled out of normalisation boxes in the following manner.*



4 *(From (Fritz, 2020, Lem. 11.11)) A disintegration of multiple wires can be done as separate disintegrations:*



Proof. 1 Obvious.

2 For the equation on the left, let $f\colon X \to \mathcal{D}_{\leq 1}(Y)$ and $g\colon A \to \mathcal{D}_{\leq 1}(B)$. Then:

$$\begin{aligned}
\Big(\mathit{Flrn} \circ (f \otimes g)\Big)(x,a) &= \sum_{y \in Y\, b \in B} \frac{(f \otimes g)(x,a)(y,b)}{\sum_{y,b} (c \otimes d)(x,a)(y,b)} \big|y,b\big\rangle \\
&= \sum_{y \in Y\, b \in B} \frac{f(x)(y) \cdot g(a)(b)}{\sum_y c(x)(y) \cdot \sum_b d(a)(b)} \big|y,b\big\rangle \\
&= \sum_{y \in Y\, b \in B} \mathit{Flrn}(f(x))(y) \cdot \mathit{Flrn}(g(a)(b)) \big|y,b\big\rangle \\
&= \Big((\mathit{Flrn} \circ f) \otimes (\mathit{Flrn} \circ g)\Big)(x,a).
\end{aligned}$$

For the equation on the right, let $g\colon X \to \mathcal{D}_{\leq 1}(Y)$ and $h\colon Y \to \mathcal{D}(Z)$. Then:

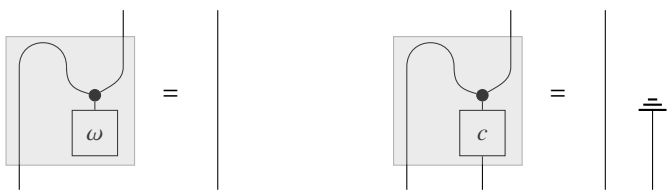
$$\begin{aligned}
\Big(\mathit{Flrn} \circ (h \odot g)\Big)(x) &= \sum_{z \in Z} \mathit{Flrn}\Big((h \odot g)(x)\Big)(z) \big|z\big\rangle \\
&= \sum_{z \in Z} \left(\sum_{y \in Y} \frac{h(y)(z) \cdot g(y)(x)}{\sum_{y,z} h(y)(z) \cdot g(x)(y)} \right) \big|z\big\rangle \\
&= \sum_{z \in Z} \left(\sum_{y \in Y} \frac{h(y)(z) \cdot g(y)(x)}{\sum_y g(x)(y)} \right) \big|z\big\rangle \\
&= \sum_{z \in Z} \left(\sum_{y \in Y} h(y)(z) \cdot \mathit{Flrn}(g(x))(y) \right) \big|z\big\rangle \;=\; \Big(h \odot (\mathit{Flrn} \circ g)\Big)(x).
\end{aligned}$$

3 The discard case follows from the previous point, since it involves pulling out a projection channel. The copy case trivially holds.

4 By instantiating the relevant extractions several times. ❑

We now come to two crucial rules from Jacobs et al. (2025) that allow us to make shaded boxes disappear.

Lemma 7.10.4 *Consider sets X, Y, where X is finite. For a distribution $\omega \in \mathcal{D}_{\mathrm{fs}}(X)$ with full support and a channel $c\colon X \to \mathcal{D}_{\mathrm{fs}}(Y)$ with pointwise full support, one has:*



(7.55)

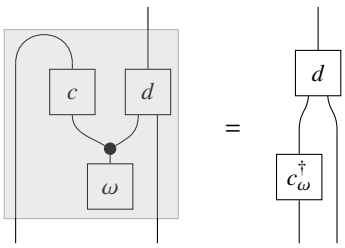
Proof. The equation on the left is a special case of the one on the right, so we prove the latter. Let's write $d := ((\overset{\circ}{\top} \circ cpr) \otimes id) \circ (id \otimes (\Delta \circ c))\colon Y \times X \to \mathcal{D}_{\leq 1}(Y)$ for the subdistribution channel inside the shaded box. Then:

$$\big(Flrn \circ d\big)(y, x) = \sum_{y' \in Y} \frac{(\Delta \circ c)(x)(y, y')}{\sum_{y''} (\Delta \circ c)(x)(y, y'')} \big|y'\rangle = \frac{c(x)(y)}{c(x)(y)} \big|y\rangle = 1|y\rangle = (id \otimes \overset{\circ}{\top})(y, x).$$

The full support requirement guarantees that $c(x)(y) \neq 0$. ❑

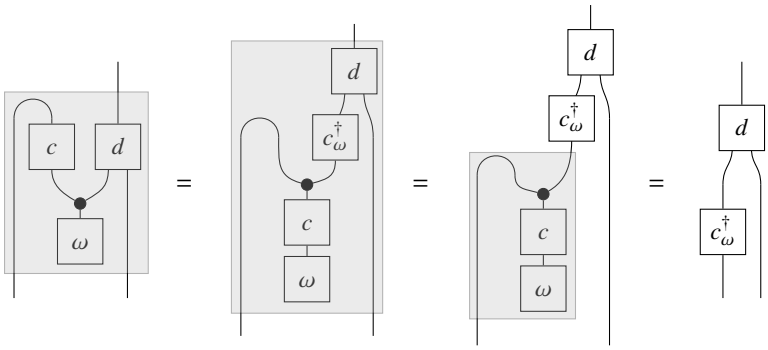
The advantage of these rules for shaded boxes is that we can use them to simplify situations and, hopefully, in the end, get rid of the shaded boxes, resulting in an ordinary string diagram. We shall see several illustrations below, including ones that we have seen before, without shaded boxes. We start with a simple one

Example 7.10.5 Consider a distribution $\omega \in \mathcal{D}(X)$ with two channels $c\colon X \to \mathcal{D}(Y)$ and $d\colon X \times U \to \mathcal{D}(Z)$. There is then an equation:



(7.56)

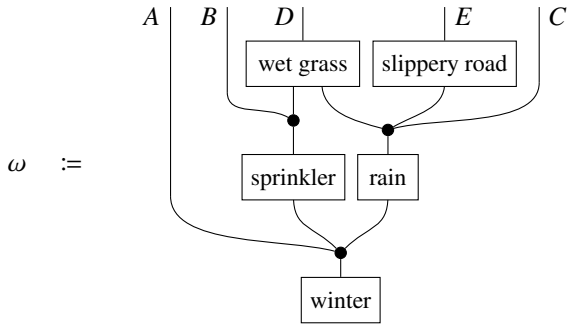
We can reason diagrammatically, first using the defining equation (7.3) for the dagger, and then Lemma 7.10.3 (2) for pulling out the d channel and its input wire on the right. The result then immediately follows from the previous lemma:



We now illustrate how to apply the above rules for shaded boxes to reason about conditioning in a Bayesian network, via what we call 'disintegration of Bayesian networks'.

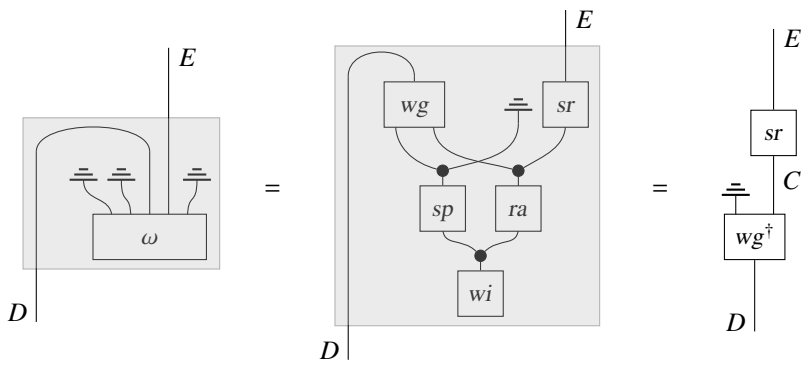
Example 7.10.6 Recall the wetness Bayesian network serving as leading example in Subsec-

tion 2.5.2. It is described below as string diagram, with all nodes accessible from the outside.



We shall use the notation and abbrevations as in Subsection 2.5.2, so $A = \{a, a^{\perp}\}$ and wg is the 'wet grass' channel, *etc.*

In this example we are interested in extracting a channel $D \rightsquigarrow E$ via disintegration, so that we find the probability of slipperiness of the road, given wetness of the grass (or not). We extract this channel $D \rightsquigarrow E$ by bending the D line around and discarding all outgoing wires except E from the above joint distribution ω, as on the left below. We then massage the diagram in such a shape that we can apply the rewrite rule for shaded boxes from Equation (7.56).



We have omitted the prior distribution for the dagger $wg^{\dagger}$ on the right-hand-side, for readability. This prior is $\rho := \langle sp, ra \rangle_*(wi)$. The marginalised dagger $\left(wg^{\dagger}_{\rho}\right)[0,1] \colon D \rightsquigarrow E$ can be expressed as a conditional probability table, of the form:

slipperiness (D)	P(rain (C))
d	$4349/6995$
$d^{\perp}$	$851/3005$

Thus, the disintegrated Bayesian network is itself a Bayesian network, given by two subsequent boxes (channels), above on the right-hand-side. It is a Bayesian network with inputs, called an 'open (causal) model' in Lorenz and Tull (2023). Feeding a distribution about the slipperiness of the road as input to this disintegrated Bayesian network allows us to compute a resulting grass wetness distribution. This can be seen as an application of Jeffrey's update rule. For instance, if we are 75% sure that the grass is wet, then we induce that the road is slippery with 38% probability. This can now be done via a prediction (pushforward) using the disintegrated Bayesian

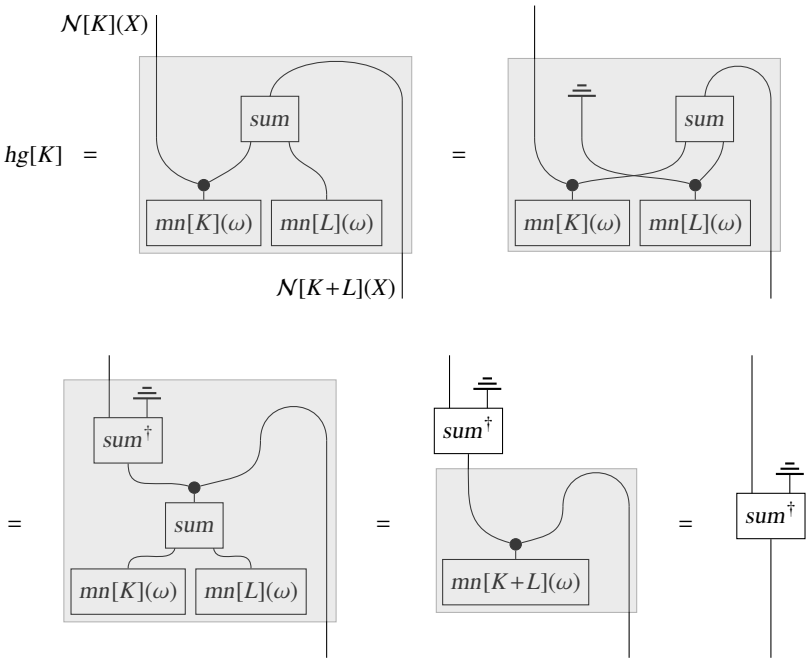


Figure 7.7 Graphical rewriting of the hypergeometric distribution as a marginal of a dagger of (multiset) sum channel, see Example 7.10.7 (1).

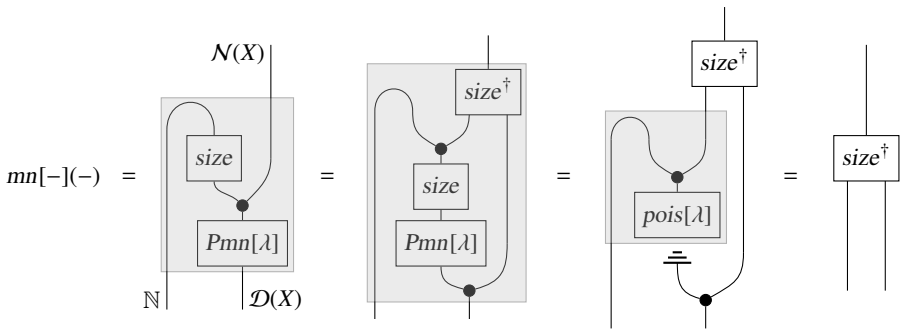


Figure 7.8 Graphical rewriting of multinomial distribution as a a dagger of a size channel, see Example 7.10.7 (2).

network:

$$\left(sr \circ (wg_\rho^\dagger)[0,1]\right)_*\left(\tfrac{3}{4}|d\rangle + \tfrac{1}{4}|d^\perp\rangle\right) = \tfrac{15805643}{42039950}|e\rangle + \tfrac{26234307}{42039950}|e^\perp\rangle \approx 0.38|e\rangle + 0.62|e^\perp\rangle.$$

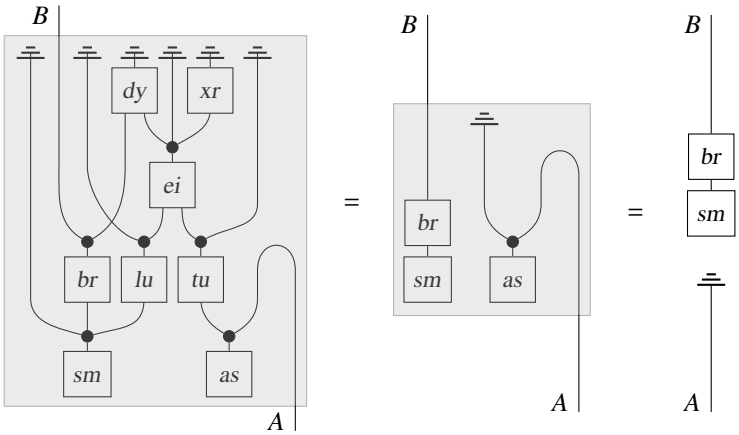
Examples 7.10.7 We redescribe in graphical form two earlier formulations of the hypergeometric and multinomial channel via updating.

1 We recall from Example 6.3.3 (3) how the hypergeometric distribution can be obtained via updating of two multinomial distributions. We can now give a graphical representation of this fact, see Figure 7.7. For readability, the prior of the dagger $sum^{\dagger}$ is ommitted, but this prior is the parallel product $mn[K](\omega) \otimes mn[L](\omega)$. The one-but-last equation uses that multinomial distributions are closed under convolution, see Exercise 3.3.10, but this is not needed in order to apply the shaded box removal.

2 In Example 6.3.3 (4) we saw how to obtain the multinomial distribution via updating of multiple Poisson distributions. The graphical description is in Figure 7.8. The $\overline{+}$ symbol arises via Lemma 3.10.4 (2). The final step uses the shaded box removal from Lemma 7.10.4. The dagger of the size channel uses a channel as prior, as in Definition 7.3.5. This prior is the Poisson multinomial channel $Pmn[\lambda] \colon \mathcal{D}(X) \rightsquigarrow \mathcal{N}(X)$. The dagger channel on the right in Figure 7.8 is thus, in explicit form, $size^{\dagger}_{Pmn[\lambda]} \colon \mathbb{N} \times \mathcal{D}(X) \rightsquigarrow \mathcal{N}(X)$. This dagger can be recognised on the right-hand-side of the equation in Example 6.3.3 (4).

Another example of a dagger with a channel as prior arises in the Asia Bayesian network that we have used earlier for conditioning queries.

Example 7.10.8 We reconsider the Asia Bayesian network from Section 6.4, described in accessible form in Figure 6.10. We look at the probability of having bronchitis, given a trip to Asia, now from a disintegration perspective, using obvious abbreviations for the channel names. Via graphical reasoning we quickly obtain that the output does not depend on the input:



The first equation arises from the annihilating power of discarding $\overline{+}$. The second equation uses a combination of Lemma 7.10.3 (2). We conclude that being in Asia does not influence the bronchitis probability. This is sometimes formulated as: A is not in the Markov blanket of B.

In Section 6.4 we have looked at the probability of lung cancer, given dyspnoea and tuberculosis. The inference via disintegration of the Bayesian network is done in Figure 7.9. The first line only involves discarding, but the second contains more action: first, Lemma 7.10.3 (4) is used to separate the two bendings, then Lemma 7.10.3 (2) is allows us to minimise the inner shaded box, which is then removed via Lemma 7.10.4. The last line in Figure 7.9 massages the situation in such a form that the dagger (7.23) appears. The diagram on the lower left shows that it does require a bit of puzzling to recognise the relevant network structure.

The final dagger box, as channel $D \times T \rightsquigarrow L$, can be formulated as a conditional probability table:

dyspnoea	tuberculosis	P(lung cancer)
d	t	0.0558
d	$t^{\perp}$	0.12
$d^{\perp}$	t	0.0506
$d^{\perp}$	$t^{\perp}$	0.0129

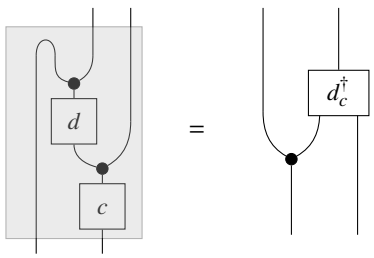
The first entry corresponds to what we have computed in Section 6.4 in a more *ad hoc* manner. As in Example 7.10.6 we can now do inference via prediction.

In Section 6.4 we have seen various examples of inference (computation of posteriors) in Bayesian networks, via updating with separate point predicates. As the examples in this section illustrate, the network disintegration approach offfers advantages.

- Disintegration gives a new "open" network that can be interpreted as a channel and that computes posteriors not only for single points but also for distributions of points.
- Certain often used disintegrations can be pre-computed; this could give an inference speed-up, especially for larger Bayesian networks.

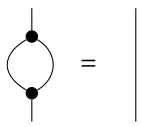
Exercises

7.10.1 Prove the following equation, either by diagrammatic or by equational reasoning (or both).

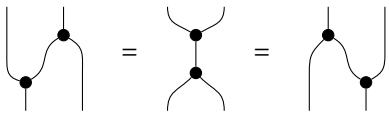


7.10.2 Show that the comparator map $cpr\colon X \times X \to \mathcal{D}_{\leq 1}(X)$ from Definition 7.10.1 satisfies the following properties (as in (Lavore and Roman, 2023, Fig. 12)).

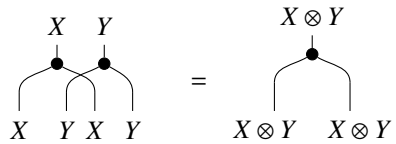
1 An equality of $\mathcal{D}$-channels:



2 An equality of $\mathcal{M}$-channels capturing the Frobenius equations:

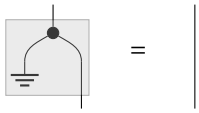


3 Another equality of $\mathcal{M}$-channels for proper interaction of comparators and tensors:



4 Two more equalities of $\mathcal{M}$-channels, for commutativity and associativity:

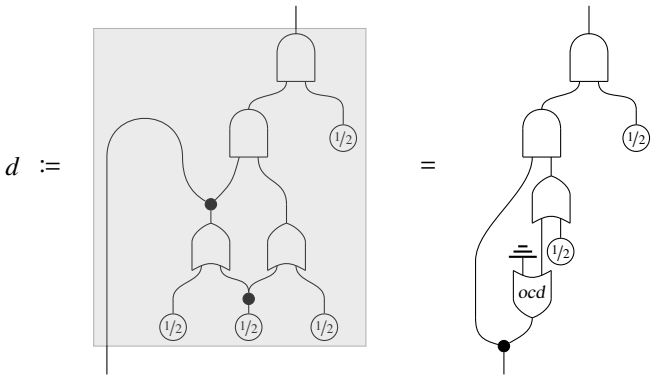
5 When X is a finite set, an equality of $\mathcal{D}$-channels:



7.10.3 Check Equation (7.24).

7.10.4 Recall the fault trees from Exercise 2.5.9. We modify the first example given there and extend it with a copy so that we can perform an update.

1 Show that the channel $d\colon \mathbf{2} \rightsquigarrow \mathbf{2}$ defined as on the left below, satisfies the equation on the right:



The upside-down or-gate is used for the or-channel-dagger $ocd := oc^{\dagger}_{\sigma\otimes\sigma}\colon \mathbf{2} \rightsquigarrow \mathbf{2}\times\mathbf{2}$, where $\sigma = \mathit{flip}(1/2)$ and where $oc\colon \mathbf{2}\times\mathbf{2} \rightsquigarrow \mathbf{2}$ is the or-channel defined in Exercise 2.5.9.

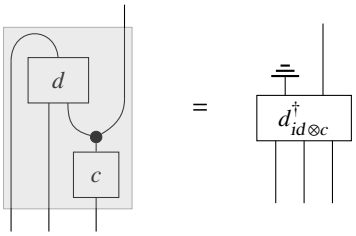
2 Check that this dagger ocd of the or-channel satisfies:

$$ocd(0) \;=\; 1|0,0\rangle \qquad \text{and} \qquad ocd(1) \;=\; \tfrac{1}{3}|1,1\rangle + \tfrac{1}{3}|1,0\rangle + \tfrac{1}{3}|0,1\rangle.$$

3 Use this description to calculate:

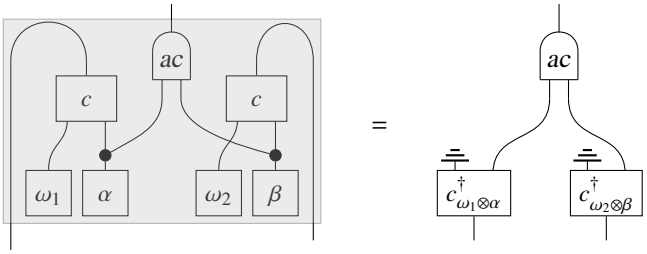
$$d(0) \;=\; 1|0\rangle \qquad \text{and} \qquad d(1) \;=\; \tfrac{5}{12}|1\rangle + \tfrac{7}{12}|0\rangle.$$

7.10.5 Show that:



7.10.6 Recall the Bayesian network from Exercise 6.4.4.

1 Show that disintegration of this network can be described as follows.



The daggers may be described in an alternative manner, as in Exercise 7.1.11.

2 Check that the channel $\mathbf{2}\times\mathbf{2} \twoheadrightarrow \mathbf{2}$ on the right-hand-side works as follows.

$$(1,1) \longmapsto \tfrac{289}{400}|1\rangle + \tfrac{111}{400}|1\rangle \qquad (0,1) \longmapsto \tfrac{51}{400}|1\rangle + \tfrac{349}{400}|1\rangle$$
$$(1,0) \longmapsto \tfrac{51}{400}|1\rangle + \tfrac{349}{400}|1\rangle \qquad (0,0) \longmapsto \tfrac{9}{400}|1\rangle + \tfrac{391}{400}|1\rangle.$$

The first outcome matches the one from Exercise 6.4.4.

7.11 Categorical aspects of Bayesian inversion

As mentioned in the beginning of this chapter the 'dagger' of a channel — *i.e.* its Bayesian inversion — can also be described categorically. It turns out to be a special 'dagger' functor. Such reversal is quite common for non-deterministic computation, see Example 7.11.1 below. The fact that this same abstract structure exists for probabilistic computation demonstrates once again that Bayesian inversion is a canonical operation — and that category theory provides a useful language for making such similarities explicit.

This final section of the chapter goes a bit deeper into the category-theoretic aspects of probabilistic computation in general and of Bayesian inversion in particular. It is not essential for the rest of this book, but provides deeper insight into the underlying structures.

Example 7.11.1 Recall the category $\mathbf{Chan}(\mathcal{P})$ of non-deterministic computations. Its objects are sets X and its morphisms $f\colon X \twoheadrightarrow Y$ are functions $f\colon X \to \mathcal{P}(Y)$. The identity morphism $\mathit{unit}_X\colon X \twoheadrightarrow X$ in $\mathbf{Chan}(\mathcal{P})$ is the singleton function $\mathit{unit}(x) = \{x\}$. Composition of $f\colon X \twoheadrightarrow Y$ and $g\colon Y \twoheadrightarrow Z$ is the function $g \circ f\colon X \twoheadrightarrow Z$ given by:

$$(g \circ f)(x) = \{z \in Z \mid \exists y \in Y.\, y \in f(x) \text{ and } z \in g(y)\}.$$

It is not hard to see that $\circ$ is associative and has *unit* as identity element. In fact, this has already been proven more generally, in Lemma 1.11.3.

There are two aspects of the category $\mathbf{Chan}(\mathcal{P})$ that we wish to illustrate, namely (1) that it has an 'inversion' operation, in the form of a dagger functor, and (2) that it is isomorphic to the category **Rel** of sets with relations between them (as morphisms). Probabilistic analogues of these two points will be described later.

1 We start from a very basic observation, namely that morphisms in $\mathbf{Chan}(\mathcal{P})$ can be reversed. There is a bijective correspondence, indicated by the double lines, between morphisms in $\mathbf{Chan}(\mathcal{P})$:

$$\frac{X \longrightarrow\!\!\!\!\!\!\circ\ Y}{Y \longrightarrow\!\!\!\!\!\!\circ\ X} \qquad \text{that is, between functions:} \qquad \frac{X \xrightarrow{f} \mathcal{P}(Y)}{Y \xrightarrow[g]{} \mathcal{P}(X)}$$

This correspondence sends $f\colon X \to \mathcal{P}(Y)$ to the function $f^\dagger\colon Y \to \mathcal{P}(X)$ with $f^\dagger(y) := \{x \mid y \in f(x)\}$. Hence $y \in f(x)$ iff $x \in f^\dagger(y)$. Similarly one sends $g\colon Y \to \mathcal{P}(X)$ to $g^\dagger\colon X \to \mathcal{P}(Y)$ via $g^\dagger(x) := \{y \mid x \in g(y)\}$. Clearly, $f^{\dagger\dagger} = f$ and $g^{\dagger\dagger} = g$.

It turns out that this dagger operation $(-)^\dagger$ interacts nicely with composition: one has $\mathit{unit}^\dagger = \mathit{unit}$ and also $(g \circ f)^\dagger = f^\dagger \circ g^\dagger$. This means that the dagger is functorial. It can be described as a functor $(-)^\dagger\colon \mathbf{Chan}(\mathcal{P}) \to \mathbf{Chan}(\mathcal{P})^{\mathrm{op}}$, which is the identity on objects: $X^\dagger = X$. The opposite $(-)^{\mathrm{op}}$ category is needed for this functor since it reverses arrows.

2 We write **Rel** for the category with sets X as objects and relations $R \subseteq X \times Y$ as morphisms $X \to Y$. The identity $X \to X$ is given by the equality relation $Eq_X \subseteq X \times X$, with $Eq_X = \{(x, x) \mid x \in X\}$. Composition of $R \subseteq X \times Y$ and $S \subseteq X \times Z$ is the 'relational' composition $S \bullet R \subseteq X \times Z$ given by:

$$S \bullet R := \{(x, z) \mid \exists y \in Y.\, R(x, y) \text{ and } S(y, z)\}.$$

It is not hard to see that we get a category in this way.

There is a 'graph' functor $G\colon \mathbf{Chan}(\mathcal{P}) \to \mathbf{Rel}$, which is the identity on objects: $G(X) = X$. On a morphism $f\colon X \to\!\!\!\!\circ\ Y$, that is, on a function $f\colon X \to \mathcal{P}(Y)$, we define $G(f) \subseteq X \times Y$ to be $G(f) = \{(x, y) \mid x \in X, y \in f(x)\}$. Then: $G(\mathit{unit}_X) = Eq_X$ and $G(g \circ f) = G(f) \bullet G(f)$.

In the other direction there is also a functor $F\colon \mathbf{Rel} \to \mathbf{Chan}(\mathcal{P})$, which is again the identity on objects: $F(X) = X$. On a morphism $R\colon X \to Y$ in **Rel**, that is, on a relation $R \subseteq X \times Y$, we define $F(R)\colon X \to \mathcal{P}(Y)$ as $F(R)(x) = \{y \mid R(x, y)\}$. This F preserves identities and composition.

These two functors G and F are each other's inverses, in the sense that:

$$F \circ G = \mathit{id}\colon \mathbf{Chan}(\mathcal{P}) \to \mathbf{Chan}(\mathcal{P}) \qquad \text{and} \qquad G \circ F = \mathit{id}\colon \mathbf{Rel} \to \mathbf{Rel}.$$

This establishes an isomorphism $\mathbf{Chan}(\mathcal{P}) \cong \mathbf{Rel}$ of categories.

Interestingly, **Rel** is also a dagger category, via the familar operation of reversal of relations: for $R \subseteq X \times Y$ one can form $R^\dagger \subseteq Y \times X$ via $R^\dagger(y, x) = R(x, y)$. This yields a functor $(-)^\dagger\colon \mathbf{Rel} \to \mathbf{Rel}^{\mathrm{op}}$, obviously with $(-)^{\dagger\dagger} = \mathit{id}$.

Moreover, the above functors G and F commute with the daggers of $\mathbf{Chan}(\mathcal{P})$ and **Rel**, in the sense that:

$$G(f^\dagger) = G(f)^\dagger \qquad \text{and} \qquad F(R^\dagger) = F(R)^\dagger.$$

We shall prove the first equation and leave the second one to the interested reader. The proof is

obtained by carefully unpacking the right definition at each stage. For a function $f\colon X \to \mathcal{P}(Y)$ and elements $x \in X$, $y \in Y$,

$$G(f^\dagger)(y,x) \iff x \in f^\dagger(y) \iff y \in f(x) \iff G(f)(x,y) \iff G(f)^\dagger(y,x).$$

We now move from non-deterministic to probabilistic computation. Our aim is to obtain analogous results, namely inversion in the form of a dagger functor on a category of probabilistic channels, and an isomorphism of this category with a category of probabilistic relations. One may expect that these results hold for the category $\mathbf{Chan}(\mathcal{D})$ of probabilistic channels. But the situation is a bit more subtle. Recall from the previous section that the dagger (Bayesian inversion) $c^\dagger_\omega\colon Y \twoheadrightarrow X$ of a probabilistic channel $c\colon X \to Y$ requires a 'prior' distribution $\omega \in \mathcal{D}(X)$ on the domain — with side-condition that $c_*(\omega)$ has full support. In order to conveniently deal with this situation we incorporate these distributions ω into the objects of our category. We follow Clerc et al. (2017) and denote this category as **Krn**; its morphisms are 'kernels'.

Definition 7.11.2 The category **Krn** of kernels has:

- objects: pairs (X,σ) where X is a finite set and $\sigma \in \mathcal{D}(X)$ is a distribution on X with full support;
- morphisms: $f\colon (X,\sigma) \to (Y,\tau)$ are probabilistic channels $f\colon X \twoheadrightarrow Y$ with $f_*(\sigma) = \tau$.

Identity maps $(X,\sigma) \to (X,\sigma)$ in **Krn** are identity channels $\mathit{unit}\colon X \twoheadrightarrow X$, given by $\mathit{unit}(x) = 1|x\rangle$, which we write simply as id. Composition in **Krn** is ordinary composition $\circ$ of channels.

Theorem 7.11.3 *Bayesian inversion forms a dagger functor* $(-)^\dagger\colon \mathbf{Krn} \to \mathbf{Krn}^{\mathrm{op}}$ *which is the identity on objects and which sends:*

$$\Big((X,\sigma) \xrightarrow{f} (Y,\tau)\Big) \longmapsto \Big((Y,\tau) \xrightarrow{f^\dagger_\sigma} (X,\sigma)\Big)$$

This functor is its own inverse: $f^{\dagger\dagger} = f$.

Proof. We first have to check that the dagger functor is well-defined, *i.e.* that the above mapping yields another morphism in **Krn**.

$$(f^\dagger_\sigma)_*(\tau) = (f^\dagger_\sigma)_*(f_*(\sigma)) \overset{(7.4)}{=} \sigma.$$

Aside: this does *not* mean that $f^\dagger \circ f = \mathit{id}$.

Identities and composition in **Krn** are preserved by Lemma 7.1.8 (3) and (2). Finally we have $(f^\dagger_\sigma)^\dagger_\tau = f$ by Lemma 7.1.8 (4). ❑

We now turn to probabilistic relations, with the goal of finding a category of such relations that is isomorphic to **Krn**. For this purpose we use *couplings*. For more information on such couplings, see *e.g.* Barthe and Hsu (2021).

Definition 7.11.4 We introduce a category **Cpl** of couplings with the same objects as **Krn**. A morphism $(X,\sigma) \to (Y,\tau)$ in **Cpl** is a joint state $\varphi \in \mathcal{D}(X \times Y)$ with $\varphi[1,0] = \sigma$ and $\varphi[0,1] = \tau$. Such a distribution which marginalises to σ and τ is called a coupling between σ and τ, see Definition 3.1.5.

Composition of $\varphi\colon (X,\sigma) \to (Y,\tau)$ and $\psi\colon (Y,\tau) \to (Z,\rho)$ is the distribution $\psi \bullet \varphi \in \mathcal{D}(X \times Z)$ defined as:

$$\psi \bullet \varphi \coloneqq \langle \varphi[1,0\,|\,0,1], \psi[0,1\,|\,1,0]\rangle_*(\tau) = \sum_{x\in X,\, z\in Z} \left(\sum_{y\in Y} \frac{\varphi(x,y)\cdot\psi(y,z)}{\tau(y)} \right) |x,z\rangle. \tag{7.57}$$

The identity coupling $Eq_{(X,\sigma)}\colon (X,\sigma) \to (X,\sigma)$ is the distribution $\Delta_*(\sigma)$.

The essence of the following result is due to Clerc et al. (2017), but there it occurs in slightly different form, namely in a setting of continuous probability. Here it is transferred to the discrete situation.

Theorem 7.11.5 *Couplings as defined above indeed form a category* **Cpl**.

1 *This category carries a dagger functor* $(-)^\dagger\colon \mathbf{Cpl} \to \mathbf{Cpl}^{\mathrm{op}}$ *which is the identity on objects; on morphisms it is defined via swapping:*

$$\Big((X,\sigma) \xrightarrow{\varphi} (Y,\tau)\Big)^\dagger \coloneqq \Big((Y,\tau) \xrightarrow{\langle \pi_2,\pi_1\rangle_*(\varphi)} (X,\sigma)\Big).$$

More concretely, this dagger is defined by swapping arguments, as in: $\varphi^\dagger = \sum_{x,y} \varphi(x,y)|y,x\rangle$.

2 *There is an isomorphism of categories* **Krn** $\cong$ **Cpl**, *in one direction by taking the graph of a channel, and in the other direction by disintegration. This isomorphism commutes with the daggers on the two categories.*

Proof. We first need to prove that $\psi \bullet \varphi$ is a distribution:

$$\begin{aligned}
\sum_{x\in X,\, z\in Z} (\psi \bullet \varphi)(x,z) &= \sum_{x\in X,\, y\in Y,\, z\in Z} \frac{\varphi(x,y)\cdot\psi(y,z)}{\tau(y)} = \sum_{y\in Y,\, z\in Z} \frac{\big(\sum_{x\in X}\varphi(x,y)\big)\cdot\psi(y,z)}{\tau(y)} \\
&= \sum_{y\in Y,\, z\in Z} \frac{\varphi[0,1](y)\cdot\psi(y,z)}{\tau(y)} \\
&= \sum_{y\in Y,\, z\in Z} \frac{\tau(y)\cdot\psi(y,z)}{\tau(y)} = \sum_{y\in Y,\, z\in Z} \psi(y,z) = 1.
\end{aligned}$$

We leave it to the reader to check that $Eq_{(X,\sigma)} = \Delta_*(\sigma)$ is neutral element for $\bullet$. We do verify that $\bullet$ is associative — and thus that **Cpl** is indeed a category. Let $\varphi\colon (X,\sigma) \to (Y,\tau)$, $\psi\colon (Y,\tau) \to (Z,\rho)$, $\chi\colon (Z,\rho) \to (W,\kappa)$ be morphisms in **Cpl**. Then:

$$\begin{aligned}
\Big(\chi \bullet (\psi \bullet \varphi)\Big)(x,w) &= \sum_{z\in Z} \frac{(\psi \bullet \varphi)(x,z)\cdot\chi(z,w)}{\rho(z)} \\
&= \sum_{y\in Y,\, z\in Z} \frac{\varphi(x,y)\cdot\psi(y,z)\cdot\chi(z,w)}{\tau(y)\cdot\rho(z)} \\
&= \sum_{y\in Y} \frac{\varphi(x,y)\cdot(\chi \bullet \psi)(y,w)}{\tau(y)} = \Big((\chi \bullet \psi) \bullet \varphi\Big)(x,w).
\end{aligned}$$

We turn to the dagger. It is obvious that $(-)^{\dagger\dagger}$ is the identity functor, and also that $(-)^\dagger$ preserves identity maps. It also preserves composition in **Cpl** since:

$$(\psi \bullet \varphi)^\dagger(z,x) = (\psi \bullet \varphi)(x,z) = \sum_{y\in Y} \frac{\varphi(x,y)\cdot\psi(y,z)}{\tau(y)} = \sum_{y\in Y} \frac{\psi^\dagger(z,y)\cdot\varphi^\dagger(y,x)}{\tau(y)} = (\varphi^\dagger \bullet \psi^\dagger)(z,x).$$

The graph operation on channels gives rise to an identity-on-objects 'graph' functor $G\colon \mathbf{Krn} \to \mathbf{Cpl}$ via:

$$G\Big((X,\sigma) \xrightarrow{f} (Y,\tau)\Big) \coloneqq \Big((Y,\tau) \xrightarrow{\langle id, f\rangle_*(\sigma)} (X,\sigma)\Big).$$

This yields a functor since:

$$\begin{aligned}
G(id_{(X,\sigma)}) &= \langle id, id\rangle_*(\sigma) = Eq_{(X,\sigma)} \\
G(g \circ f)(x,z) &= \langle id, g \circ f\rangle_*(\sigma)(x,z) = \sigma(x) \cdot (g \circ f)(x)(z) \\
&= \sum_{y \in Y} \sigma(x) \cdot f(x)(y) \cdot g(y)(z) \\
&= \sum_{y \in Y} \frac{\sigma(x) \cdot f(x)(y) \cdot \tau(y) \cdot g(y)(z)}{\tau(y)} \\
&= \sum_{y \in Y} \frac{\langle id, f\rangle_*(\sigma)(x,y) \cdot \langle id, g\rangle_*(\tau)(y,z)}{\tau(y)} \\
&= (\langle id, g\rangle_*(\tau) \bullet \langle id, f\rangle_*(\sigma))(x,z) = (G(g) \bullet G(f))(x,z).
\end{aligned}$$

In the other direction we define a functor $F\colon \mathbf{Cpl} \to \mathbf{Krn}$ which is the identity on objects and uses disintegration on morphisms: for $\varphi\colon (X,\sigma) \to (Y,\tau)$ in $\mathbf{Cpl}$ we get a channel $F(\varphi) \coloneqq \varphi[0,1 \,|\, 1,0]\colon X \rightsquigarrow Y$ which satisfies, by construction (7.11):

$$\varphi = \langle id, \varphi[0,1 \,|\, 1,0]\rangle_*(\varphi[1,0]) = \langle id, F(\varphi)\rangle_*(\sigma) = GF(\varphi).$$

Moreover, $F(\varphi)$ is a morphism $(X,\sigma) \to (Y,\tau)$ in $\mathbf{Krn}$ since:

$$F(\varphi)_*(\sigma) = \varphi[0,1 \,|\, 1,0]_*(\varphi[1,0]) = \Big(\langle id, \varphi[0,1 \,|\, 1,0]\rangle_*(\varphi[1,0])\Big)[0,1] \overset{(7.11)}{=} \varphi[0,1] = \tau.$$

We still need to prove that F preserves identities and composition. This follows by uniqueness of disintegration. We get $F(Eq_{(X,\sigma)}) = id_{F(X,\sigma)}$ from:

$$\langle id, F(Eq_{(X,\sigma)})\rangle_*(\sigma) = Eq_{(X,\sigma)} = \Delta_*(\sigma) = \langle id, id\rangle_*(\sigma).$$

Next, the equation $F(\psi \bullet \varphi) = F(\psi) \circ F(\varphi)$ follows from:

$$\begin{aligned}
\langle id, F(\psi \bullet \varphi)\rangle_*(\sigma) &= \psi \bullet \varphi && \text{by definition} \\
&\overset{(7.57)}{=} \langle \varphi[1,0 \,|\, 0,1], \psi[0,1 \,|\, 1,0]\rangle_*(\tau) \\
&= (id \otimes \psi[0,1 \,|\, 1,0])_*\Big(\langle \varphi[1,0 \,|\, 0,1], id\rangle_*(\varphi[1,0 \,|\, 0,1]_*(\sigma))\Big) \\
&= (id \otimes F(\psi))_*\Big(\langle F(\varphi)^\dagger, id\rangle_*(F(\varphi)_*(\sigma))\Big) && \text{by Exercise 7.2.8} \\
&\overset{(7.11)}{=} (id \otimes F(\psi))_*(\langle id, F(\varphi)\rangle_*(\sigma)) \\
&= \langle id, F(\psi) \circ F(\varphi)\rangle_*(\sigma).
\end{aligned}$$

We have seen that, by construction, $G \circ F$ is the identity functor on the category $\mathbf{Cpl}$. In the other direction we also have $F \circ G = id\colon \mathbf{Krn} \to \mathbf{Krn}$. This follows directly from uniqueness of disintegration.

In the end we like to show that the functors G and F commute with the daggers, on kernels

and couplings. This works as follows. First, for $f\colon (X,\sigma) \to (Y,\tau)$ we have:

$$\begin{aligned} G(f^\dagger)(y,x) \;=\; \langle id, f^\dagger\rangle_*(\tau)(y,x) &= \tau(y)\cdot f^\dagger_\sigma(y)(x) \\ &\overset{(7.1)}{=} f_*(\sigma)(y)\cdot \frac{\sigma(x)\cdot f(x)(y)}{f_*(\sigma)(y)} \\ &= \sigma(x)\cdot f(x)(y) \\ &= \langle id, f\rangle_*(\sigma)(x,y) \;=\; G(f)(x,y) \;=\; G(f)^\dagger(y,x). \end{aligned}$$

Next, for $\varphi\colon (X,\sigma) \to (Y,\tau)$ in **Cpl**,

$$\begin{aligned} F(\varphi^\dagger)(y)(x) \;=\; \varphi^\dagger[0,1\,|\,1,0](y)(x) &\overset{(7.12)}{=} \frac{\varphi^\dagger(y,x)}{\varphi^\dagger[1,0](y)} \\ &= \frac{\varphi(x,y)}{\varphi[0,1](y)} \\ &= \varphi[1,0\,|\,0,1](y)(x) \\ &= (\varphi[0,1\,|\,1,0])^\dagger(y)(x) \qquad \text{by Exercise 7.2.8} \\ &= F(\varphi)^\dagger(y)(x). \end{aligned}$$

❑

We have done a lot of work in order to be able to say that **Krn** and **Cpl** are isomorphic dagger categories, or, more informally, that there is a one-one correspondence between probabilistic computations (channels) and probabilistic relations.

Exercises

7.11.1 Prove in the context of the powerset channels of Example 7.11.1 that:

1. $unit^\dagger = unit$ and $(g \circ f)^\dagger = f^\dagger \circ g^\dagger$.
2. $G(unit_X) = Eq_X$ and $G(g \circ f) = G(f) \bullet G(f)$.
3. $F(Eq_X) = unit_X$ and $F(S \bullet R) = F(S) \circ F(R)$.
4. $(S \bullet R)^\dagger = R^\dagger \bullet S^\dagger$.
5. $F(R^\dagger) = F(R)^\dagger$.

7.11.2 1. Define tensors $\otimes$ on the category of kernels, as a functor $\mathbf{Krn} \times \mathbf{Krn} \to \mathbf{Krn}$.

2. Prove that the dagger functor on **Krn** preseves these tensors: $(f \otimes g)^\dagger = f^\dagger \otimes g^\dagger$.

7.11.3 Prove the equation = in the definition (7.57) of composition $\bullet$ in the category **Cpl**. Give also a string diagrammatic description of $\bullet$.

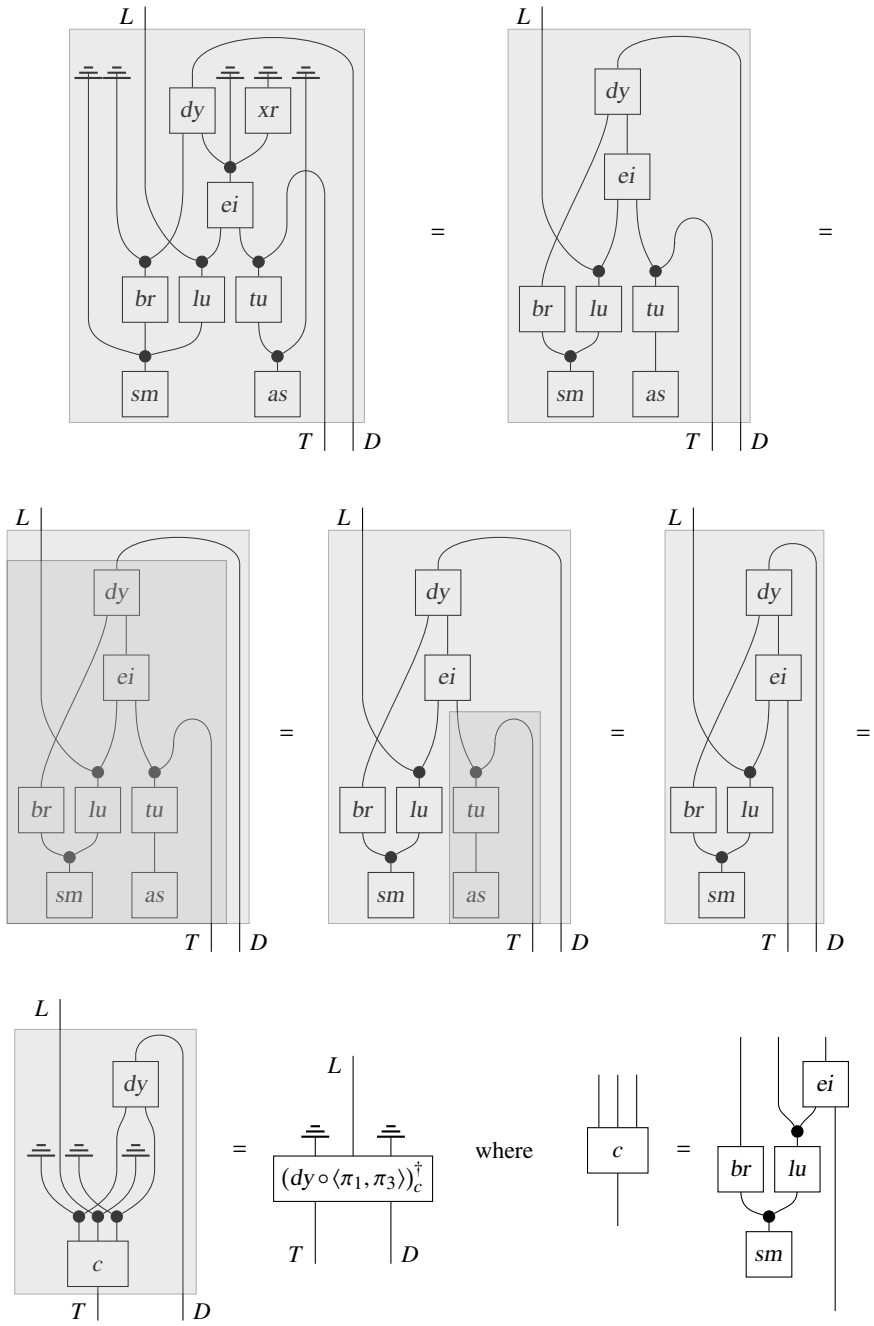


Figure 7.9 Disintegration of the Asia Bayesian network from Example 7.10.8, in order to get the lung cancer probability given dyspnoea and tuberculosis.

8

Learning of Distributions and Channels

In Chapters 6 and 7 much attention has been paid to updating of distributions. This chapter goes a step further and also considers updating of channels. Such channel updating is an essential ingredient of many basic algorithms in probabilistic learning, for instance in the class of algorithms called expectation maximisation (EM) and in learning of hidden Markov models (HMMs). These algorithms typically learn from multiple data items, in the form of a multiset (for EM) or a list (for HMMs). Therefore, the topic of learning from multiple items is also included in this chapter. It will lead to a more systematic approach to the update rules of Pearl and Jeffrey, that we saw earlier in Section 7.9.

The first section of this chapter considers a simple question: suppose we have multiple predicates, say p_1 and p_2, in presence of a single distribution ω. How should we evaluate the joint validity of these predicates?. There is a difference between the validity of the conjunction, below on the left, and the multiplication of separate validities, on the right.

$$\begin{aligned} \omega \models (p_1 \mathbin{\&} p_2) &= (\omega \models p_1) \cdot (\omega|_{p_1} \models p_2) \\ &= (\omega \models p_2) \cdot (\omega|_{p_2} \models p_1) \end{aligned} \qquad (\omega \models p_1) \cdot (\omega \models p_2) \tag{8.1}$$

The equations on the left are instances of Bayes' product rule, see Theorem 6.1.4 (1). The difference may be described in terms of dependent or independent evaluation (of the validities). In presence of channels the number of validity formulations increases rapidly, since predicate transformation and conjunction do not commute. These differences can be expressed in programming languages, like ProbLog and WebPPL, but in a rather indirect manner. This section illustrates how a proper logical language is helpful to be aware of the various differences and to make (the right) choices.

Next, Section 8.2 is also introductory in nature. It concentrates on a basic result for increasing sum expressions, of the form $\sum_y F(x, y)$, for a function F. This applies to basic validities, like $\omega \models p = \sum_x \omega(x) \cdot p(x)$, but also to more complicated sum expressions. The result gives a recipe for a higher outcome, which can often be determined via Lagrange's multiplier method. In this way we rediscover known validity increases, but we also find an essential new increase result, namely Proposition 8.2.5.

The subsequent two sections 8.3 and 8.4 introduce a new perspective on 'information' that we wish to learn from. It is called evidence and takes the form of a (natural) multiset $\sum_i n_i | p_i \rangle$ of factors p_i (or observables, or predicates). We have seen Bayesian updating, which involves conditioning a distribution with a single factor. At this stage we introduce two formulations, both for validity and for updating, with multiple factors. These two approaches are associated with Pearl and with Jeffrey, leading two four forms in total of Pearl / Jeffrey validity / updating, starting from evidence. It is shown that Pearl updating increases Pearl validity, and similarly that

Jeffrey updating increases Jeffrey validity. Such a valdity increase need not happen when the 'wrong' form of updating is used. The earlier validities $\omega \models (p_1 \mathbin{\&} p_2)$ and $(\omega \models p_1) \cdot (\omega \models p_2)$ in (8.1) are instance of Pearl validity and of Jeffrey validity, of the two factors p_1 and p_2, now considered as evidence multiset $1|p_1\rangle + 1|p_2\rangle$.

The next three sections 8.5 – 8.7 concentrate on expectation maximisation (EM). This is a powerful (unsupervised) learning technique that can find a convex mixture of (a fixed number of) distributions in data. The algorithm aims to find a mixture of distributions with a minimal divergence from the data distribution (obtained via frequentist learning). The data forms a multiset of points, also called point evidence. The mixture weights / probabilities form a distribution on a 'latent' set of features, whereas the distributions in the mixture form a channel, defined on the set of features. Expectation maximisation thus involves learning of both a mixture distribution and a channel. This learning happens iteratively, in multiple steps. A single update of the mixture distribution is called an E-step and happens via Jeffrey's rule. The M-step of the algorithm involves the update of the channel. As we shall see, this happens via a double-dagger construction, with a 'twist' that depends on the situation at hand. Section 8.5 introduces the basic idea of expectation maximisation in greater detail and explains how the double dagger forms an 'ideal' solution. The subsequent two sections elaborate examples, first without parameterisation (in Section 8.6) and then with parameters (in Section 8.7). The latter form is common in the literature.

The final two section 8.8 and 8.9 of this chapter are about learning of Markov chains and of hidden Markov models. The data now consists of a list of factors, that capture what is observed at successive steps. The algorithm for doing this is called Baum-Welch (after Baum et al. (1970)) and involves a non-trivial application of the sum-increase method from Section 8.2, not only for learning a new distribution, but also for learning new (transition and emmission) channels. Several examples are included and the connection to a standard (Python) implementation of Baum-Welch is made.

8.1 When to use conjunctions

Early on we have seen that predicate transformation does not preserve conjunction & of predicates (and thus observables), see Exercise 4.5.1. This means that if we have a channel c and multiple predicates q_i on the channel's codomain, there are choices to make about how to combine them, for instance do conjunction first as in $c^*(q_1 \mathbin{\&} q_2)$, or do predicate transformation first, as in $c^*(q_1) \mathbin{\&} c^*(q_2)$.

For completeness, this non-preservation holds for proper probabilistic channels, and not for deterministic channels. These deterministic channels do preserve conjunctions, see Proposition 4.4.11 and also Exercise 4.5.5.

This section elaborates two examples to set the scene for the rest of this chapter. The first illustration is a slightly simplified version of the example that is used in Jacobs (2024) — without the multiple occurrences of the same predicate. The second example below adds a channel.

Example 8.1.1 Let $\omega = \frac{1}{20}|d\rangle + \frac{19}{20}|d^\perp\rangle$ be a distribution about the occurrence of a certain disease d in a population. It gives what medical people call the prevalence, *i.e.* the prior disease distribution. There is a test available for disease d. It is not perfect: a positive test outcome does not mean that the person being tested actually has the disease. The situation is captured by the predicate $pt = \frac{9}{10} \cdot \mathbf{1}_d + \frac{2}{5} \cdot \mathbf{1}_{d^\perp}$. It gives 90% disease certainty and 40% non-disease certainty.

Recall that in a predicate these probabilities need not add up to one. The corresponding negative test predicate is $nt = pt^{\perp} = \frac{1}{10} \cdot \mathbf{1}_d + \frac{3}{5} \cdot \mathbf{1}_{d^{\perp}}$. Now we ask the question: suppose we have both a positive and a negative test outcome. How likely is this, that is, what is the probability of two such tests?

(a) The most obvious approach is to compute the validity of the conjunction:

$$\omega \models pt \mathbin{\&} nt = \tfrac{1}{20} \cdot \tfrac{9}{10} \cdot \tfrac{1}{10} + \tfrac{19}{20} \cdot \tfrac{2}{5} \cdot \tfrac{3}{5} = \tfrac{93}{400} = 0.2325.$$

(b) However, one could also take the product of the validities:

$$(\omega \models pt) \cdot (\omega \models nt) = (\tfrac{1}{20} \cdot \tfrac{9}{10} + \tfrac{19}{20} \cdot \tfrac{1}{10}) \cdot (\tfrac{1}{20} \cdot \tfrac{2}{5} + \tfrac{19}{20} \cdot \tfrac{3}{5}) = \tfrac{391}{1600} = 0.244375.$$

The difference in outcomes is not dramatic, but the important thing is that there is a difference — and thus a choice and a question: when to apply which computation?

In this simple situation the matter is not difficult. We should distinguish whether the tests are applied to the same individual, or to two random persons from the population. In the first case (a) the test outcomes are dependent, whereas in the second cases they are independent. This dependence in the first case may also be expressed via Bayes' product rule from Theorem 6.1.4 (1). One can look at the validity of the first, say positive test and evaluate the second (negative) test in the updated distribution (for the same individual). Doing this in a different order gives the same outcome, namely the conjunction of the two test predicates:

$$(\omega \models pt) \cdot (\omega|_{pt} \models nt) = \omega \models pt \mathbin{\&} nt = (\omega \models nt) \cdot (\omega|_{nt} \models pt).$$

In the second case (b) the validity calculations are done with two separate copies of the distribution ω, corresponding to independent evaluation, for two arbitrary individuals from the population.

In the beginning of this section we mentioned the non-preservation of conjunction & by predicate transformation. The difference between the outcomes in the above two points (a), (b) is in fact an instance of this phenomenon. Recall from Exercise 4.5.8 that when we regard the distribution $\omega \in \mathcal{D}(\{d, d^{\perp}\})$ as a channel $\omega \colon \mathbf{1} \rightsquigarrow \{d, d^{\perp}\}$, then predicate transformation $\omega^{*} \colon Pred(\{d, d^{\perp}\}) \to Pred(\mathbf{1}) \cong [0, 1]$ is validity: $\omega^{*}(p) = \omega \models p$. Thus we can describe the computation in (a) as a conjunction followed by predicate transformation:

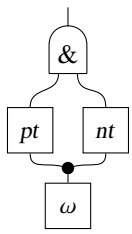
$$\omega \models pt \mathbin{\&} nt = \omega^{*}(pt \mathbin{\&} nt).$$

The outcome in (b) is predicate transformation followed by conjunction, where we use that conjunction for predicates in $Pred(\mathbf{1}) \cong [0, 1]$ is simply multiplication:

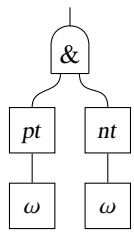
$$(\omega \models pt) \cdot (\omega \models nt) = \omega^{*}(pt) \mathbin{\&} \omega^{*}(nt).$$

There is yet another way to describe the difference, via string diagrams

(a) =



(b) =



(8.2)

```
1/20 :: disease(X).

9/10 :: positive_test(X) :- disease(X).
2/5  :: positive_test(X) :- not disease(X).

1/10 :: negative_test(X) :- disease(X).
3/5  :: negative_test(X) :- not disease(X).

conjunction_a :- positive_test(1), negative_test(1).
conjunction_b :- positive_test(2), negative_test(3).

query(conjunction_a).
query(conjunction_b).
```

Figure 8.1 ProbLog representation of the two combinations of medical tests in Example 8.1.1.

Here we use the test predicates as channels $pt, nt\colon \{d, d^{\perp}\} \twoheadrightarrow \mathbf{2}$, and combine the results at the top via conjunction, using the and-channel $ac\colon \mathbf{2} \times \mathbf{2} \twoheadrightarrow \mathbf{2}$ from Exercise 2.5.9.

We can also formulate the two options (a), (b) in the probabilistic logic programming language ProbLog, using parametrisation of the distribution and predicates to keep them apart, see the code fragment in Figure 8.1. Notice that for option (a) we use the same instantiation 1 of the distribution, whereas for option (b) we use different instantions 2, 3. The queries at the end of the program in Figure 8.1 prove the same probabilities that we computed above, of the form:

```
conjunction_a: 0.2325          conjunction_b: 0.244375
```

In our next example we add a channel. It leads to many more options.

Example 8.1.2 We now start with a distribution $\omega = \frac{9}{20}|h\rangle + \frac{11}{20}|\ell\rangle$ describing the probabilities of high (h) and low (ℓ) inflation. There is an associated channel $c\colon \{h, \ell\} \twoheadrightarrow \{m, i, s\}$ about the sectors in which to invest: m for minerals, i for industry, s for services. Let's say that the channel has the following form:

$$c(h) \;=\; \tfrac{2}{5}|m\rangle + \tfrac{1}{4}|i\rangle + \tfrac{7}{20}|s\rangle \qquad\qquad c(\ell) \;=\; \tfrac{9}{20}|m\rangle + \tfrac{3}{20}|i\rangle + \tfrac{2}{5}|s\rangle.$$

We are given three pieces of advice, in the form of predicates $q_1, q_2, q_3 \in \mathit{Pred}(\{m, i, s\})$, namely:

$$\begin{aligned} q_1 &= \tfrac{3}{20} \cdot \mathbf{1}_m + \tfrac{19}{20} \cdot \mathbf{1}_i + \tfrac{9}{10} \cdot \mathbf{1}_s \\ q_2 &= \tfrac{7}{20} \cdot \mathbf{1}_m + \tfrac{4}{5} \cdot \mathbf{1}_i + \tfrac{3}{4} \cdot \mathbf{1}_s \\ q_3 &= \tfrac{3}{4} \cdot \mathbf{1}_m + \tfrac{1}{20} \cdot \mathbf{1}_i + \tfrac{1}{4} \cdot \mathbf{1}_s. \end{aligned}$$

(The numbers are arbitrary and are not meant to make economic sense.)

We now ask ourselves: how to evaluate this combination of advice? Figure 8.2 describes six options as string diagrams. The conjunction channels on top are omitted, for simplicity. The options (b), (d) and (e) allows variations via permutations of the predicates q_i, but these variations are non-structural and thus not listed.

The validities of these six options are calculated as follows.

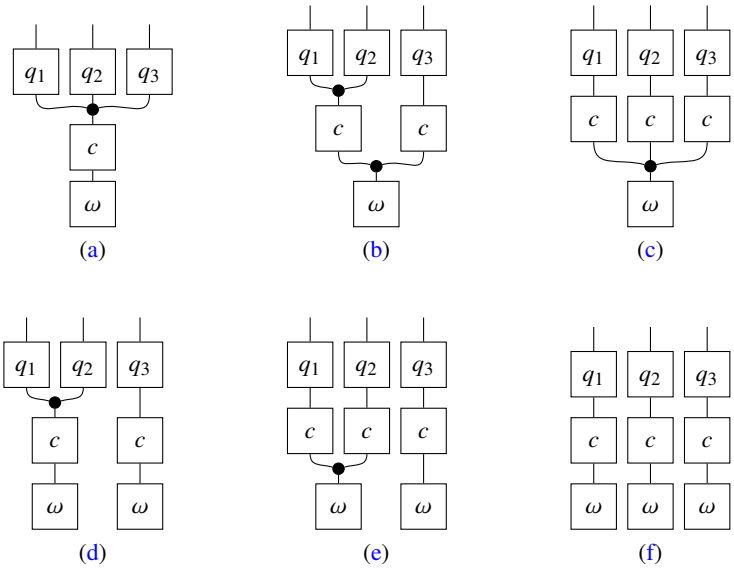


Figure 8.2 The six possible ways of combining three advice predicates q_i in presence of a channel c that are considered in Example 8.1.2, described as string diagrams. The three wires at the top should all go into a ternary conjunction channel that is omitted for simplicity.

(a) This first case involves an 'early' conjunction of all predicates:

$$\omega \models c^*(q_1 \mathbin{\&} q_2 \mathbin{\&} q_3) \;=\; 0.08795.$$

(b) Next the early conjunction is applied to the first two predicates only. Its validity is twice as high as the previous one:

$$\omega \models c^*(q_1 \mathbin{\&} q_2) \mathbin{\&} c^*(q_3) \;=\; 0.18027.$$

(c) Next we apply the conjunction to all transformed predicates:

$$\omega \models c^*(q_1) \mathbin{\&} c^*(q_2) \mathbin{\&} c^*(q_3) \;=\; 0.14699.$$

(d) We now start using two copies of the distribution ω, with an early evaluation of the first two predicates:

$$\big(\omega \models c^*(q_1 \mathbin{\&} q_2)\big) \cdot \big(\omega \models c^*(q_3)\big) \;=\; 0.18071.$$

(e) Next we do the conjunction later:

$$\big(\omega \models c^*(q_1) \mathbin{\&} c^*(q_2)\big) \cdot \big(\omega \models c^*(q_3)\big) \;=\; 0.14716.$$

(f) Finally, we evaluate in three separate copies of the distribution:

$$\big(\omega \models c^*(q_1)\big) \cdot \big(\omega \models c^*(q_2)\big) \cdot \big(\omega \models c^*(q_3)\big) \;=\; 0.14732.$$

We see that some outcomes are close, but also that there is considerable variation. One can imagine that the values vary more if there are more than three predicates involved.

In Example 8.1.1 we have explained the different formulations in terms of dependence and independence. Is there a a similar explanation here? We try between (b) and (e). We can imagine that the advice comes from investment advisors working for a particular agency. In case of (e) one can think of advice q_1 and q_2 coming from two advisors working for the same agency and thus involving a dependence, whereas advice q_3 comes from an advisor working for a different, independent agency. For the formulation (e) we can think of ω describing the inflation in a particular country. Thus, when there are two copies of ω around, the advice may come from different countries, where in the first one country it comes from advisors working for different agencies. In this way different interpretations may be given for all six formulations (a) – (f).

Finally we also look at how to formulate these six options in ProbLog. This is possible, in an indirect manner, via ProbLog's parametrisation mechanism, see Figure 8.3. It is not the case that the language offers predicate transformation as a separate operator that can be applied in a specific manner, in relation to conjunction. Below we list the definitions of `advice_x` for the six different options and the resulting probability when the ProbLog program from Figure 8.3 is run with an additional statement `query(advice_x).`

- `advice_a :- advice_1(1,1), advice_2(1,1), advice_3(1,1)`
  `advice_a : 0.087945937`
- `advice_b :- advice_1(2,1), advice_2(2,1), advice_3(2,2)`
  `advice_b : 0.18027122`
- `advice_c :- advice_1(3,1), advice_2(3,2), advice_3(3,3)`
  `advice_c : 0.14699121`
- `advice_d :- advice_1(4,1), advice_2(4,1), advice_3(5,1)`
  `advice_d : 0.18071254`
- `advice_e :- advice_1(6,1), advice_2(7,1), advice_3(7,2).`
  `advice_e : 0.14715937`
- `advice_f :- advice_1(8,1), advice_2(9,1), advice_3(10,1)`
  `advice_f : 0.14732341`

We see that the listed probabilities match with what we have computed in the above points (a) – (f). The numbers `advice_1(i,j)` in the above definitions of `advice_x` refer to the i-th instance of the inflation distribution and the j-th instance of the investment channel. If the i's are the same in advice `advice_x`, then the same distribution is used, like in points (a) – (c). Different j's refer to different predicate transformations, like in all points except (a). Thus, the numbers i, j in the above six bullet points match the descriptions in Figure 8.2.

The conclusion that we draw from the illustrations in this section is that combining multiple predicates is a subtle matter, certainly in presence of predicate transformation. We have used conjunction in the example combinations, but similar differences in validity show up with disjunction — via the or-channel *oc* from Exercise 2.5.9. In order to distinguish the different combination options it is essential that one has an expressive language that supports the different formulations in a transparent manner. We claim that string diagrams, see Figure 8.3, form such a language with visually clear distinctions. The assoicated textual formulations using validity $\models$, predicates transformation c^* and conjunction & in points (a) – (f) mimicks these differences. The various options can be encoded in ProbLog, but only indirectly, via a tricky use of its parametrisation mechanism. The differences may also be expressed in the probabilistic programming language

```
9/20 :: inflation(high,X); 11/20 :: inflation(low,X)
   :- inflation(X).

inflation(1). inflation(2). inflation(3). inflation(4). inflation(5).
inflation(6). inflation(7). inflation(8). inflation(9). inflation(10).

2/5   :: invest(minerals,X,Y);
1/4   :: invest(industry,X,Y);
7/20  :: invest(service,X,Y) :- inflation(high,X).
9/20  :: invest(minerals,X,Y);
3/20  :: invest(industry,X,Y);
2/5   :: invest(service,X,Y) :- inflation(low,X).

3/20  :: advice1(X,Y) :- invest(minerals,X,Y).
19/20 :: advice1(X,Y) :- invest(industry,X,Y).
9/10  :: advice1(X,Y) :- invest(service,X,Y).

7/20  :: advice2(X,Y) :- invest(minerals,X,Y).
4/5   :: advice2(X,Y) :- invest(industry,X,Y).
3/4   :: advice2(X,Y) :- invest(service,X,Y).

3/4   :: advice3(X,Y) :- invest(minerals,X,Y).
1/20  :: advice3(X,Y) :- invest(industry,X,Y).
1/4   :: advice3(X,Y) :- invest(service,X,Y).
```

Figure 8.3 Definitions in ProbLog, in parametrised form, of the inflation distribution, the investment options channel, and the three advice predicates from Example 8.1.2.

WebPPL via different sampling variables, see Exercise 8.1.2 below. In both cases the different combinations are not captured by different logical formulas. Similarly, in traditional notation in probability there is no obvious (logical) way to express the different combinations. This may be the reason why the differences have not been clearly identified earlier. This is a serious issue, because important decisions affecting people may depend on how multiple data about them and others is evaluated in a particular situation.

In the course of this chapter we will see that there are different forms of updating (learning) associated with the different options for validity computations that we saw above. Thus, in learning from multiple data, it is important how to organise this data, since different combinations will produce different outcomes.

Exercises

8.1.1 Consider the disease prevalence distribution $\omega = \frac{1}{20}|d\rangle + \frac{19}{20}|d^{\perp}\rangle$ from Example 8.1.1 with the associated positive and negative test predicates pt and nt. Assume that a doctor sees three patients in one afternoon, with the following consecutive test results.

patient 1	**patient 2**	**patient 3**
pt	*nt*	*nt*
nt	*nt*	*nt*
nt		*pt*

For some reason, *e.g.* lack of time, only two tests were done for the second patient.

1 Write a string diagram for the validity of the combined tests, like in Figure 8.2.

2 You may wish to actually compute this validity, as: $\frac{41291937}{6400000000} \approx 0.00645$.

8.1.2 1 Implement the distribution ω, the channel c and the three predicates q_1, q_2, q_3 from Example 8.1.2 in WebPPL.

2 The validities in Example 8.1.2 (a) and (b) can now be obtained as:

```
var val_a = Math.exp(Infer(function() {
        var x = sample(w);
        var y = sample(c(x));
        return q1(y) & q2(y) & q3(y) }).score(1))
var val_b = Math.exp(Infer(function() {
        var x = sample(w);
        var y1 = sample(c(x));
        var y2 = sample(c(x));
        return q1(y1) & q2(y1) & q3(y2) }).score(1))
```

Write your own implementations of the other validities from Example 8.1.2 (c) – (f) yourself.

8.2 A general sum-increase method

We recall the 'learning makes us wiser' inequality $\omega|_p \models p \geq \omega \models p$ from Theorem 6.1.6. This result, and similar inequalities between validities in Proposition 6.3.6, was derived from the inequality $(\omega \models p^2) \geq (\omega \models p)^2$ in Corollary 5.1.4 (1). This section introduces a more general approach to derive such inequalities, stemming from (Baum et al., 1970, Thm. 2.1). It forms the basis for many learning methods using validity increases. This approach involves a generic sum-increase result, see Lemma 8.2.1 below. It applies in many situations, with the help of some basic real analysis. This will be illustrated, among others, by re-deriving, in a completely different manner, some of the earlier inequalities — including the most basic one $\omega|_p \models p \geq \omega \models p$ from Theorem 6.1.6.

Lemma 8.2.1 *Let X, Y be finite sets, and let $F \colon X \times Y \to \mathbb{R}_{\geq 0}$ be a given function. For each $x \in X$, write $F_1(x) := \sum_{y \in Y} F(x, y)$ for the 'marginal' sum that we wish to increase. Now fix $x \in X$ and assume that there is an $x' \in X$ with:*

$$x' \in \operatorname*{argmax}_{z \in X} G(x, z) \qquad \textit{where} \qquad G(x, z) := \sum\nolimits_{y \in Y} F(x, y) \cdot \ln\big(F(z, y)\big).$$

Then $F_1(x') \geq F_1(x)$.

We often say informally that the element x' is (or works) better than x.

Proof. We fix $x \in X$ and let $x' \in X$ be an element where $G(x,-)\colon X \to \mathbb{R}_{\geq 0}$ takes its maximum. We need to prove that x' satisfies $F_1(x') \geq F_1(x)$. This obviously holds when $F_1(x) = 0$. When $F_1(x) \neq 0$ we Jensen's inequality:

$$\begin{aligned}
\ln\left(\frac{F_1(x')}{F_1(x)}\right) = \ln\left(\sum_{y\in Y} \frac{F(x',y)}{F_1(x)}\right) &= \ln\left(\sum_{y\in Y} \frac{F(x,y)}{F_1(x)} \cdot \frac{F(x',y)}{F(x,y)}\right) \\
&\geq \sum_{y\in Y} \frac{F(x,y)}{F_1(x)} \cdot \ln\left(\frac{F(x',y)}{F(x,y)}\right) \qquad \text{by Lemma 2.10.3 (3)} \\
&= \frac{1}{F_1(x)} \cdot \sum_{y\in Y} F(x,y) \cdot \Big(\ln\big(F(x',y)\big) - \ln\big(F(x,y)\big)\Big) \\
&= \frac{1}{F_1(x)} \cdot \Big(G(x,x') - G(x,x)\Big) \geq 0.
\end{aligned}$$

❑

In general, this inequality $\geq$ in Lemma 8.2.1 gives a strict increase, like in Jensen's inequality in Lemma 2.10.3. It forms the basis for many forms of learning. An actual maximum x' can in many situation be determined analytically — typically via the Lagrange multiplier method — but it need not be unique and it can be a local maximum. This method will be illustrated in the remainder of this section. We will see that several previous validity increase results can be obtained via the above lemma.

In our first application of this lemma we re-derive the 'getting wiser' inequality $\omega|_p \models p \geq \omega \models p$ from Theorem 6.1.6.

Example 8.2.2 Suppose we have a validity $\omega \models p$, where $\omega \in \mathcal{D}(Y)$ and $p \in Fact(Y)$, which we would like to increase, by changing the distribution ω, while keeping the factor p fixed. We assume that the set Y is finite. This may always be achieved by restricting Y to be the support of ω.

We apply Lemma 8.2.1 with function $F\colon \mathcal{D}(Y) \times Y \to \mathbb{R}_{\geq 0}$ given by $F(\omega, y) = \omega(y) \cdot p(y)$. Then $F_1(\omega) = \sum_y F(\omega, y) = \sum_y \omega(y) \cdot p(y) = \omega \models p$.

We have $G(\omega, \omega') = \sum_y \omega(y) \cdot p(y) \cdot \ln\big(\omega'(y) \cdot p(y)\big)$, as defined in Lemma 8.2.1. We wish to find a maximum of the function $G(\omega, -)\colon \mathcal{D}(Y) \to \mathbb{R}_{\geq 0}$. Let $Y = \{y_1, \ldots, y_n\}$; we will use variables v_i for the values $\omega'(y_i)$ that we wish to find, by seeing $G(\omega, -)$ as a function $\mathbb{R}^n \to \mathbb{R}_{\geq 0}$ with constraints.

These constraints are handled via the Lagrange multiplier method for finding the maximum (see *e.g.* (Bishop, 2006, §2.2)). We keep the distribution ω fixed and consider a new function H, also known as the Lagrangian, with an additional parameter κ.

$$\begin{aligned}
H(\vec{v}, \kappa) &:= G(\omega, \vec{v}) + \kappa \cdot \left(1 - \sum_{1\leq i\leq n} v_i\right) \\
&= \sum_{1\leq i\leq n} \omega(y_i) \cdot p(y_i) \cdot \ln\Big(v_i \cdot p(y_i)\Big) + \kappa \cdot \left(1 - \sum_{1\leq i\leq n} v_i\right).
\end{aligned}$$

The partial derivatives of H are:

$$\begin{aligned}
\frac{\partial H}{\partial v_i}(\vec{v}, \kappa) &= \frac{\omega(y_i) \cdot p(y_i)}{v_i \cdot p(y_i)} \cdot p(v_i) - \kappa = \frac{\omega(y_i) \cdot p(y_i)}{v_i} - \kappa \\
\frac{\partial H}{\partial \kappa}(\vec{v}, \kappa) &= 1 - \sum_{1\leq i\leq n} v_i
\end{aligned}$$

Setting all these derivatives to zero yields:

$$1 \;=\; \sum_{1 \le i \le n} v_i \;=\; \sum_{1 \le i \le n} \frac{\omega(y_i) \cdot p(y_i)}{\kappa} \;=\; \frac{\omega \models p}{\kappa}.$$

Hence $\kappa \;=\; \omega \models p$ and thus:

$$v_i \;=\; \frac{\omega(y_i) \cdot p(y_i)}{\kappa} \;=\; \frac{\omega(y_i) \cdot p(y_i)}{\omega \models p} \;=\; \omega|_p(y_i).$$

Hence, Lemma 8.2.1 gives as 'better' distribution the updated version $\omega|_p$ of ω with p. In Theorem 6.1.6 we already saw that $\omega|_p$ is a 'better' distribution, via the inequality $\omega|_p \models p \;\ge\; \omega \models p$. This is now re-comfirmed via a different method.

Next we show how to get a better channel.

Example 8.2.3 Let a (fixed) distribution $\omega \in \mathcal{D}(X)$ and factor $q \in \mathit{Fact}(Y)$ be given, on finite sets X, Y. We will now try to find a 'better' version of a channel $c\colon X \twoheadrightarrow Y$ in a validity situation $c_*(\omega) \models q \;=\; \omega \models c^*(q)$. We define the function $F\colon \mathcal{D}(Y)^X \times X \times Y \to \mathbb{R}_{\ge 0}$ as:

$$F(c, x, y) \;:=\; \omega(x) \cdot c(x)(y) \cdot q(y)$$

Then $F_1(c) = \sum_{x,y} F(c, x, y) = \sum_y c_*(\omega)(y) \cdot q(y) = c_*(\omega) \models q$.

Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. We shall use variables z_{ij} for the values $c'(x_i)(y_i)$ of the better channel c' that we seek. We can thus describe the function $G(c, -)\colon \mathbb{R}^{n \times m} \to \mathbb{R}_{\ge 0}$ like in Lemma 8.2.1 as:

$$\begin{aligned} G(c, \vec{z}) \;&=\; \sum_{i,j} F(c, x_i, y_j) \cdot \ln\Big(F(\vec{z}, x_i, y_j)\Big) \\ &=\; \sum_{i,j} \omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j) \cdot \ln\Big(\omega(x_i) \cdot z_{ij} \cdot q(y_j)\Big). \end{aligned}$$

In this situation we use the Lagrange multiplier method with n additional variables $\vec{\kappa}$ in a new function:

$$H(\vec{z}, \vec{\kappa}) \;:=\; \sum_{i,j} \omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j) \cdot \ln\Big(\omega(x_i) \cdot z_{ij} \cdot q(y_j)\Big) \;+\; \sum_i \kappa_i \cdot \Big(1 - \sum_j z_{ij}\Big).$$

The partial derivatives of this function H are:

$$\begin{aligned} \frac{\partial H}{\partial z_{ij}}(\vec{z}, \vec{\kappa}) \;&=\; \frac{\omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j)}{\omega(x_i) \cdot z_{ij} \cdot q(y_j)} \cdot \omega(x_i) \cdot q(y_j) \;-\; \kappa_i \\ &=\; \frac{\omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j)}{z_{ij}} \;-\; \kappa_i \\ \frac{\partial H}{\partial \kappa_i}(\vec{z}, \vec{\kappa}) \;&=\; 1 - \sum_j z_{ij}. \end{aligned}$$

Setting all these derivatives to zero yields for each i,

$$1 \;=\; \sum_j z_{ij} \;=\; \sum_j \frac{\omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j)}{\kappa_i} \;=\; \frac{\omega(x_i) \cdot (c(x_i) \models q)}{\kappa_i}.$$

Hence $\kappa_i = \omega(x_i) \cdot (c(x_i) \models q)$ so that:

$$z_{ij} \;=\; \frac{\omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j)}{\kappa_i} \;=\; \frac{\omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j)}{\omega(x_i) \cdot (c(x_i) \models q)}$$
$$= \frac{c(x_i)(y_j) \cdot q(y_j)}{c(x_i) \models q} \;=\; c(x_i)|_q(y_j) \;=\; c|_q(x_i)(y_j).$$

Thus, the 'better' channel than c in the situation $c_*(\omega) \models q$ is the channel $c|_q$ defined via pointwise updating $(c|_q)(x) = c(x)|_q$, see Definition 6.1.1 (3). This matches Proposition 6.3.6 (1) where we already saw the inequality $(c|_q)_*(\omega) \models q \;\geq\; c_*(\omega) \models q$.

In the description of Lemma 8.2.1 the set X may be a product of other sets, so that the sum-increase approach also works for multiple channels and/or distributions — giving a better version for each of them. This will be illustrated next.

Example 8.2.4 We again consider a situation $c_*(\omega) \models q$ as in Example 8.2.2, but now we wish to get better versions of both the channel $c\colon X \rightsquigarrow Y$ and the distribution $\omega \in \mathcal{D}(X)$. In this situation the factor q is fixed and the sets X, Y are finite. The approach is much as before, except that we now use an additional argument in the function $F\colon \mathcal{D}(Y)^X \times \mathcal{D}(X) \times X \times Y \to \mathbb{R}_{\geq 0}$ as:

$$F(c, \omega, x, y) \;\coloneqq\; \omega(x) \cdot c(x)(y) \cdot q(y)$$

Then $F_1(c, \omega) = \sum_{x,y} F(c, \omega, x, y) = \sum_y c_*(\omega)(y) \cdot q(y) = c_*(\omega) \models q$.

Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. We shall use variables z_{ij} for $c'(x_i)(y_i)$ and w_i for the $\omega'(x_i)$ of the better channel c' and distribution ω' that we seek. We can thus describe the function $G(c, \omega, -, -)\colon \mathbb{R}^{n \times m} \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ as:

$$G(c, \omega, \vec{z}, \vec{w}) \;=\; \sum_{i,j} F(c, \omega, x_i, y_j) \cdot \ln\Big(F(\vec{z}, \vec{w}, x_i, y_j)\Big)$$
$$= \sum_{i,j} \omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j) \cdot \ln\Big(w_i \cdot z_{ij} \cdot q(y_j)\Big).$$

We again use the Lagrange multiplier method, now with variables $\vec{\kappa}, \lambda$ in the function:

$$H(\vec{z}, \vec{w}, \vec{\kappa}, \lambda) \;\coloneqq\; \sum_{i,j} \omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j) \cdot \ln\Big(w_i \cdot z_{ij} \cdot q(y_j)\Big)$$
$$+ \sum_i \kappa_i \cdot \Big(1 - \sum_j z_{ij}\Big) \;+\; \lambda \cdot \Big(1 - \sum_i w_i\Big).$$

The partial derivatives of H are:

$$\frac{\partial H}{\partial z_{ij}}(\vec{z}, \vec{w}, \vec{\kappa}, \lambda) \;=\; \frac{\omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j)}{w_i \cdot z_{ij} \cdot q(y_j)} \cdot w_i \cdot q(y_j) \;-\; \kappa_i \;=\; \frac{\omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j)}{z_{ij}} \;-\; \kappa_i$$
$$\frac{\partial H}{\partial w_i}(\vec{z}, \vec{w}, \vec{\kappa}, \lambda) \;=\; \sum\nolimits_j \frac{\omega(x_i) \cdot c(x_i)(y_j) \cdot q(y_j)}{w_i \cdot z_{ij} \cdot q(y_j)} \cdot z_{ij} \cdot q(y_j) \;-\; \lambda \;=\; \frac{\omega(x_i) \cdot c^*(q)(x_i)}{w_i} \;-\; \lambda$$
$$\frac{\partial H}{\partial \kappa_i}(\vec{z}, \vec{w}, \vec{\kappa}, \lambda) \;=\; 1 - \sum_j z_{ij}$$
$$\frac{\partial H}{\partial \lambda}(\vec{z}, \vec{w}, \vec{\kappa}, \lambda) \;=\; 1 - \sum_i w_i.$$

By taking all these derivatives to be zero we can derive $z_{ij} = c|_q(x_i)(y_j)$, as in Example 8.2.3. This

is left as an exercise to the reader. We prove that we can also derive $w_i = \omega|_{c^*(q)}(x_i)$, by setting the ∂w derivatives to zero and solving:

$$1 = \sum_i w_i = \sum_i \frac{\omega(x_i) \cdot c^*(q)(x_i)}{\lambda} = \frac{\omega \models c^*(q)}{\lambda}.$$

Thus $\lambda = \omega \models c^*(q)$ and so:

$$w_i = \frac{\omega(x_i) \cdot c^*(q)(x_i)}{\lambda} = \frac{\omega(x_i) \cdot c^*(q)(x_i)}{\omega \models c^*(q)} = \omega|_{c^*(q)}(x_i).$$

Thus, we rediscover the inequality $(c|_q)_*(\omega|_{c^*(q)}) \models q \geq c_*(\omega) \models q$ from Proposition 6.3.6 (1), in which better versions of both the channel c and the state ω are used.

The previous illustrations involved inequalities that we already knew. We continue with a new result that turns out to be quite useful.

Proposition 8.2.5 *Let $\omega \in \mathcal{D}(X)$ be distribution on a finite set X, with $N > 0$ many factors $p_1, \ldots, p_N$, also on X, all with non-zero validity $\omega \models p_i$.*

1 The convex sum of updated distributions $\omega' := \sum_i \frac{1}{N} \cdot \omega|_{p_i}$ satisfies:

$$\prod_i (\omega' \models p_i) \geq \prod_i (\omega \models p_i).$$

2 For $k_1, \ldots, k_N \in \mathbb{N}$ write $k = \sum_i k_i$ and $\omega' = \sum_i \frac{k_i}{k} \cdot \omega|_{p_i}$, where we assume $k > 0$. Then:

$$\prod_i (\omega' \models p_i)^{k_i} \geq \prod_i (\omega \models p_i)^{k_i}.$$

Proof. 1 We define a function $F(\omega, -)\colon X^N \to \mathbb{R}_{\geq 0}$, on $\vec{x} \in X^N$ as:

$$F(\omega, \vec{x}) := \prod_{1 \leq i \leq N} \omega(x_i) \cdot p_i(x_i).$$

Then by distributivity of multiplication over addition:

$$\sum_{\vec{x} \in X^N} F(\omega, \vec{x}) = \prod_{1 \leq i \leq N} \sum_{x \in X} \omega(x) \cdot p_i(x) = \prod_{1 \leq i \leq N} (\omega \models p_i).$$

Let $X = \{x_1, \ldots, x_n\}$ and let w_i be a variable for the new distributions ω' at x_i. We use the function H be given by:

$$H(\vec{w}, \lambda) := \sum_{\vec{x} \in X^N} F(\omega, \vec{x}) \cdot \ln\left(\prod_{1 \leq i \leq n} w_i \cdot p_i(x_i)\right) + \lambda \cdot \left(1 - \sum_{1 \leq i \leq n} w_i\right).$$

Its partial derivatives are:

$$\begin{aligned}
\frac{\partial H}{\partial w_k}(\vec{w}, \lambda) &= \sum_{\vec{x} \in X^{N-1}} \frac{\sum_{1 \leq i \leq N-1} F(\omega, x_1, \ldots, x_{i-1}, x_k, x_i, \ldots, x_{N-1})}{w_k} - \lambda \\
&= \frac{\sum_{1 \leq i \leq N} \omega(x_k) \cdot p_i(x_k) \cdot \prod_{j \neq i} (\omega \models p_j)}{w_k} - \lambda \\
\frac{\partial H}{\partial \lambda}(\vec{w}, \lambda) &= 1 - \sum_{1 \leq i \leq n} w_i.
\end{aligned}$$

Setting these to zero gives:

$$1 = \sum_{1 \leq k \leq n} w_k = \frac{\sum_{1 \leq i \leq N} (\omega \models p_i) \cdot \prod_{j \neq i} (\omega \models p_j)}{\lambda} = \frac{N \cdot \prod_{1 \leq j \leq N} (\omega \models p_j)}{\lambda}.$$

Hence $\lambda = N \cdot \prod_{1 \leq j \leq N} (\omega \models p_j)$ so that:

$$w_k = \frac{\sum_{1 \leq i \leq N} \omega(x_k) \cdot p_i(x_k) \cdot \prod_{j \neq i} (\omega \models p_j)}{\lambda} = \sum_{1 \leq i \leq N} \frac{1}{N} \cdot \frac{\omega(x_k) \cdot p_i(x_k)}{\omega \models p_i} = \sum_{1 \leq i \leq N} \frac{1}{N} \cdot \omega|_{p_i}(x_k).$$

Thus, the better distribution $\omega' = \sum_i w_i |x_i\rangle$ is $\sum_i \frac{1}{N} \cdot \omega|_{p_i}$.

2 Directly from (1), by writing $\prod_i (\omega' \models p_i)^{k_i}$ as a k-fold product, for $k = \sum_i k_i$. ❑

Remark 8.2.6 The validity increase in Proposition 8.2.5 (2) is closely related to the divergence decrease in Jeffrey's rule from Theorem 7.9.3 (2). In fact, the proposition provides an alternative proof of the decrease, for the special case when the evidence distribution $\tau \in \mathcal{D}(Y)$ in Jeffrey's rule is fractional, say of the form $\tau = \sum_i \frac{k_i}{k} |y_i\rangle$, where $k = \sum_i k_i > 0$ and $Y = \bigcup_i \{y_i\}$, so that τ has full support. For a prior distribution $\omega \in \mathcal{D}(X)$ and a channel $c\colon X \rightsquigarrow Y$, as in Proposition 8.2.5, we consider the predicates $p_i := c^*(\mathbf{1}_{y_i})$. The posterior ω' defined there is now a Jeffrey update, since:

$$\omega' = \sum_i \frac{k_i}{k} \cdot \omega|_{p_i} = \sum_i \tau(y_i) \cdot \omega|_{c^*(\mathbf{1}_{y_i})} \overset{(7.1)}{=} \sum_i \tau(y_i) \cdot c^\dagger_\omega(y_i) = (c^\dagger_\omega)_*(\tau).$$

We now consider the following sequence of implications, starting from the inequality in Proposition 8.2.5 (2), to which we apply the (order-presering) natural logarithm ln — like in Exercise 3.4.1.

$$\begin{aligned}
&\prod_i (\omega' \models p_i)^{k_i} \geq \prod_i (\omega \models p_i)^{k_i} \\
&\Longrightarrow \sum_i k_i \cdot \ln(\omega' \models p_i) \geq \sum_i k_i \cdot \ln(\omega \models p_i) \\
&\Longrightarrow -\sum_i \frac{k_i}{k} \cdot \ln(\omega' \models p_i) - H(\tau) \leq -\sum_i \frac{k_i}{k} \cdot \ln(\omega \models p_i) - H(\tau) \\
&\Longrightarrow \sum_i \tau(y_i) \cdot \ln\left(\frac{1}{\omega' \models c^*(\mathbf{1}_{y_i})}\right) + \sum_i \tau(y_i) \cdot \ln(\tau(y_i)) \\
&\qquad \leq \sum_i \tau(y_i) \cdot \ln\left(\frac{1}{\omega' \models c^*(\mathbf{1}_{y_i})}\right) + \sum_i \tau(y_i) \cdot \ln(\tau(y_i)) \\
&\Longrightarrow \sum_i \tau(y_i) \cdot \ln\left(\frac{\tau(y_i)}{c_*(\omega')(y_i)}\right) \leq \sum_i \tau(y_i) \cdot \ln\left(\frac{\tau(y_i)}{c_*(\omega)(y_i)}\right) \\
&\Longrightarrow D_{KL}\big(\tau, c_*(\omega')\big) \leq D_{KL}\big(\tau, c_*(\omega)\big).
\end{aligned}$$

Thus we can derive Jeffrey's divergence decrease (in Theorem 7.9.3 (2)) from Proposition 8.2.5 (2) — for the special case that the evidence distribution τ is fractional.

In the remainder of this section we reconsider the earlier investment example with its six different evaluations of the validity multiple predicates in presence of predicate transformation.

Example 8.2.7 Recall the situation in Exercise 8.1.2 with a prior inflation distribution $\omega \in \mathcal{D}(\{h, \ell\})$, an investment channel $c\colon \{h, \ell\} \rightsquigarrow \{m, i, s\}$, and three predicates $q_1, q_2, q_3 \in Pred(\{m, i, s\})$. We saw six different possible validity interpretations to evaluate these predicates, labeled as (a) – (f). We show how to increase these validities via updating. The first three cases (a) – (c) can be

done in a standard way via Pearl's rule — from Definition 7.9.1 (1). For the last three cases we use the new inequality of Proposition 8.2.5 (2).

Below, we describe in each of the six cases an appropriate updated posterior distribution, with the associated increase validity. The increases are tiny, in each case, but that is not so relevant.

(a) This first case involves the validity $\omega \models c^*(q_1 \,\&\, q_2 \,\&\, q_3)$. This a clear case of Pearl updating, leading to the following posterior inflation distribution.

$$\omega_{(a)} \coloneqq \omega|_{c^*(q_1 \,\&\, q_2 \,\&\, q_3)} = 0.4314|h\rangle + 0.5686|\ell\rangle.$$

The new and old validities are:

$$\omega_{(a)} \models c^*(q_1 \,\&\, q_2 \,\&\, q_3) = 0.08807 \geq 0.08795 = \omega \models c^*(q_1 \,\&\, q_2 \,\&\, q_3).$$

(b) The next validity involves of the form $\omega \models c^*(q_1 \,\&\, q_2) \,\&\, c^*(q_3)$. This can be handled similarly, via one update with a conjunction, or via two updates, see Lemma 6.1.7 (3):

$$\omega_{(b)} \coloneqq \omega|_{c^*(q_1 \,\&\, q_2) \,\&\, c^*(q_3)} = \omega|_{c^*(q_1 \,\&\, q_2)}|_{c^*(q_3)} = 0.4466|h\rangle + 0.5534|\ell\rangle.$$

This time there is a really small increase:

$$\begin{aligned}\omega_{(b)} \models c^*(q_1 \,\&\, q_2) \,\&\, c^*(q_3) &= 0.18028 \\ &\geq 0.18027 = \omega \models c^*(q_1 \,\&\, q_2) \,\&\, c^*(q_3).\end{aligned}$$

(c) The third case involves a validity of the form $\omega \models c^*(q_1) \,\&\, c^*(q_2) \,\&\, c^*(q_3)$, so that Pearl's rule stil applies:

$$\omega_{(c)} \coloneqq \omega|_{c^*(q_1) \,\&\, c^*(q_2) \,\&\, c^*(q_3)} = \omega|_{c^*(q_1)}|_{c^*(q_2)}|_{c^*(q_3)} = 0.4519|h\rangle + 0.5481|\ell\rangle.$$

We now have an even smaller increase:

$$\begin{aligned}\omega_{(c)} \models c^*(q_1) \,\&\, c^*(q_2) \,\&\, c^*(q_3) &= 0.146993 \geq 0.146991 \\ &= \omega \models c^*(q_1) \,\&\, c^*(q_2) \,\&\, c^*(q_3).\end{aligned}$$

(d) At this stage things are getting more interesting since the validity in this case involves two copies of the distribution: $(\omega \models c^*(q_1 \,\&\, q_2))\cdot(\omega \models c^*(q_3))$. In this situation Proposition 8.2.5 (1) applies, with $N = 2$ predicates $p_1 = c^*(q_1 \,\&\, q_2)$ and $p_2 = c^*(q_3)$. According to the proposition we can increase the product of validities via a convex sum of updates, which here takes the form:

$$\omega_{(d)} \coloneqq \tfrac{1}{2} \cdot \omega|_{c^*(q_1 \,\&\, q_2)} + \tfrac{1}{2} \cdot \omega|_{c^*(q_3)} = 0.4469|h\rangle + 0.5531|\ell\rangle.$$

We then also have a small increase:

$$\begin{aligned}\big(\omega_{(d)} \models c^*(q_1 \,\&\, q_2)\big) \cdot \big(\omega_{(d)} \models c^*(q_3)\big) &= 0.18072 \geq 0.18071 \\ &= \big(\omega \models c^*(q_1 \,\&\, q_2)\big) \cdot \big(\omega \models c^*(q_3)\big).\end{aligned}$$

(e) Next we again have a product of two validities, but of a slightly different form, namely $(\omega \models c^*(q_1) \,\&\, c^*(q_2)) \cdot (\omega \models c^*(q_3))$. This leads to an update:

$$\omega_{(e)} \coloneqq \tfrac{1}{2} \cdot \omega|_{c^*(q_1) \,\&\, c^*(q_2)} + \tfrac{1}{2} \cdot \omega|_{c^*(q_3)} = \tfrac{1}{2} \cdot \omega|_{c^*(q_1)}|_{c^*(q_2)} + \tfrac{1}{2} \cdot \omega|_{c^*(q_3)} = 0.451|h\rangle + 0.549|\ell\rangle.$$

Again there is a small increase:

$$\begin{aligned}\big(\omega_{(e)} \models c^*(q_1) \,\&\, c^*(q_2)\big) \cdot \big(\omega_{(e)} \models c^*(q_3)\big) &= 0.147161 \geq 0.147159 \\ &= \big(\omega \models c^*(q_1) \,\&\, c^*(q_2)\big) \cdot \big(\omega \models c^*(q_3)\big).\end{aligned}$$

(f) The final case involves evaluation in three separate copies of the distribution: $(\omega \models c^*(q_1)) \cdot (\omega \models c^*(q_2)) \cdot (\omega \models c^*(q_3))$. We now apply Proposition 8.2.5 (1) with three predicates, giving as posterior:

$$\omega_{(f)} \coloneqq \tfrac{1}{3} \cdot \omega|_{c^*(q_1)} + \tfrac{1}{3} \cdot \omega|_{c^*(q_2)} + \tfrac{1}{3} \cdot \omega|_{c^*(q_3)} = 0.4507|h\rangle + 0.5493|\ell\rangle.$$

Also in this final case there is an increase:

$$\begin{aligned}&\big(\omega_{(f)} \models c^*(q_1)\big) \cdot \big(\omega_{(f)} \models c^*(q_2)\big) \cdot \big(\omega_{(f)} \models c^*(q_3)\big)\\ &= 0.147324 \geq 0.147323 = \big(\omega \models c^*(q_1)\big) \cdot \big(\omega \models c^*(q_2)\big) \cdot \big(\omega \models c^*(q_3)\big)\end{aligned}$$

In the next section we look more systematically at such updates, with multisets of factors as evidence.

Exercises

8.2.1 Recall the prior disease distribution ω and the positive and negative tests pt, nt from Example 8.1.1. Apply in this situation the different forms of updating used in Example 8.2.7 and show that it yields a 'dependent' update ω_D and an 'independent' update ω_I of the form:

$$\omega_D \coloneqq \omega|_{pt \,\&\, nt} \qquad\qquad \omega_I \coloneqq \tfrac{1}{2} \cdot \omega|_{pt} + \tfrac{1}{2} \cdot \omega|_{nt}.$$

1 Calculate these distributions ω_D and ω_I, using the formulas for ω, pt and nt from Example 8.1.1.

2 Calculate the 'prior' validities given by the two validity diagrams in (8.2), using the prior distribution ω.

3 Calculated also the two 'posterior' validities, for ω_D in the string diagram on the left in (8.2) and for ω_I in the diagram on the right. Notice the increases of validities with respect to the previous point.

8.2.2 Check that the better channel that appears in Example 8.2.4 is $c|_q$, like in Example 8.2.3.

8.2.3 Let a distribution $\omega \in \mathcal{D}(X)$ and a factor $p\colon \mathcal{N}[K](X) \to \mathbb{R}_{\geq 0}$ be given, for some number $K \geq 0$. Show that the approach of Lemma 8.2.1 for finding a 'better' distribution ω' for the validity $mn[K](\omega) \models p$ yields:

$$\omega' = \mathit{Flrn}_*\Big(mn[K](\omega)\big|_p\Big),$$

so that $mn[K](\omega) \models p \leq mn[K](\omega') \models p$ for this ω'.

Check that Example 3.4.3 (1) is a special case, when p is a point predicate $\mathbf{1}_\varphi$.

8.2.4 Let a distributions $\sigma, \tau \in \mathcal{D}(X)$ and channels $c, d\colon X \rightsquigarrow Y$ be given.

1 Prove that for any $y \in Y$,

$$\sum_{x \in X} c^\dagger_\sigma(y)(x) \cdot \ln\left(\frac{\tau(x) \cdot d(x)(y)}{\sigma(x) \cdot c(x)(y)}\right) \geq 0 \Longrightarrow d_*(\tau)(y) \geq c_*(\sigma)(y).$$

This result is a folklore probabilistic variation of Lemma 8.2.1.

2 Use this implication to find a better distribution $\omega' \in \mathcal{D}(X)$ for the validity $\omega \models p$ of a predicate p, by considering this predicate as a channel $p\colon X \rightsquigarrow \mathbf{2}$.

8.2.5 Let $\omega \in \mathcal{D}(X)$ and $p \in \mathit{Fact}(X)$ be given, with non-zero validity $\omega \models p$.

1 Use the inequality $(\omega \models p) \cdot D_{KL}(\omega|_p, \omega) \geq 0$, obtained from Proposition 2.10.6 (1), to prove:

$$\sum_{x \in X} \omega(x) \cdot p(x) \cdot \ln\left(\frac{\omega|_p(x)}{\omega(x)}\right) \geq 0.$$

2 Use this inequality to obtain the inequality $\omega|_p \models p \geq \omega \models p$ from Theorem 6.1.6, in the style of the proof of Lemma 8.2.1, by showing that the natural logarithm of their fraction is non-negative.x

3 In Example 8.2.2 the updated distribution $\omega|_p$ is constructed with an inequality:

$$\sum_{x \in X} \omega(x) \cdot p(x) \cdot \ln\big(\omega|_p(x) \cdot p(x)\big) \geq \sum_{x \in X} \omega(x) \cdot p(x) \cdot \ln\big(\omega(x) \cdot p(x)\big).$$

Deduce this inequality from the first item of this Exercise.

4 Use the first item also to derive the inequality:

$$\sum_{x \in X} \omega(x) \cdot p(x) \cdot \ln(p(x)) \geq \ln(\omega \models p).$$

Hint: Recall that $r^r \geq r$ holds for $r \in \mathbb{R}_{\geq 0}$.

5 Deduce the following formulation: for finitely many numbers $r_i \in [0, 1]$ with $\sum_i r_i = 1$ and for $a_i \in \mathbb{R}_{\geq 0}$ one has:

$$\prod_i a_i^{r_i \cdot a_i} \geq \sum_i r_i \cdot a_i.$$

8.2.6 In Proposition 6.12.8 we have seen an updated distribution $\omega' = \omega|_{e^p}$ that gives an increase of mutual information. Here we show that this distribution ω' can also be obtained via the sum-increase lemma of this section. This approach is used in the original source (Blahut, 1972, Thm. 1). We assume a finite set X and a channel $c \colon X \to\!\!\!\!\!\circ\; Y$. Define for $\omega \in \mathcal{D}(X)$ and $d \colon Y \to\!\!\!\!\!\circ\; X$,

$$J_c(\omega, d) := \sum_{x \in X, y \in Y} \omega(x) \cdot c(x)(y) \cdot \ln\left(\frac{d(y)(x)}{\omega(x)}\right).$$

1 Show that $MI(\langle id, c\rangle_*(\omega)) = J_c(\omega, c^\dagger_\omega)$.

2 Prove that for fixed $\omega \in \mathcal{D}(X)$,

$$c^\dagger_\omega \in \operatorname*{argmax}_{d \colon Y \to\!\!\!\!\!\circ\; X} J_c(\omega, d).$$

Hint: Use (2.77).

3 Now fix a distribution $\omega \in \mathcal{D}(X)$ and a channel $d \colon Y \to\!\!\!\!\!\circ\; X$. Show that using Lemma 8.2.1 to find a 'better' distribution $\omega' \in \mathcal{D}(X)$, that is with $J_c(\omega', d) \geq J_c(\omega, d)$, yields:

$$\omega' = Flrn\left(\sum_{x \in X} \prod_{y \in Y} d(y)(x)^{c(x)(y)} \,|x\rangle\right).$$

4 Check that when we apply the previous point with $d = c^\dagger_\omega$ we get, as in Proposition 6.12.8,

$$\omega' = \omega|_{e^p} \quad \text{where} \quad p = D_{KL}(c(-), c_*(\omega)).$$

5 Let $\omega \in \mathcal{D}(X)$ be given. Check the following sequence of inequalities proving an increase of mutual information.

$$\begin{aligned} MI(\langle id, c\rangle_*(\omega)) &= J_c(\omega, c^\dagger_\omega) && \text{by point (1)} \\ &\leq J_c(\omega', c^\dagger_\omega) && \text{via points (3) and (4)} \\ &\leq J_c(\omega', c^\dagger_{\omega'}) && \text{by point (2)} \\ &= MI(\langle id, c\rangle_*(\omega')). \end{aligned}$$

8.3 Validity and updating with multisets of factors

In this section we take a more systematic look at what we have seen in the previous two sections, following Jacobs (2024). We define 'evidence', as basis for updating a distribution, in terms of a (natural) multiset of factors. Evidence thus involves multiple data items for which the order does not matter — but the multiplicities of the items do matter.

We first precisely define such evidence, with different appearances.

Definition 8.3.1 1 What we call evidence is a natural multiset of factors over the same underlying set. Thus, if X is this common set, then evidence is a multiset of the form $\psi \in \mathcal{N}(Fact(X))$.

2 We use the following three special cases.

- When all factors in the multiset are predicates, the evidence may be identified with a multiset $\psi \in \mathcal{N}(Pred(X))$.
- When evidence $\psi \in \mathcal{N}(Pred(X))$ consists only of point predicates $\mathbf{1}_x$, this ψ may be identified with a multiset $\psi \in \mathcal{N}(X)$ over X. We may then call ψ point evidence.
- A single factor $p\colon X \to \mathbb{R}_{\geq 0}$ may also be called evidence, as singleton multiset $1|p\rangle$.

3 Given a channel $c\colon Z \rightsquigarrow X$ with evidence $\psi \in \mathcal{N}(Fact(X))$ on its codomain, one can turn it into evidence $c^{**}(\psi) \in \mathcal{N}(Fact(Z))$ on the channel's domain via a combination of funcoriality (of $\mathcal{N}$) and observation transformation $c^*\colon Fact(X) \to Fact(Z)$. The result will be written as:

$$c^{**}(\psi) \coloneqq \mathcal{N}(c^*)(\psi) = \sum_{p \in supp(\psi)} \psi(p) \Big| c^*(p) \Big\rangle. \tag{8.3}$$

This operation $c^{**}\colon \mathcal{N}(Fact(X)) \to \mathcal{N}(Fact(Z))$ will be called evidence transformation, along the channel c.

In Example 8.1.1 we had positive and negative test predicates, pt and nt, which we can now describe as a multiset $1|pt\rangle + 1|nt\rangle$ of evidence. In Example 8.1.2 the three forms of advice form evidence of the form $1|q_1\rangle + 1|q_2\rangle + 1|q_3\rangle$.

In ordered form, as a list, point evidence is also called a feature vector, especially in machine learning.

We turn to validity of evidence in a particular distribution. We generalise the two forms that we used for the medical test example, see the different graphical descriptions in (8.2). We associate these two forms with Jeffrey and Pearl, where the diagram on the right in (8.2) captures Jeffrey validity and the diagram on left is for what we call Pearl validity. The justification for using the same names as for the two forms of updating in Section 7.9 follows later below.

Definition 8.3.2 Let $\omega \in \mathcal{D}(X)$ be a distribution and let $\psi \in \mathcal{N}(Fact(X))$ be an evidence multiset of size $K = \|\psi\|$.

1 The *Jeffrey validity* $\omega \models_{\mathrm{J}} \psi$ of ψ in ω is defined as:

$$\begin{aligned} \omega \models_{\mathrm{J}} \psi &:= (\psi) \cdot \prod_{p \in supp(\psi)} (\omega \models p)^{\psi(p)} \\ &= (\psi) \cdot \left(\underbrace{\omega \otimes \cdots \otimes \omega}_{K \text{ times}} \models \bigotimes_{p \in supp(\psi)} \underbrace{p \otimes \cdots \otimes p}_{\psi(p) \text{ times}} \right). \end{aligned} \tag{8.4}$$

This last equation holds by Lemma 4.4.8. Recall that the multiset coefficient (ψ) is $\frac{\|\psi\|!}{\prod_p \psi(p)!}$.

2 The *Pearl validity* $\omega \models_{\mathrm{P}} \psi$ of ψ in distribution ω is:

$$\begin{aligned} \omega \models_{\mathrm{P}} \psi &:= (\psi) \cdot \left(\omega \models \mathop{\&}_{p \in supp(\psi)} \underbrace{p \,\&\, \cdots \,\&\, p}_{\psi(p) \text{ times}} \right) \\ &= (\psi) \cdot \left(\mathcal{D}(\Delta)(\omega) \models \bigotimes_{p \in supp(\psi)} \underbrace{p \otimes \cdots \otimes p}_{\psi(p) \text{ times}} \right). \end{aligned} \tag{8.5}$$

The diagonal $\Delta \colon X \to X^K$ is the K-fold copier $x \mapsto (x, \ldots, x)$. The last equation holds by Exercise 4.5.9. We shall abbreviate the (sequential) conjunction of evidence as:

$$\&\,\psi := \mathop{\&}_{p \in supp(\psi)} p^{\psi(p)} \quad \text{where, in general} \quad p^n = p \,\&\, \cdots \,\&\, p. \tag{8.6}$$

We first observe that Jeffrey and Pearl validity coincide for singleton evidence. For the two-element evidence $\psi = 1|pt\rangle + 1|nt\rangle$ from Example 8.1.1 and distribution ω we see differences. Jeffrey evidence is $\omega \models_{\mathrm{J}} \psi = (\omega \models pt) \cdot (\omega \models nt)$, as on the right in (8.2), corresponding to independent tests, performed on two arbitrary individuals from the population. In contrast, the Pearl validity $\omega \models_{\mathrm{P}} \psi = \omega \models pt \,\&\, nt$, as on the left in (8.2), corresponds to dependent tests, performed on the same individual.

Jeffrey validity of point evidence can be expressed in terms of multinomial distributions, as will be shown below. Pearl validity of point evidence is zero, except for a trivial singleton multiset, since $\mathbf{1}_x \,\&\, \mathbf{1}_y = \mathbf{0}$ when $x \neq y$.

Lemma 8.3.3 *Consider a distribution $\omega \in \mathcal{D}(X)$ with point evidence $\psi \in \mathcal{N}(X)$.*

1 Jeffrey validity equals multinomial probability:

$$\omega \models_{\mathrm{J}} \psi = mn[K](\omega)(\psi).$$

2 As a result, the highest Jeffrey validity of the point evidence ψ is reached at the distribution $Flrn(\psi)$, that is:

$$Flrn(\psi) \in \operatorname*{argmax}_{\omega \in \mathcal{D}(X)} \omega \models_{\mathrm{J}} \psi.$$

Proof. 1 Since:

$$\omega \models_{\mathrm{J}} \psi \overset{(8.4)}{=} (\psi) \cdot \prod_{x \in X} \left(\omega \models \mathbf{1}_x\right)^{\psi(x)} = (\psi) \cdot \prod_{x \in X} \omega(x)^{\psi(x)} \overset{(2.54)}{=} mn[K](\omega)(\psi).$$

2 By the previous point and Example 3.4.3 (1). ❑

Remark 8.3.4 The coefficients $\binom{\psi}{} = \frac{\|\psi\|!}{\prod_p \psi(p)!}$ in the formulas (8.4) and (8.5) are included in order to take the different orders in which the evidence can be presented into account, see Proposition 1.7.2. These numbers $\binom{\psi}{}$ involve multiple factorials and thus may become too large to calculate when the evidence ψ consists of hundreds of items (or more). Correspondingly, the terms $\prod_p (\omega \models p)^{\psi(p)}$ and $\omega \models \&\psi$ in (8.4) and (8.5) may become unmanageably small. A way out in the Jeffrey case is to ignore the coefficient $\binom{\psi}{}$ and work with what is called the log-likelihood. We use special variations $\underset{\text{JL}}{\models}$ and $\underset{\text{PL}}{\models}$ of the Jeffrey and Pearl validity signs $\underset{\text{J}}{\models}$ and $\underset{\text{P}}{\models}$ with an extra 'L' for logarithm. They are defined in the following way.

$$\omega \underset{\text{JL}}{\models} \psi \coloneqq \ln\left(\prod_{p \in supp(\psi)} (\omega \models p)^{\psi(p)}\right) = \sum_{p \in supp(\psi)} \psi(p) \cdot \ln\big(\omega \models p\big). \tag{8.7}$$

For the Pearl one uses a slightly different formulation:

$$\begin{aligned} \omega \underset{\text{PL}}{\models} \psi \coloneqq \omega \models \ln(\&\psi) &= \omega \models \ln\left(\prod_{p \in supp(\psi)} \ln(p)^{\psi(p)}\right) \\ &= \omega \models \left(\sum_{p \in supp(\psi)} \psi(p) \cdot \ln(p)\right) = \sum_{p \in supp(\psi)} \psi(p) \cdot \big(\omega \models \ln(p)\big). \end{aligned} \tag{8.8}$$

This log-version of Pearl is less common.

We now give general formulations for the Jeffrey and Pearl update mechanisms, with evidence. We have used the vertical bar | for Bayesian updating in the notation $\omega|_p$. For the Jeffrey and Pearl versions with evidence we make subtle variations: we change the vertical bar | into $\rfloor$ for the Jeffrey update and into $\rceil$ for the Pearl update, referring respectively to the letters 'J' and 'P'.

Definition 8.3.5 Let $\omega \in \mathcal{D}(X)$ be a distribution with an evidence multiset of factors $\psi \in \mathcal{N}(Fact(X))$.

1 The Jeffrey update $\omega\rfloor_\psi \in \mathcal{D}(X)$ of the distribution ω with the evidence ψ is defined as a convex sum of Bayesian updates with separate factors:

$$\omega\rfloor_\psi \coloneqq \sum_{p \in supp(\psi)} Flrn(\psi)(p) \cdot \omega|_p. \tag{8.9}$$

2 The Pearl update $\omega\rceil_\psi \in \mathcal{D}(X)$ is the Bayesian update with a single factor, namely with the conjunction $\&\psi = \&_p\, p^{\psi(p)}$, as in:

$$\omega\rceil_\psi \coloneqq \omega\big|_{\&\psi} = \omega\big|_{\&_p\, p^{\psi(p)}}. \tag{8.10}$$

For the special case where ψ is singleton evidence $1|p\rangle$, the Jeffrey and Pearl updates with ψ coincide, and are the same as a Bayesian update with p. In the cognitive setting of (Smith et al., 2022, §1.2), updating after observing a single piece of (point) evidence is called perception, whereas learning is updating after many observations.

Recall Subsection 7.4.1 where we elaborated an example of learning with missing data: the missing items where handled not by a point predicate but by a truth predicate. There we defined a joint distribution of the form $\tau = \sum_i \frac{1}{5} \cdot unif|_{q_i}$. We now recognise it has a Jeffrey update with evidence $\sum_i 1|q_i\rangle$. A bit later on, in Subsection 7.4.2, we also used a Jeffrey update, but now with evidence in which some predicates occur multiple times. The two examples used there were taken from the literature and illustrate that the above form of Jeffrey updating is used, implicitly.

In Exercise 8.2.1 we have seen the two different update mechanisms, where the Jeffrey form was called 'independent' and the Pearl form 'dependent'. In Section 6.9, on learning the bias of a coin from a number of coin flips, we have used Pearl's updating since all the data applied to the same coin. In general, we use validity and updating according to Jeffrey in situations where the data are 'independent', that is, hold for separate instances (draws) from ω. In contrast, validity and updating according to Pearl are used in dependent situations, where the data apply to the same instance, see Exercise 8.1.2.

The remainder of this section we review the earlier three advice illustration in the light of the above update definitions $\rfloor$ and $\rceil$. Further investigation of the properties of these two forms of updating is postponed to the next section. There we will see expected results, like Jeffrey (resp. Pearl) updating increases Jeffrey (resp. Pearl) validity, see Theorems 8.4.6 and 8.4.7.

Example 8.3.6 Recall the three forms of advice predicates q_1, q_2, q_3 from Examples 8.1.2 and 8.2.7, in presence of a channel c and a distribution ω. We review the six different validities and updates, now using the terminology of Definition 8.3.5, with the same labels (a) – (f) that occurred before.

(a) This first case involves the validity $\omega \models c^*(q_1 \,\&\, q_2 \,\&\, q_3)$ and the update $\omega|_{c^*(q_1 \,\&\, q_2 \,\&\, q_3)}$. We now recognise it as a case of Pearl validity $\omega \models_{\mathrm{P}} \psi_{(a)}$ and Pearl update $\omega\rceil_{\psi_{(a)}}$ for the singleton evidence:

$$\psi_{(a)} \;=\; 1\big|c^*(q_1 \,\&\, q_2 \,\&\, q_3)\rangle \;=\; c^{**}(1|q_1 \,\&\, q_2 \,\&\, q_3\rangle).$$

Since Pearl and Jeffrey validity and updates coincides on singleton evidence, this case can also be described in terms of Jeffrey valdity and update.

(b) The next case involves validity $\omega \models_{\mathrm{P}} \psi_{(b)} \;=\; \omega \models c^*(q_1 \,\&\, q_2) \,\&\, c^*(q_3)$ with an Pearl update $\omega\rceil_{\psi_{(b)}} = \omega|_{c^*(q_1 \,\&\, q_2) \,\&\, c^*(q_3)}$ for the two-element evidence multiset:

$$\begin{aligned}\psi_{(b)} \;&=\; 1\big|c^*(q_1 \,\&\, q_2)\rangle + 1\,\big|\,c^*(q_3)\big\rangle \\ &=\; c^{**}(1|q_1 \,\&\, q_2\rangle) + c^{**}(1|q_3\rangle) \;=\; c^{**}(1|q_1 \,\&\, q_2\rangle + 1|q_3\rangle.\end{aligned}$$

(c) In the third case involves a validity of the form $\omega \models c^*(q_1) \,\&\, c^*(q_2) \,\&\, c^*(q_2)$. It is Pearl validity for the evidence multiset with three elements:

$$\begin{aligned}\psi_{(c)} \;&=\; 1\big|c^*(q_1)\rangle + 1\big|c^*(q_2)\rangle + 1\big|c^*(q_3)\rangle \\ &=\; c^{**}(1|q_1\rangle) + c^{**}(1|q_2\rangle) + c^{**}(1|q_3\rangle) \;=\; c^{**}(1|q_1\rangle + 1|q_2\rangle + 1|q_3\rangle).\end{aligned}$$

The update used in this case is indeed the corresponding one of Pearl, with three conjunctions: $\omega\rceil_{\psi_{(c)}} = \omega|_{c^*(q_1) \,\&\, c^*(q_2) \,\&\, c^*(q_3)}$.

(d) The next validity is of the form: $(\omega \models c^*(q_1 \,\&\, q_2)) \cdot (\omega \models c^*(q_3))$. This is Jeffrey validity $\omega \models_{\mathrm{J}} \psi_{(b)}$ for the evidence used in point (b). The corresponding Jeffrey update $\omega\rfloor_{\psi_{(b)}} = \frac{1}{2} \cdot \omega|_{c^*(q_1 \,\&\, q_2)} + \frac{1}{2} \cdot \omega|_{c^*(q_3)}$ is indeed the one that we used in Example 8.2.7 (d).

(e) We now look at a validity $(\omega \models c^*(q_1) \,\&\, c^*(q_2)) \cdot (\omega \models c^*(q_3))$. This is Jeffrey validity $\omega \models_{\mathrm{J}} \psi_{(e)}$ for the subtly different evidence multiset:

$$\psi_{(e)} \;=\; 1\big|c^*(q_1) \,\&\, c^*(q_2)\rangle + 1\big|c^*(q_3)\rangle.$$

The corresponding Jeffrey update $\omega\rfloor_{\psi_{(e)}}$ is $\frac{1}{2} \cdot \omega|_{c^*(q_1) \,\&\, c^*(q_2)} + \frac{1}{2} \cdot \omega|_{c^*(q_3)}$.

(f) The final case uses validity $(\omega \models c^*(q_1)) \cdot (\omega \models c^*(q_2)) \cdot (\omega \models c^*(q_3))$. It is Jeffrey validity $\omega \models_{\mathrm{J}} \psi_{(c)}$ with corresponding Jeffrey update: $\omega|_{\psi_{(c)}} = \frac{1}{3} \cdot \omega|_{c^*(q_1)} + \frac{1}{3} \cdot \omega|_{c^*(q_2)} + \frac{1}{3} \cdot \omega|_{c^*(q_3)}$.

We see that all six different forms of validities (see Figure 8.2) and updates that we distinguished before can be accomodated by using suitably different multisets of predicates as evidence.

Exercises

8.3.1 Continue Exercise 8.1.1 and describe the formula for the posterior disease probability distribution that the doctor may compute. Check that this posterior associated with the example described there is: $\frac{67}{25071}|d\rangle + \frac{25004}{25071}|d^\perp\rangle \approx 0.00267|d\rangle + 0.997|d^\perp\rangle$.

8.3.2 Let $c\colon X \rightsquigarrow Y$ and $d\colon Y \rightsquigarrow Z$ be two (composable) channels. Show that evidence transformation (8.3) satisfies:

$$(d \circ c)^{**} = c^{**} \circ d^{**} \qquad \text{and also} \qquad id^{**} = id.$$

8.3.3 Let $\omega \in \mathcal{D}(X)$, $c\colon X \rightsquigarrow Y$ and $\psi \in \mathcal{N}(Fact(Y))$ be given.

1 Show that Jeffrey validity satisfies the following analogue of Proposition 4.5.3.

$$\omega \models_{\mathrm{J}} c^{**}(\psi) = c_*(\omega) \models_{\mathrm{J}} \psi.$$

2 Assume now that ψ is point evidence. Consider the Jeffrey log-validity (8.7) and prove its relation below to KL-divergence and entropy, where $K = \|\psi\|$.

$$\omega \models_{\mathrm{JL}} c^{**}(\psi) = -K \cdot D_{KL}\Big(Flrn(\psi),\, c_*(\omega)\Big) - K \cdot H\Big(Flrn(\psi)\Big).$$

Notice the inverse relationship, for point evidence: in/de-creasing the log-validity corresponds to de/in-creasing the KL-divergence beteen the (normalised) evidence and the prediction.

8.3.4 Let $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ be two distributions with joint point evidence $\psi \in \mathcal{N}(X \times Y)$. We write $\psi_1 = \psi[1,0] = \mathcal{N}(\pi_1)(\psi) \in \mathcal{N}(X)$ and $\psi_2 = \psi[0,1] = \mathcal{N}(\pi_2)(\psi) \in \mathcal{N}(Y)$ for the two (point) evidence marginals. Show that:

$$\omega \otimes \rho \models_{\mathrm{J}} \psi = \frac{(\psi)}{(\psi_1) \cdot (\psi_2)} \cdot \Big(\omega \models_{\mathrm{J}} \psi_1\Big) \cdot \Big(\rho \models_{\mathrm{J}} \psi_2\Big).$$

8.3.5 Let $\omega, \rho \in \mathcal{D}(X)$ be two distributions with a number $r \in [0,1]$. Show that the Jeffrey validity of an evidence multiset $\psi \in \mathcal{N}[L](Fact(X))$ of size L, in the convex sum distribution, is:

$$r \cdot \omega + (1-r) \cdot \rho \models_{\mathrm{J}} \psi = \sum_{0 \le K \le L} bn[L](r)(K) \cdot \sum_{\varphi \le_K \psi} \Big(\omega \models_{\mathrm{J}} \varphi\Big) \cdot \Big(\rho \models_{\mathrm{J}} \psi - \varphi\Big).$$

8.3.6 Fix a set X and consider the set $\mathcal{N}(Fact(X))$ of evidence multisets on X. It is a commutative monoid, via multiset addition, see Lemma 1.6.3. In convolution style we introduce a multiplicative commutative monoid structure $\bullet$, for $\varphi, \psi \in \mathcal{N}(Fact(X))$, as:

$$\varphi \bullet \psi := \mathcal{N}(\&)(\varphi \otimes \psi) = \sum_{p,q \in Fact(X)} \varphi(p) \cdot \psi(q) \,\big|\, p \,\&\, q\rangle.$$

1 Check that $\bullet$ with unit $1|\mathbf{1}\rangle \in \mathcal{N}(Fact(X))$ forms a commutative monoid.

2 Check that $\varphi \bullet (-)$ and $(-) \bullet \psi$, as functions $\mathcal{N}(Fact(X)) \to \mathcal{N}(Fact(X))$, preserve the additive monoid structure. Conclude that $\mathcal{N}(Fact(X))$ is a commutative semiring.

3 Consider the set $\mathbb{R}^{\mathcal{D}(X)}$ of functions with the map $\alpha\colon \mathcal{N}(Fact(X)) \times \mathbb{R}^{\mathcal{D}(X)} \to \mathbb{R}^{\mathcal{D}(X)}$ defined, roughly as the 'averaged conditioning' from (Lang et al., 2025, §2):

$$\alpha(\varphi, f)(\omega) := \sum_{p \in Fact(X)} \varphi(p) \cdot (\omega \models p) \cdot f(\omega|_p).$$

Show that α is a monoid action, with respect to the multiplicative monoid structure $\bullet$ on $\mathcal{N}(Fact(X))$.

4 Show that $\alpha(\varphi, -)\colon \mathbb{R}^{\mathcal{D}(X)} \to \mathbb{R}^{\mathcal{D}(X)}$ and also $\alpha(-, f)\colon \mathcal{N}(Fact(X)) \to \mathbb{R}^{\mathcal{D}(X)}$ preserve the additive monoid structures.

(Abstractly we can now say that $\mathcal{N}(Fact(X))$, as a commutative semiring, is a commutative monoid in the category **CMon** of commutative monoids and that α is an action in **CMon**.)

5 Call a multiset $\varphi \in \mathcal{N}(Fact(X))$ a *test* if $\sum_p \varphi(p) \cdot p = \mathbf{1}$. Check that the multiplicative monoid structure restricts to tests.

8.3.7 Recall the election poll from Example 6.9.4.

1 Describe the polling data given there as point evidence.

2 Observe that the example contains a Pearl update with a transformation of this point evidence. It leads to a chance that candidate A wins of over 60%.

3 The chance that A wins after a Jeffrey update in this situation approximates 32%. Proving this requires a variation on Lemma 1.8.8. We describe the main steps.

- Let set X have $N > 0$ elements, with $a \in X$, and let K be an even number. Show that:

$$\sum_{\varphi \in \mathcal{N}[K](X),\, Flrn(\varphi)(a) > \frac{1}{2}} \varphi(a) = \left(1 + \frac{1}{N}\right) \cdot \frac{K}{2} \cdot \left(\!\!\binom{N}{K/2}\!\!\right).$$

And for $b \in X$, $b \neq a$,

$$\sum_{\varphi \in \mathcal{N}[K](X),\, Flrn(\varphi)(a) > \frac{1}{2}} \varphi(b) = \frac{K}{2N} \cdot \left(\!\!\binom{N}{K/2}\!\!\right).$$

- For fixed $N > 0$,

$$\lim_{K \to \infty} \frac{\left(\!\!\binom{N}{K/2}\!\!\right)}{\left(\!\!\binom{N}{K}\!\!\right)} = \frac{1}{2^{N-1}}.$$

- Show, in the setting of Example 6.9.4 that as the fractional distribution parameter $K \to \infty$, one has:

$$\upsilon|_{i^*(\mathbf{1}_a)} \models aw \to \tfrac{1}{2} \qquad \upsilon|_{i^*(\mathbf{1}_b)} \models aw \to \tfrac{1}{8} \qquad \upsilon|_{i^*(\mathbf{1}_c)} \models aw \to \tfrac{1}{8}.$$

Use this to show that $\upsilon|_{i^{**}(\psi)} \models aw \to \frac{32}{100}$.

Which update mechanism is most appropriate to learn from polling information?

8.4 Properties of updating with multisets of factors

The previous section has defined Jeffrey and Pearl style validity and updating for evidence, in the form of a multiset of factors. This section digs deeper into these operations and describes several basic properties. The situation is illustrated in several examples from the literature. The examples show that the update rules of Pearl and Jeffrey are common. But mostly, they are not recognised and identified as two different update mechanisms.

We first look at the common update cases where evidence comes via transformation along a channel from point evidence. This brings us closer to the earlier forms of the update rules of Jeffrey and Pearl — as will be discussed subsequently, in Remark 8.4.5.

Proposition 8.4.1 *Consider a (prior) distribution $\omega \in \mathcal{D}(X)$ and a channel $c\colon X \multimap Y$ together with point evidence $\varphi \in \mathcal{N}(Y)$ on the channel's codomain Y. In that case we can form evidence $c^{**}(\varphi) = \sum_y \varphi(y)\big|c^*(\mathbf{1}_y)\rangle$ on the domain X via transformation along the channel c.*

1 *In this situation the Jeffrey update with the transformed evidence can be expressed via the dagger $c^\dagger_\omega\colon Y \multimap X$ of the channel, as:*

$$\omega\big|_{c^{**}(\varphi)} = \big(c^\dagger_\omega\big)_*(\mathit{Flrn}(\varphi)). \tag{8.11}$$

This is an instance of Jeffrey's update rule in Definition 7.9.1 (2).

2 *The Pearl update with the transformed evidence $c^{**}(\varphi)$ can expressed via the multinomial channel $mn[K](c) := mn[K] \circ c\colon X \to \mathcal{D}(Y) \to \mathcal{D}(\mathcal{N}[K](Y))$, for $K = \|\varphi\|$, namely as:*

$$\omega\big|^{\flat}_{c^{**}(\varphi)} = \omega\big|_{mn[K](c)^*(\mathbf{1}_\varphi)} = mn[K](c)^\dagger_\omega(\varphi). \tag{8.12}$$

This is an instance of Pearl's update rule in Definition 7.9.1 (1) involving an update with a point predicate $\mathbf{1}_\varphi \in \mathit{Pred}(\mathcal{N}[K](Y))$.

Proof. 1 Since:

$$\begin{aligned}
\omega\big|_{c^{**}(\varphi)} &\overset{(8.9)}{=} \sum_{p \in \mathit{supp}(\psi)} \mathit{Flrn}(c^{**}(\varphi))(p) \cdot \omega\big|_{c^{**}(\varphi)(p)} \\
&= \sum_{y \in \mathit{supp}(\varphi)} \mathit{Flrn}(\varphi)(y) \cdot \omega\big|_{c^*(\mathbf{1}_y)} \overset{(7.1)}{=} \sum_{y \in \mathit{supp}(\varphi)} \mathit{Flrn}(\varphi)(y) \cdot c^\dagger_\omega(y) = \big(c^\dagger_\omega\big)_*(\mathit{Flrn}(\varphi)).
\end{aligned}$$

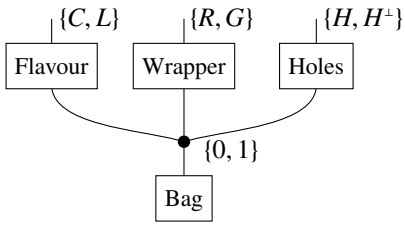
2 We translate from Bayesian update to Pearlian update:

$$\begin{aligned}
\omega\big|_{mn[K](c)^*(\mathbf{1}_\varphi)} &\overset{(6.1)}{=} \sum_{x \in X} \frac{\omega(x) \cdot mn[K](c)^*(\mathbf{1}_\varphi)(x)}{\omega \models mn[K](c)^*(\mathbf{1}_\varphi)} \big|x\rangle \\
&= \sum_{x \in X} \frac{\omega(x) \cdot mn[K](c(x))(\varphi)}{mn[K](c)_*(\omega)(\varphi)} \big|x\rangle \\
&= \sum_{x \in X} \frac{\omega(x) \cdot (\varphi) \cdot \prod_y c(x)(y)^{\varphi(y)}}{\sum_z \omega(z) \cdot (\varphi) \cdot \prod_y c(z)(y)^{\varphi(y)}} \big|x\rangle \\
&= \sum_{x \in X} \frac{\omega(x) \cdot (\&_y\, (c^*(\mathbf{1}_y))^{\varphi(y)})(x)}{\sum_z \omega(z) \cdot (\&_y\, (c^*(\mathbf{1}_y))^{\varphi(y)})(z)} \big|x\rangle \overset{(6.1)}{=} \omega\big|_{\&_y\, (c^*(\mathbf{1}_y))^{\varphi(y)}} = \omega\big|^{\flat}_{c^{**}(\varphi)}.
\end{aligned}$$

❑

The next illustration from the literature uses Jeffrey's update — implicitly in its original formulation. We show here that the update can be described in terms of a pushfoward along a dagger channel, as in the first point of the above proposition.

Example 8.4.2 Consider the following Bayesian network from (Russell and Norvig, 2003, §20.3).



(8.13)

It involves two bags, labelled 0 and 1, containing candies. These candies can have one of two flavours, cherry (C) or lime (L), with a red (R) or green (G) wrapper, and with a hole (H) or not ($H^\perp$).

The network (8.13) involves a prior distribution $\rho = \frac{3}{5}|0\rangle + \frac{2}{5}|1\rangle \in \mathcal{D}(\mathbf{2})$ and three channels, written as $f\colon \mathbf{2} \rightsquigarrow \{C, L\}$, $w\colon \mathbf{2} \rightsquigarrow \{R, G\}$, $h\colon \mathbf{2} \rightsquigarrow \{H, H^\perp\}$. They have equal probabilities in (Russell and Norvig, 2003, (20.8)):

$$\begin{aligned} f(0) &= \tfrac{6}{10}|C\rangle + \tfrac{4}{10}|L\rangle & f(1) &= \tfrac{4}{10}|C\rangle + \tfrac{6}{10}|L\rangle \\ w(0) &= \tfrac{6}{10}|R\rangle + \tfrac{4}{10}|G\rangle & w(1) &= \tfrac{4}{10}|R\rangle + \tfrac{6}{10}|G\rangle \\ h(0) &= \tfrac{6}{10}|H\rangle + \tfrac{4}{10}|H^\perp\rangle & h(1) &= \tfrac{4}{10}|H\rangle + \tfrac{6}{10}|H^\perp\rangle. \end{aligned}$$

Here we combine these three channels into a single 3-tuple channel $\langle f, w, h\rangle = (f \otimes w \otimes h) \circ \Delta_3\colon \mathbf{2} \rightsquigarrow \{C, L\} \times \{R, G\} \times \{H, H^\perp\}$, as in Diagram (8.13). Explicitly:

$$\begin{aligned} \langle f, w, h\rangle(0) &= f(0) \otimes w(0) \otimes h(0) \\ &= \tfrac{27}{125}|C, R, H\rangle + \tfrac{18}{125}|C, R, H^\perp\rangle + \tfrac{18}{125}|C, G, H\rangle + \tfrac{12}{125}|C, G, H^\perp\rangle \\ &\quad + \tfrac{18}{125}|L, R, H\rangle + \tfrac{12}{125}|L, R, H^\perp\rangle + \tfrac{12}{125}|L, G, H\rangle + \tfrac{8}{125}|L, G, H^\perp\rangle \\ \langle f, w, h\rangle(1) &= f(1) \otimes w(1) \otimes h(1) \\ &= \tfrac{8}{125}|C, R, H\rangle + \tfrac{12}{125}|C, R, H^\perp\rangle + \tfrac{12}{125}|C, G, H\rangle + \tfrac{18}{125}|C, G, H^\perp\rangle \\ &\quad + \tfrac{12}{125}|L, R, H\rangle + \tfrac{18}{125}|L, R, H^\perp\rangle + \tfrac{18}{125}|L, G, H\rangle + \tfrac{27}{125}|L, G, H^\perp\rangle. \end{aligned}$$

The point evidence in Russell and Norvig (2003) involves 1000 data elements in total, about combinations of flavours, wrappers and holes. From these data we like to learn an updated distribution on bags, giving the probabilities that the candies that produce these data come from bag 0 or 1. The data forms point evidence $\psi \in \mathcal{N}(\{C, L\} \times \{R, G\} \times \{H, H^\perp\})$ with the following multiplicities.

$$\begin{aligned} \psi = {} & 273|C, R, H\rangle + 93|C, R, H^\perp\rangle + 104|C, G, H\rangle + 90|C, G, H^\perp\rangle \\ & + 79|L, R, H\rangle + 100|L, R, H^\perp\rangle + 94|L, G, H\rangle + 167|L, G, H^\perp\rangle. \end{aligned} \tag{8.14}$$

The update performed in Russell and Norvig (2003) is of Jeffrey's type, matching the independence of the data elements. The original formulation does not make explicit what update rule is used, but we can describe this update via a dagger, like in Section 7.9, or as Jeffrey update $\rfloor$ from Definition 8.3.5 (1). In the latter case we have to first transform the evidence multiset ψ into a multiset in $\mathcal{N}(\{0, 1\})$, consisting of transformed point predicates along the tuple channel $\langle f, w, h\rangle$.

The posterior $\rho' \in \mathcal{D}(\{0,1\})$ can then be described as Jeffrey update:

$$\begin{aligned}
\rho' \;=\; \rho\big|_{\langle f,w,h\rangle^{**}(\psi)} &\overset{(8.11)}{=} \Big(\langle f,w,h\rangle^{\dagger}_{\rho}\Big)_{*}(Flrn(\psi)) \\
&= \tfrac{273}{1000}\cdot\rho\big|_{\langle f,w,h\rangle^{*}(\mathbf{1}_{(C,R,H)})} + \cdots + \tfrac{167}{1000}\cdot\rho\big|_{\langle f,w,h\rangle^{*}(\mathbf{1}_{(L,G,H^{\perp})})} \\
&= \tfrac{30891}{50440}|0\rangle + \tfrac{19549}{50440}|1\rangle \;\approx\; 0.6124|0\rangle + 0.3876|1\rangle.
\end{aligned}$$

This probability 0.6124 for bag 0 is exactly as computed in (Russell and Norvig, 2003, §20.3). As an aside: Pearl's rule gives a different outcome, see Example 8.4.14. It is computed in an indirect manner, via a switch to logarithms.

In theory one could show an increase of Jeffrey validity of ψ, from prior ρ to posterior ρ'. However, the coefficient $\mathbf{(}\psi\mathbf{)}$ is too large to actually compute. What we can do is calculate the Jeffrey log-likelihoods (8.7), without the coefficient. This shows a small increase:

$$\begin{aligned}
\rho \models_{\mathrm{JL}} \langle f,w,h\rangle^{**}(\psi) &= \sum_{t\in supp(\psi)} \psi(t)\cdot\ln\Big(\rho \models \langle f,w,h\rangle^{*}(\mathbf{1}_t)\Big) \;=\; -2044.26 \\
\rho' \models_{\mathrm{JL}} \langle f,w,h\rangle^{**}(\psi) &= \sum_{t\in supp(\psi)} \psi(t)\cdot\ln\Big(\rho' \models \langle f,w,h\rangle^{*}(\mathbf{1}_t)\Big) \;=\; -2043.66.
\end{aligned}$$

This same prior Jeffrey log-likelihood occurs in Russell and Norvig (2003).

The previous example from the literature uses Jeffrey's update mechanism. There are also examples where Pearl's approach is used.

Example 8.4.3 In (Jaynes, 2003, §4.2) the following situation is described.

We have 11 automatic machines turning out widgets, which pour out of the machines into 11 boxes. This example corresponds to a very early stage in the development of widgets, because ten of the machines produce one in six defective. The 11th machine is even worse; it makes one in three defective.

The letter A is for the machine that performs worst and the letter B for any of the others. This gives a prior machine distribution $\omega = \frac{1}{11}|A\rangle + \frac{10}{11}|B\rangle$. If we use the letter d for defective, and $d^{\perp}$ for non-defective, then we can recognise a machine channel $mc\colon \{A,B\} \twoheadrightarrow \{d,d^{\perp}\}$ in the above description with:

$$mc(A) \;=\; \tfrac{1}{3}|d\rangle + \tfrac{2}{3}|d^{\perp}\rangle \qquad\qquad mc(B) \;=\; \tfrac{1}{6}|d\rangle + \tfrac{5}{6}|d^{\perp}\rangle.$$

The situation in Jaynes (2003) further involves a test of 12 produced widgets out of which 5 turn out to be defective. It is claimed that the posterior probability for A is around 0.4. There, the outcome is calculated via logarithms, as in Example 8.4.13 below.

Here we try to reconstruct this outcome in the present setting in order to find out which form of updating is used in Jaynes (2003). The tests translate into point evidence $\psi = 5|d\rangle + 7|d^{\perp}\rangle \in \mathcal{N}(\{d,d^{\perp}\})$. We can transform it along the machine channel mc to evidence $mc^{**}(\psi) \in \mathcal{N}(Pred(\{A,B\}))$ of the form:

$$mc^{**}(\psi) \;=\; 5\Big|mc^{*}(\mathbf{1}_d)\Big\rangle + 7\Big|mc^{*}(\mathbf{1}_{d^{\perp}})\Big\rangle \;=\; 5\Big|\tfrac{1}{3}\cdot\mathbf{1}_A + \tfrac{1}{6}\cdot\mathbf{1}_B\Big\rangle + 7\Big|\tfrac{2}{3}\cdot\mathbf{1}_A + \tfrac{5}{6}\cdot\mathbf{1}_B\Big\rangle.$$

The Pearl update with this evidence leads to the claimed posterior since:

$$\omega\big|^{\flat}_{mc^{**}(\psi)} \;=\; \omega\Big|_{(mc^{*}(\mathbf{1}_d))^5 \,\&\, (mc^{*}(\mathbf{1}_{d^{\perp}}))^7} \;\approx\; 0.402\Big|A\Big\rangle + 0.598\Big|B\Big\rangle.$$

In contrast, Jeffrey's update gives a quite different outcome:

$$\omega\big|_{mc^{**}(\psi)} \;=\; \tfrac{5}{12}\cdot\omega\big|_{mc^{*}(\mathbf{1}_d)} + \tfrac{7}{12}\cdot\omega\big|_{mc^{*}(\mathbf{1}_{d^{\perp}})} \;=\; (mc^{\dagger}_{\omega})_{*}(Flrn(\psi)) \;\approx\; 0.113\Big|A\Big\rangle + 0.887\Big|B\Big\rangle.$$

We may conclude that Jaynes (2003) performed a Pearl update.

Here is another example of Jeffrey learning along a channel.

Example 8.4.4 In Prescher (2005) a challenge is presented to learn the biases of two dices, when there is information about the sums of their outcomes. The prior dice distributions are:

$$\begin{aligned}\omega_1 &= 0.18|1\rangle + 0.19|2\rangle + 0.16|3\rangle + 0.13|4\rangle + 0.17|5\rangle + 0.17|6\rangle\\ \omega_2 &= 0.22|1\rangle + 0.23|2\rangle + 0.13|3\rangle + 0.16|4\rangle + 0.14|5\rangle + 0.12|6\rangle.\end{aligned}$$

There is a deterministic dice sum channel $ds\colon P \times P \rightsquigarrow S$ where $P = \{1, 2, \ldots, 6\}$ is the set of pips and $S = \{2, 3, \ldots, 12\}$ the set of their sums. This channel is given by $ds(x, y) = 1|x+y\rangle$. The point evidence $\psi \in \mathcal{N}(S)$ is of the following form, see (Prescher, 2005, Ex. 5).

$$\begin{aligned}\psi = {} & 3790|2\rangle + 7508|3\rangle + 10217|4\rangle + 10446|5\rangle + 12003|6\rangle + 17732|7\rangle\\ & + 13923|8\rangle + 8595|9\rangle + 6237|10\rangle + 5876|11\rangle + 3673|12\rangle.\end{aligned}$$

The Jeffrey update along the dice sum channel ds yields a joint distribution $\tau \in \mathcal{D}(P \times P)$ the form:

$$\tau := (\omega_1 \otimes \omega_2)\big|_{ds^{**}(\psi)} = \sum_{y \in S} \mathit{Flrn}(\psi)(y) \cdot (\omega_1 \otimes \omega_2)\big|_{ds^{*}(\mathbf{1}_y)}.$$

The learned dice distributions appear as its marginals, written as $\omega_1' = \tau[1, 0]$ and $\omega_2' = \tau[0, 1]$. They take the following form, corresponding to the tables for p_{11} and p_{12} in (Prescher, 2005, p. 24).

$$\begin{aligned}\omega_1' &= 0.168|1\rangle + 0.182|2\rangle + 0.156|3\rangle + 0.123|4\rangle + 0.173|5\rangle + 0.198|6\rangle\\ \omega_2' &= 0.207|1\rangle + 0.223|2\rangle + 0.126|3\rangle + 0.153|4\rangle + 0.146|5\rangle + 0.145|6\rangle.\end{aligned}$$

In this case we may conclude that Prescher (2005) performed a Jeffrey update.

Definition 8.3.5 uses the names of Jeffrey and Pearl for updates that seem to differ from the earlier updates in Definition 7.9.1 that were also asociated with Jeffrey and Pearl. Proposition 8.4.1 allows us to elaborate on the similarities.

Remark 8.4.5 The coincidence of the two forms of updating on the left and right in Equation (8.11) justifies why we extend the use of the name 'Jeffrey's rule' from its original in Definition 7.9.1 (2) to the seemingly different formulation (8.9) that has (8.11) as special case.

The situation is a bit subtle, since there is no perfect match: the earlier formulation of Jeffrey's rule in Definition 7.9.1 (2) with a dagger is not always an instance of the new formulation (8.9). The earlier version involves a pushforward $(c_\omega^\dagger)_*(\tau)$ of an *arbitrary* distribution $\tau \in \mathcal{D}(Y)$. This τ does not have to be a fractional distribution $\mathit{Flrn}(\varphi)$, as in (8.11). As we shall see, the proof that Jeffrey's update (8.9) gives an increase of Jeffrey validity — see Theorem 8.4.6 (1) below — relies on Proposition 8.2.5 (2), which involves natural numbers k_i, and not arbitrary (non-negative) real numbers. Therefore, the pushforward $(c_\omega^\dagger)_*(\mathit{Flrn}(\varphi))$ in Equation (8.11) arises as instance of Jeffrey's rule from Definition 7.9.1 (2) only for a fractional distribution $\mathit{Flrn}(\varphi)$ and not for an arbitrary distribution τ. We consider this difference small enough to use the name Jeffrey's rule for all cases.

(In order to obtain a perfect match we need a version of Proposition 8.2.5 (2) where the numbers $k_i \geq 0i$ are may be real-valued.)

In the Pearl case, the original update rule from Definition 7.9.1 (1) is always a special case

of the newest formulation (8.10): suppose we have a single factor q on Y which we like to use in a Pearl update $\omega|_{c^*(q)}$ as in Definition 7.9.1 (1). This factor q can be considered as singleton evidence $\varphi = 1|q\rangle \in \mathcal{N}(Fact(Y))$. Transforming it along a channel $c\colon X \rightsquigarrow Y$ gives $\psi = c^{**}(\varphi) = 1\big|c^*(q)\rangle$. Hence:

$$\omega\big|^{\flat}_{c^{**}(\varphi)} \overset{(2)}{=} \omega\Big|_{\&(c^{**}(\varphi))} = \omega\Big|_{\&(1|c^*(q)\rangle)} = \omega\Big|_{c^*(q)}.$$

The latter is the Pearl update from Definition 7.9.1 (1).

We continue with some properties of the Jeffrey and Pearl updates with evidence, from Definition 8.3.5. These properties are in line with the ones in Proposition 7.9.7. We begin with the Jeffrey case.

Theorem 8.4.6 *Let $\omega \in \mathcal{D}(X)$ be a distribution with evidence $\psi \in \mathcal{N}(\mathsf{Fact}(X))$.*

1 *Jeffrey updating increases the Jeffrey validity:*

$$\omega\rceil_{\psi} \models_{\mathcal{J}} \psi \;\geq\; \omega \models_{\mathcal{J}} \psi.$$

2 *In the special case that ψ is point evidence, inhabiting $\mathcal{N}(X)$, Jeffrey updating gives a reduction of KL-divergence:*

$$D_{KL}(\mathsf{Flrn}(\psi), \omega\rceil_{\psi}) \;\leq\; D_{KL}(\mathsf{Flrn}(\psi), \omega).$$

3 *More of the same evidence $\psi \in \mathcal{N}(\mathsf{Fact}(X))$ does not change Jeffrey updating: for each $n \geq 1$,*

$$\omega\rceil_{n\cdot\psi} = \omega\rceil_{\psi} \qquad \textit{where} \qquad n\cdot\psi = \psi + \cdots + \psi.$$

4 *Jeffrey updating is sensitive to the order of updating: for more evidence $\chi \in \mathcal{N}(\mathsf{Fact}(X))$, in general,*

$$\omega\rceil_{\psi}\rceil_{\chi} \;\neq\; \omega\rceil_{\chi}\rceil_{\psi}.$$

5 *Let ω be 'fractional', that is, the probablities in ω are fractions, and let $N \in \mathbb{N}$ be such that $N \cdot \omega$ is a multiset. Consider this $N \cdot \omega \in \mathcal{N}(X)$ as point evidence. Then, updating with this point evidence has no effect:*

$$\omega\rceil_{N\cdot\omega} = \omega.$$

6 *A fractional convex sum of Jeffrey updates can be turned into a single Jeffrey update: for numbers $n_i \in \mathbb{N}$ with $\sum_i n_i = n$ and for evidence multisets $\psi_i \in \mathcal{N}(\mathsf{Fact}(X))$,*

$$\tfrac{n_1}{n}\cdot\omega\rceil_{\psi_1} + \cdots + \tfrac{n_L}{n}\cdot\omega\rceil_{\psi_L} = \omega\rceil_{\psi} \qquad \textit{for}\;\; \psi = n_1\cdot\psi_1 + \cdots + n_L\cdot\psi_L.$$

Item (5) says that you learn nothing if you use the information from the distribution ω as evidence. More informally, you learn nothing in Jeffrey style from what you already know, as in the second equation in Proposition 7.9.7 (2). The fact that ω is required to be fractional is not essential. It is needed because we have defined multisets of evidence to have only natural numbers as multiplicities. One could drop this limitation and allow non-negative real numbers. Then one can use the distribution $\omega \in \mathcal{D}(X)$ as a multiset of point predicates itself — without multiplication with the number N — and write $\omega\rceil_{\omega} = \omega$. The same thing applies in item (6).

Proof. Let the evidence multinomial $\psi \in \mathcal{N}(Fact(X))$ have size K.

1 We write the multiset ψ as accumulation $\psi = acc(p_1, \ldots, p_K)$ of a sequence of factors, including multiple occurrences, with each $p \in supp(\psi)$ occurring $\psi(p)$ many times in this sequence $p_1, \ldots, p_K$. We form a posterior distribution ω' like in Proposition 8.2.5 (2):

$$\omega' \coloneqq \sum_{1 \leq i \leq K} \tfrac{1}{K} \cdot \omega|_{p_i} = \prod_{p \in supp(\psi)} \tfrac{\psi(p)}{K} \cdot \omega|_p = \prod_{p \in supp(\psi)} Flrn(\psi)(p) \cdot \omega|_p = \omega\rfloor_\psi.$$

Using the inequality from Proposition 8.2.5 (2) we now get:

$$\begin{aligned} \omega' \models_J \psi &\overset{(8.4)}{=} (\psi) \cdot \prod_{p \in supp(\psi)} (\omega' \models p)^{\psi(p)} = (\psi) \cdot \prod_i (\omega' \models p_i) \\ &\geq (\psi) \cdot \prod_i (\omega \models p_i) \\ &= (\psi) \cdot \prod_{p \in supp(\psi)} (\omega \models p)^{\psi(p)} \overset{(8.4)}{=} \omega \models_J \psi. \end{aligned}$$

2 First we combinine the validity increase of the previous point with the fact that Jeffrey validity of point evidence equals a multinomial draw probability, see Lemma 8.3.3 (1).

$$mn[K](\omega\rfloor_\psi)(\psi) = \omega\rfloor_\psi \models_J \psi \geq \omega \models_J \psi = mn[K](\omega)(\psi)$$

Next, Exercise 3.4.1 gives the required divergence decrease:

$$D_{KL}(Flrn(\psi), \omega\rfloor_\psi) \leq D_{KL}(Flrn(\psi), \omega).$$

3 Since $Flrn(n \cdot \psi) = Flrn(\psi)$.

4 Exercise 7.9.5 can be re-used.

5 Suppose we can write $\omega = \sum_i \frac{n_i}{N} |x_i\rangle$, with $n_i \in \mathbb{N}$ and $N = \sum_i n_i$. Then $N \cdot \omega$ is the multiset $\sum_i n_i |x_i\rangle$, which we identify with the point evidence $\sum_i n_i \big|\mathbf{1}_{x_i}\big\rangle$. Hence, by Lemma 6.1.7 (2),

$$\omega\rfloor_{N \cdot \omega} = \sum_i Flrn(N \cdot \omega)(x_i) \cdot \omega|_{\mathbf{1}_{x_i}} = \sum_i \tfrac{n_i}{N} \cdot 1|x_i\rangle = \omega. \qquad \square$$

We turn to Pearl's style of updating and its basic properties. They differ in essential ways from the properties of Jeffrey's update rule.

Theorem 8.4.7 *Let $\omega \in \mathcal{D}(X)$ be a distribution with evidence $\psi, \chi \in \mathcal{N}(Fact(X))$.*

1 Pearl updating increases the Pearl validity:

$$\omega|^{\flat}_{\psi} \models_P \psi \geq \omega \models_P \psi.$$

2 The 'product' and 'Bayes' analogues of Theorem 6.1.4 (1), (2) hold for Pearl updating:

$$\omega|^{\flat}_{\psi} \models_P \chi = \frac{\omega \models_P \psi + \chi}{\omega \models_P \psi} = \frac{(\omega|^{\flat}_{\chi} \models_P \psi) \cdot (\omega \models_P \chi)}{\omega \models_P \psi},$$

where $\psi + \chi$ is the pointwise sum of multisets.

3 Pearl updates compose and have the empty multiset neutral element, as in:

$$\omega|^{\flat}_{\psi}|^{\flat}_{\chi} = \omega|^{\flat}_{\psi+\chi} \qquad\qquad \omega|^{\flat}_{\mathbf{0}} = \omega.$$

This may be described in terms of a monoid action, see Exercise 8.4.1.

4 Pearl updating is insensitive to the order of updating:

$$\omega|^{\flat}_{\psi}|^{\flat}_{\chi} = \omega|^{\flat}_{\chi}|^{\flat}_{\psi}.$$

5 *One learns nothing via Pearl updates with uniform factors, that is, with scalars $s \cdot \mathbf{1}$ of the truth predicate, for $s > 0$. Explicitly,*

$$\omega\big|^{\flat}_{\psi} \;=\; \omega \qquad \textit{for } \psi = \sum\nolimits_i \psi(i)\big|s_i \cdot \mathbf{1}\rangle \textit{ with } s_i > 0.$$

Proof. 1 Directly by Theorem 6.1.6:

$$\omega\big|^{\flat}_{\psi} \models_{\mathrm{P}} \psi \;=\; \omega|_{\&\psi} \models \&\psi \;\geq\; \omega \models \&\psi \;=\; \omega \models_{\mathrm{P}} \psi.$$

2 We use Equation (8.19) from Exercise 8.4.1 and Theorem 6.1.4 (1):

$$\omega\big|^{\flat}_{\psi} \models_{\mathrm{P}} \chi \;=\; \omega|_{\&\psi} \models \&\chi \;=\; \frac{\omega \models (\&\psi)\,\&\,(\&\chi)}{\omega \models \&\psi} \;=\; \frac{\omega \models \&(\psi+\chi)}{\omega \models \&\psi} \;=\; \frac{\omega \models_{\mathrm{P}} \psi+\chi}{\omega \models_{\mathrm{P}} \psi}.$$

Similarly, using Theorem 6.1.4 (2) one obtains:

$$\begin{aligned}\omega\big|^{\flat}_{\psi} \models_{\mathrm{P}} \chi \;=\; \omega|_{\&\psi} \models \&\chi \;&=\; \frac{\omega \models (\&\psi)\,\&\,(\&\chi)}{\omega \models \&\psi} \\ &=\; \frac{(\omega|_{\&\chi} \models \&\psi)\cdot(\omega \models \&\chi)}{\omega \models \&\psi} \;=\; \frac{(\omega\big|^{\flat}_{\chi} \models_{\mathrm{P}} \psi)\cdot(\omega \models_{\mathrm{P}} \chi)}{\omega \models_{\mathrm{P}} \psi}.\end{aligned}$$

3 Via Lemma 6.1.7 (3):

$$\omega\big|^{\flat}_{\psi}\big|^{\flat}_{\chi} \;=\; \omega|_{\&\psi}|_{\&\chi} \;=\; \omega|_{(\&\psi)\,\&\,(\&\chi)} \overset{(8.19)}{=} \omega|_{\&(\psi+\chi)} \;=\; \omega\big|^{\flat}_{\psi+\chi}.$$

4 We use the previous point and the commutativity of addition + of multisets, in:

$$\omega\big|^{\flat}_{\psi}\big|^{\flat}_{\chi} \;=\; \omega\big|^{\flat}_{\psi+\chi} \;=\; \omega\big|^{\flat}_{\chi+\psi} \;=\; \omega\big|^{\flat}_{\chi}\big|^{\flat}_{\psi}.$$

5 Let the evidence be of the form $\psi = \sum_i \psi(i)\big|s_i \cdot \mathbf{1}\rangle$ with $s_i > 0$. Then $\&\psi = \&_i(s_i \cdot \mathbf{1})^{\psi(i)} = s \cdot \mathbf{1}$, where $s = \prod_i s_i^{\psi(i)} > 0$. Then, by Lemma 6.1.7 (4), (3), in:

$$\omega\big|_{\psi} \;=\; \omega|_{\&\psi} \;=\; \omega|_{s\cdot\mathbf{1}} \;=\; \omega|_{\mathbf{1}} \;=\; \omega. \qquad \square$$

Having seen separate applications of Jeffrey's and of Pearl's approaches in the literature we turn to an illustration where both can be used, in a suitable combination.

Example 8.4.8 Reference Do and Batzoglou (2008) explains 'expectation maximisation' (EM), mainly by providing an example. Subsection 8.6.2 will dive deeper into the EM aspects, but here we already explore the example. It involves classification of a coin in one of two classes, depending on its bias. There are five separate batches of data (multisets), each with ten coin outcomes head (H) and tail (T).

In the example there are two coins with different biases, namely $\frac{3}{5}$ and $\frac{1}{2}$. We translate this situation into a coin channel $c\colon \mathbf{2} \rightsquigarrow \{H, T\}$, where $\mathbf{2} = \{0, 1\}$, that captures the two coins with their different biases:

$$c(0) \;=\; \tfrac{3}{5}|H\rangle + \tfrac{2}{5}|T\rangle \qquad \text{and} \qquad c(1) \;=\; \tfrac{1}{2}|H\rangle + \tfrac{1}{2}|T\rangle.$$

Thus, the distribution $c(0)$ shows a slight bias towards head, whereas $c(1)$ is unbiased. By updating a prior distribution on $\mathbf{2}$ along c the resulting 'mixture' distribution on $\mathbf{2}$ tells whether the first or the second coin in c best fits the data. The idea is that if the data contains more heads than tails, then the first coin $c(0)$ is most likely: the learned distribution $r|0\rangle + (1-r)|1\rangle \in \mathcal{D}(\mathbf{2})$ will then have $r > \frac{1}{2}$.

The five batches of point evidence on $\{H, T\}$ are:

$$\psi_1 \;=\; 5|H\rangle + 5|T\rangle \qquad \psi_2 \;=\; 9|H\rangle + 1|T\rangle \qquad \psi_3 \;=\; 8|H\rangle + 2|T\rangle$$
$$\psi_4 \;=\; 4|H\rangle + 6|T\rangle \qquad \psi_5 \;=\; 7|H\rangle + 3|T\rangle.$$

Together they form a single multiset of evidence:

$$\Psi \;=\; 1\big|\psi_1\big\rangle + 1\big|\psi_2\big\rangle + 1\big|\psi_3\big\rangle + 1\big|\psi_4\big\rangle + 1\big|\psi_5\big\rangle \;\in\; \mathcal{N}\Big(\mathcal{N}[10](\{H, T\})\Big).$$

The type of this evidence suggests that we should not update along the original coin channel $c\colon \mathbf{2} \rightsquigarrow \{H, T\}$, but along $mn[10](c) := mn[10] \circ c\colon \mathbf{2} \rightsquigarrow \mathcal{N}[10](\{H, T\})$. The Jeffrey update of the prior uniform channel $\upsilon \in \mathcal{D}(\mathbf{2})$ is then of the form:

$$\begin{aligned}
\upsilon\big|_{mn[10](c)^{**}(\Psi)} &\overset{(8.9)}{=} \tfrac{1}{5}\cdot \upsilon|_{mn[10](c)^*(\mathbf{1}_{\psi_1})} + \tfrac{1}{5}\cdot \upsilon|_{mn[10](c)^*(\mathbf{1}_{\psi_2})} + \tfrac{1}{5}\cdot \upsilon|_{mn[10](c)^*(\mathbf{1}_{\psi_3})} \\
&\qquad + \tfrac{1}{5}\cdot \upsilon|_{mn[10](c)^*(\mathbf{1}_{\psi_4})} + \tfrac{1}{5}\cdot \upsilon|_{mn[10](c)^*(\mathbf{1}_{\psi_5})} \\
&\overset{(8.12)}{=} \tfrac{1}{5}\cdot \upsilon\big|^{\flat}_{c^{**}(\psi_1)} + \tfrac{1}{5}\cdot \upsilon\big|^{\flat}_{c^{**}(\psi_2)} + \tfrac{1}{5}\cdot \upsilon\big|^{\flat}_{c^{**}(\psi_3)} + \tfrac{1}{5}\cdot \upsilon\big|^{\flat}_{c^{**}(\psi_4)} + \tfrac{1}{5}\cdot \upsilon\big|^{\flat}_{c^{**}(\psi_5)} \\
&= \tfrac{1}{5}\cdot \upsilon|_{(c^*(\mathbf{1}_H))^5 \,\&\, (c^*(\mathbf{1}_T))^5} + \tfrac{1}{5}\cdot \upsilon|_{(c^*(\mathbf{1}_H))^9 \,\&\, (c^*(\mathbf{1}_T))^1} + \tfrac{1}{5}\cdot \upsilon|_{(c^*(\mathbf{1}_H))^8 \,\&\, (c^*(\mathbf{1}_T))^2} \\
&\qquad + \tfrac{1}{5}\cdot \upsilon|_{(c^*(\mathbf{1}_H))^4 \,\&\, (c^*(\mathbf{1}_T))^6} + \tfrac{1}{5}\cdot \upsilon|_{(c^*(\mathbf{1}_H))^7 \,\&\, (c^*(\mathbf{1}_T))^3} \\
&= 0.597|0\rangle + 0.403|1\rangle.
\end{aligned}$$

Thus, this Jeffrey update along $mn[10](c)$ involves a convex sum of five Pearl updates — a bit like in Example 8.3.6 (e). The table below describes these five updates; they match the (intermediate) results in (Do and Batzoglou, 2008, Fig. 1).

point evidence ψ_i	**Pearl updates** $\upsilon\big|^{\flat}_{c^{**}(\psi_i)}$
$5\|H\rangle + 5\|T\rangle$	$0.449\|0\rangle + 0.551\|1\rangle$
$9\|H\rangle + 1\|T\rangle$	$0.805\|0\rangle + 0.195\|1\rangle$
$8\|H\rangle + 2\|T\rangle$	$0.733\|0\rangle + 0.267\|1\rangle$
$4\|H\rangle + 6\|T\rangle$	$0.352\|0\rangle + 0.648\|1\rangle$
$7\|H\rangle + 3\|T\rangle$	$0.647\|0\rangle + 0.353\|1\rangle$

(8.15)

This illustration will be investigated further in Subsection 8.6.2, where we will not only update the mixture distribution υ but also the coin channel c.

We have a few more points, as afterthoughts. We have seen that Jeffrey updating increases Jeffrey validity and similarly for Pearl. The first point below demonstrates that these increases may fail if we mix up the mechanisms of Jeffrey and Pearl.

Remark 8.4.9 1 In Theorem 8.4.6 (1) and Theorem 8.4.7 (1) we have seen that the approaches of Jeffrey and Pearl each lead to an increase of validity, for the right combination of Jeffrey updating with Jeffrey validity and similarly for Pearl. Mixing validity and update of different kinds may lead to a decrease of validity. This will be illustrated for the prior distribution desease distribution $\omega = \frac{1}{20}|d\rangle + \frac{19}{20}|d^{\perp}\rangle$ from Example 8.1.1 with the positive and negative test predicates pt and nt defined there. Using the three-test evidence $\psi = 2|pt\rangle + 1|nt\rangle$ one has validity decreases after (the wrong kind of) updates:

$$\omega\big|^{\flat}_{\psi} \models_{\mathrm{J}} \psi \;=\; 0.3081 \;<\; 0.3116 \;=\; \omega \models_{\mathrm{J}} \psi$$
$$\omega\big|_{\flat\psi} \models_{\mathrm{P}} \psi \;=\; 0.2847 \;<\; 0.2858 \;=\; \omega \models_{\mathrm{P}} \psi.$$

Thus, if we wish to get wiser from learning we need to use the right combination of validity and updating. Hence, awareness of the different approaches of Jeffrey and Pearl is relevant in practical situations.

2 In Theorem 7.9.3 (2) we have seen a decrease of KL-divergence via Jeffrey updating with a distribution τ, of the form $\omega \mapsto (c_\omega^\dagger)_*(\tau)$. When this τ happens to be fractional, say $\tau = \mathit{Flrn}(\psi)$, this decrease of KL-divergence takes the following form:

$$D_{KL}\Big(\mathit{Flrn}(\varphi),\ c_*(\omega')\Big) \ \leq\ D_{KL}\Big(\mathit{Flrn}(\varphi),\ c_*(\omega)\Big), \tag{$*$}$$

where $\omega' := (c_\omega^\dagger)_*(\mathit{Flrn}(\varphi))$. We claim that this inequality $(*)$ can also be derived from the Jeffrey validity increase of Theorem 8.4.6 (1), using that ω' equals the Jeffrey update $\omega\big|_{c^{**}(\varphi)}$ by (8.11).

This works as follows. First we note that:

$$\begin{aligned}
mn[K](c_*(\omega'))(\varphi) &= c_*(\omega') \models_{\mathrm{J}} \varphi && \text{see Lemma 8.3.3 (1)}\\
&= \omega' \models_{\mathrm{J}} c^{**}(\varphi) && \text{by Exercise 8.3.3}\\
&\geq \omega \models_{\mathrm{J}} c^{**}(\varphi) && \text{by Theorem 8.4.6 (1) since } \omega' = \omega\big|_{c^{**}(\varphi)}\\
&= c_*(\omega) \models_{\mathrm{J}} \varphi && \text{by Exercise 8.3.3 again}\\
&= mn[K](c_*(\omega))(\varphi).
\end{aligned}$$

The inequality $(*)$ now follows directly from Exercise 3.4.1.

3 Pearl updates compose and commute, see Theorem 8.4.7 (3), (4), but the situation is more delicate for Jeffrey updates. In Remark 7.9.8 we have seen a combination of multiple updates via tensors. We show how this combination (also) works in the current setting, for point evidence. So let $\sigma \in \mathcal{D}(X)$ be a distribution with a channel $c\colon X \rightsquigarrow Y$ and two multisets of point evidence $\varphi, \psi \in \mathcal{N}(Y)$. We form evidence $\chi \in \mathcal{N}(\mathit{Pred}(X))$ of the following form:

$$\chi = \sum_{y,z \in Y} \varphi(y) \cdot \psi(z) \Big| c^*(\mathbf{1}_y) \mathbin{\&} c^*(\mathbf{1}_z)\Big\rangle.$$

Then it is not hard to show that:

$$\sigma\big|_\chi = \Big(\langle c, c\rangle_\sigma^\dagger\Big)_*(\mathit{Flrn}(\varphi) \otimes \mathit{Flrn}(\psi)).$$

4 A next topic is: what happens if one tries to update an arbitrary distribution $\omega \in \mathcal{D}(X)$ with inconsistent evidence? We take this inconsistent evidence to be of the form $\psi = 1\big|\mathbf{1}_U\rangle + 1\big|\mathbf{1}_{\neg U}\rangle$, with two sharp predicates $\mathbf{1}_U$ and $(\mathbf{1}_U)^\perp = \mathbf{1}_{\neg U}$, for a non-empty subset $U \subseteq \mathit{supp}(\omega)$ with non-empty complement $\neg U = \mathit{supp}(\omega) \setminus U$. The Jeffrey update with such inconsistent evidence yields a convex combination of the two updates:

$$\omega\big|_\psi = \tfrac{1}{2} \cdot \omega|_{\mathbf{1}_U} + \tfrac{1}{2} \cdot \omega|_{\mathbf{1}_{\neg U}}. \tag{8.16}$$

This can be interpreted as a 'superposition' of the two independently updated distributions. Thus, the two distributions $\omega|_{\mathbf{1}_U}$ and $\omega|_{\mathbf{1}_{\neg U}}$ have disjoint supports, but happily live together, independently, in the Jeffrey update (8.16). This reinforces the idea that Jeffrey's approach involves a (convex) combination of independent updates.

The Pearl update would be of the form $\omega\big|^{\flat}_\psi = \omega|_{\&\psi} = \omega|_{\mathbf{1}_U \,\&\, \mathbf{1}_{\neg U}}$. We notice that the conjunction $\&\psi = \mathbf{1}_U \mathbin{\&} \mathbf{1}_{\neg U} = \mathbf{1}_{U \cap \neg U} = \mathbf{1}_\emptyset = \mathbf{0}$ yields falsity. Updating with the predicate $\mathbf{0}$ is impossible, since its validity is 0. Thus, the dependent character of Pearl's approach makes updating with inconsistent evidence impossible.

8.4.1 Odds of binary distributions for Pearl updating

Binary distributions, on a two-element set, are nice and simple. Updating such a distribution can be done efficiently, via what is called the 'odds' of the distribution. We show how it can be used to calculate Pearl updates in situations with many data elements.

Definition 8.4.10 A distribution ω is called binary if its support has two elements, say $supp(\omega) = \{x_0, x_1\}$. When there is an (obvious) order on this support, say $x_0 \leq x_1$ with $\omega(x_1) \neq 0$, then the odds of ω is the fraction of probabilities:

$$od(\omega) \coloneqq \frac{\omega(x_0)}{\omega(x_1)} \in \mathbb{R}_{\geq 0}.$$

In case also $\omega(x_0) \neq 0$ we can form the *log-odds* as:

$$lod(\omega) \coloneqq \ln\Big(od(\omega)\Big) = \ln\left(\frac{\omega(x_0)}{\omega(x_1)}\right).$$

We make explicit how to reconstruct a binary distribution from its odds and log-odds.

Lemma 8.4.11 *Let a binary distribution $\omega = r|x_0\rangle + (1-r)|x_1\rangle$ be given with odds a and log-odss b. Then we can find r, and thus ω, from a and b as:*

$$r = \frac{a}{1+a} \qquad\qquad r = \frac{e^b}{1+e^b}.$$

Proof. Clearly, if $a = \frac{\omega(x_0)}{\omega(x_1)} = \frac{r}{1-r}$, then $a - r \cdot a = (1-r) \cdot a = r$, so that $a = r + r \cdot a = r \cdot (1+a)$, giving $r = \frac{a}{1+a}$. If b is the log-odds of ω, then e^b is the odds, so the second equation follows from the first. ❑

The odds and log-odds are useful in the context of Pearlian updating along a channel.

Proposition 8.4.12 *Let $\omega \in \mathcal{D}(X)$ be a binary distribution, where we assume for convenience that $supp(\omega) = X = \{0, 1\}$. We assume a channel $c\colon X \rightsquigarrow Y$ with evidence $\psi \in \mathcal{N}(Fact(Y))$. Let's write ω' for the Pearl update $\omega|^{\flat}_{c^{**}(\psi)}$ of ω with the transformed evidence.*

1 The posterior odds and log-odds, of the updated distribution ω', can be described as:

$$\begin{aligned} od(\omega') &= od(\omega) \cdot \prod_{q \in supp(\psi)} \left(\frac{c^*(q)(0)}{c^*(q)(1)}\right)^{\psi(q)} \\ lod(\omega') &= lod(\omega) + \sum_{q \in supp(\psi)} \psi(q) \cdot \ln\left(\frac{c^*(q)(0)}{c^*(q)(1)}\right). \end{aligned} \tag{8.17}$$

2 In the special case when the evidence ψ consists of point evidence, so $\psi \in \mathcal{N}(Y)$, this becomes:

$$\begin{aligned} od(\omega') &= od(\omega) \cdot \prod_{y \in supp(\psi)} \left(\frac{c(0)(y)}{c(1)(y)}\right)^{\psi(y)} \\ lod(\omega') &= lod(\omega) + \sum_{y \in supp(\psi)} \psi(y) \cdot \ln\left(\frac{c(0)(y)}{c(1)(y)}\right). \end{aligned} \tag{8.18}$$

Proof. 1 The result follows because the validities used for normalisation cancel each other out:

$$\begin{aligned} od(\omega') &= \frac{\omega|_{\&\, c^{**}(\psi)}(0)}{\omega|_{\&\, c^{**}(\psi)}(1)} = \frac{\omega(0) \cdot \&\, c^{**}(\psi)(0)}{\omega \models \&\, c^{**}(\psi)} \cdot \frac{\omega \models \&\, c^{**}(\psi)}{\omega(1) \cdot \&\, c^{**}(\psi)(1)} \\ &= \frac{\omega(0)}{\omega(1)} \cdot \frac{\prod_q (c^*(q)(0))^{\psi(q)}}{\prod_q (c^*(q)(1))^{\psi(q)}} = od(\omega) \cdot \prod_{q \in supp(\psi)} \left(\frac{c^*(q)(0)}{c^*(q)(1)}\right)^{\psi(q)}. \end{aligned}$$

As a result,

$$lod(\omega') = \ln\left(od(\omega) \cdot \prod_{q \in supp(\psi)} \left(\frac{c^*(q)(0)}{c^*(q)(1)}\right)^{\psi(q)}\right) = lod(\omega) + \sum_{q \in supp(\psi)} \psi(q) \cdot \ln\left(\frac{c^*(q)(0)}{c^*(q)(1)}\right).$$

2 This follows directly from the previous point by using that for a point predicate $\mathbf{1}_y \in Pred(Y)$ one has $c^*(\mathbf{1}_y)(i) = c(i)(y)$, for $i = 0, 1$. ❑

For large evidence multisets ψ, computing a Pearl update in its original form (8.10) can become unmangeable because the products $\&_p\ p^{\psi(p)}$ may become too small. However, the logarithmic formulas in (8.17) and (8.18) may work. This will be illustrated in two earlier examples.

Example 8.4.13 In Example 8.4.3 we have used a binary prior $\omega = \frac{1}{11}|A\rangle + \frac{10}{11}|B\rangle$, say with $A \leq B$. Its odds and log-odds are:

$$od(\omega) = \frac{\omega(A)}{\omega(B)} = \frac{1/11}{10/11} = \frac{1}{10} \qquad\qquad lod(\omega) = \ln\left(\tfrac{1}{10}\right) = -\ln(10).$$

In Example 8.4.3 further had a 'machine' channel $mc\colon \{A, B\} \twoheadrightarrow \{d, d^\perp\}$ with point evidence $\psi = 5|d\rangle + 7|d^\perp\rangle$. Let's abbreviate the Pearl update as $\omega' = \omega^{\flat}|_{mc^{**}(\psi)}$. We compute its log-odds as in (8.18):

$$\begin{aligned} lod(\omega') &= lod(\omega) + 5 \cdot \ln\left(\frac{mc(A)(d)}{mc(B)(d)}\right) + 7 \cdot \ln\left(\frac{mc(A)(d^\perp)}{mc(B)(d^\perp)}\right) \\ &= -\ln(10) + 5 \cdot \ln\left(\frac{1/3}{1/6}\right) + 7 \cdot \ln\left(\frac{2/3}{5/6}\right) = -\ln(10) + \ln(2) + \ln\left(\frac{4}{5}\right) \approx -0.3989. \end{aligned}$$

The associated posterior probability is, as in Lemma 8.4.11, $\omega'(A) \approx 0.402$. This outcome coincides with what we computed in Example 8.4.3. The log-odds approach that we use in the above follow-up version of the example closely follows the method from Jaynes (2003), with a minor difference: we use the natural logarithm ln with base e instead of the one with base 10.

Example 8.4.14 In Example 8.4.2 we skipped the calculation of the Pearl update because the evidence multiset ψ is too big for the original formula (8.10). Here we show how to do this Pearl

update via log-odds:

$$\begin{aligned}
lod\Big(\rho\big|^{P}_{\langle f,w,h\rangle^{**}(\psi)}\Big) &\overset{(8.18)}{=} lod(\rho) + \sum_{y\in supp(\psi)} \psi(y)\cdot \ln\left(\frac{\langle f,w,h\rangle(0)(y)}{\langle f,w,h\rangle(1)(y)}\right) \\
&= \ln\left(\frac{6/10}{4/10}\right) + 273\cdot\ln\left(\frac{27/125}{8/125}\right) + 93\cdot\ln\left(\frac{18/125}{12/125}\right) \\
&\qquad + 104\cdot\ln\left(\frac{18/125}{12/125}\right) + 90\cdot\ln\left(\frac{12/125}{18/125}\right) + 79\cdot\ln\left(\frac{18/125}{12/125}\right) \\
&\qquad + 100\cdot\ln\left(\frac{12/125}{18/125}\right) + 94\cdot\ln\left(\frac{12/125}{18/125}\right) + 167\cdot\ln\left(\frac{8/125}{27/125}\right) \\
&= \ln\left(\tfrac{3}{2}\right) + 273\cdot\ln\left(\tfrac{27}{8}\right) + 93\cdot\ln\left(\tfrac{3}{2}\right) + 104\cdot\ln\left(\tfrac{3}{2}\right) + 90\cdot\ln\left(\tfrac{2}{3}\right) + 79\cdot\ln\left(\tfrac{3}{2}\right) \\
&\qquad + 100\cdot\ln\left(\tfrac{2}{3}\right) + 94\cdot\ln\left(\tfrac{2}{3}\right) + 167\cdot\ln\left(\tfrac{8}{27}\right) \\
&\approx 0.4 + 332.1 + 37.7 + 42.2 - 36.5 + 32.0 - 40.5 - 38.1 - 203.1 \;=\; 126.2
\end{aligned}$$

This high number yields as Pearl update: $\rho\big|^{P}_{\langle f,w,h\rangle^{**}(\psi)} = 1|0\rangle$. This certainty of bag 0 comes from the fact that the evidence ψ in (8.14) is dominated by its first factor $273|C,R,H\rangle$. It corresponds to the highest number 332.1 in the above penultimate line. This pushes the Pearl update towards bag 0. Recall from Example 8.4.2 that the Jeffrey update yields quite a different outcome.

Exercises

8.4.1 1 Show that the operation $\&\colon \mathcal{M}(Fact(X)) \to Fact(X)$ from (8.6) is in fact a free extension, as in Proposition 1.6.5. Conclude that & is a map of commutative monoids, and thus satisfies:

$$\&(\varphi+\psi) \;=\; (\&\varphi)\,\&\,(\&\psi) \qquad\qquad \&(\mathbf{0}) \;=\; \mathbf{1}. \tag{8.19}$$

2 Recall the subset $Fact_*(X) \subseteq Fact(X)$ of non-zero factors from (6.3) and show that & restricts to $\&\colon \mathcal{M}(Fact_*(X)) \to Fact_*(X)$.

3 Consider Pearl updating as a map $(-)\big|^{P}_{(-)}\colon \mathcal{D}(X)\times\mathcal{M}(Fact_*(X)) \to \mathcal{D}(X)$. Check that as such it is an monoid action, using the equations from Theorem 8.4.7 (3).

8.4.2 Consider the situation as in Proposition 8.4.1 with a distribution $\omega\in\mathcal{D}(X)$, a channel $c\colon X \rightsquigarrow Y$, and point evidence $\varphi\in\mathcal{N}[K](Y)$. Show that Jeffrey and Pearl validity of the transformed evidence can both be expressed via multinomials, in a slightly different manner (see also Jacobs and Stein (2023c)), as formulated below.

1 $\omega \models_{J} c^{**}(\varphi) \;=\; mn[K](c_*(\omega))(\varphi)$.

2 $\omega \models_{P} c^{**}(\varphi) \;=\; mn[K](c)_*(\omega)(\varphi)$, where $mn[K](c) := mn[K]\circ c$.

8.4.3 (From Jacobs and Stein (2023c)) Let a distribution $\omega\in\mathcal{D}(X)$ and a channel $c\colon X\rightsquigarrow Y$ be given, with point evidence $\varphi\in\mathcal{N}[K](Y)$.

1 First notice that combining Proposition 8.4.1 (1) with Exercise 7.1.7 yields:

$$\omega\big|_{c^{**}(\varphi)} \;=\; (c^{\dagger}_{\omega})_*(Flrn(\varphi)) \;=\; Flrn_*\Big(\mathcal{N}[K](c)^{\dagger}_{mn[K](\omega)}(\varphi)\Big) \;=\; Flrn_*\Big(mn[K](\omega)\big|_{\mathcal{N}[K](c)^*(\mathbf{1}_{\varphi})}\Big),$$

where we use $\mathcal{N}[K](c)$ extended to channels, as in (3.40).

2 Now use Theorem 3.4.4 (2) to check that the Jeffrey update can be characterised as:

$$\omega\big|_{c^{**}(\varphi)} \in \operatorname*{argmin}_{\rho\in\mathcal{D}(X)} D_{KL}\Big(mn[K](\omega)\big|_{\mathcal{N}[K](c)^*(\mathbf{1}_{\varphi})},\; mn[K](\rho)\Big).$$

8.4.4 Say that evidence $\psi \in \mathcal{N}(Fact(X))$ *matches* (or is a match) when the sum of all factors in the support of ψ is below the truth predicate $\mathbf{1}$, that is, if $\sum_{p \in supp(\psi)} p \leq \mathbf{1}$. This implies that all factors in ψ are predicates, and thus that ψ inhabits the set $\mathcal{N}(Pred(X))$. We call ψ a *perfect match* when $\sum_{p \in supp(\psi)} p = \mathbf{1}$.

1 Show that matching evidence $\psi \in \mathcal{N}(Pred(X))$ be matching evidence satisfies:

$$\omega \models_{\mathrm{J}} \psi \leq 1 \qquad \text{and} \qquad \omega \models_{\mathrm{P}} \psi \leq 1.$$

2 Now let $P \subseteq Pred(X)$ be a finite subset of predicates that add up to one: $\sum_{p \in P} p = \mathbf{1}$. Show that for each $K \geq 1$, the sums of Jeffrey and Pearl validities of all (perfectly matching) evidence over P of size K add up to one and form distributions of the form:

$$\sum_{\varphi \in \mathcal{N}[K](P)} (\omega \models_{\mathrm{J}} \varphi) \big|\varphi\rangle = \sum_{\varphi \in \mathcal{N}[K](P)} (\varphi) \cdot \prod_{p \in P} (\omega \models p)^{\varphi(p)} \big|\varphi\rangle$$
$$\sum_{\varphi \in \mathcal{N}[K](P)} (\omega \models_{\mathrm{P}} \varphi) \big|\varphi\rangle = \sum_{\varphi \in \mathcal{N}[K](P)} (\varphi) \cdot (\omega \models \mathop{\&}_{p \in P} p^{\varphi(p)}) \big|\varphi\rangle.$$

8.4.5 Let $\omega, \omega' \in \mathcal{D}(X)$ be two distributions, with an evidence multiset $\psi \in \mathcal{N}(Fact(X))$. Show that:

$$\omega \models_{\mathrm{J}} \psi \leq \omega' \models_{\mathrm{J}} \psi \iff Flrn(\psi) \models \ln\left(\frac{\omega \models -}{\omega' \models -}\right) \leq 0.$$

8.5 Ideal expectation maximisation via double daggers

Expectation maximisation (EM) is a powerful unsupervised method in machine learning that is used to uncover structure in data. It is the topic of the current and the subsequent two sections. This section describes the ideas behind expectation maximisation, via an update not only of a distribution but also of a channel, in order to reduce the divergence between the data distribution and the prediction. The distribution update is done via Jeffrey's rule. The update of the channel is done via a new double dagger construction. We have seen that double reversal of a channel c yields the original c, see Lemma 7.1.8 (4). This works when the first reversal uses a prior ω and the second reversal uses $c_*(\omega)$ as prior. In expectation maximisation the prior for the second reversal is given by the data — and is thus not of the form $c_*(\omega)$. Hence the double dagger is a new channel, different from the original one. This section concentrates on what we call an 'ideal' version of EM. It gives a perfect match between the data distribution and the prediction in one single round. This is too good to be true. In the two next sections we shall describe variations on this ideal version of EM, which are less ideal and require multiple iterations.

In general, expectation maximisation tries to recognise a mixture of distributions in data. In essence, it works as follows. Suppose we have a multiset $\psi \in \mathcal{N}(Y)$ of data points — forming point evidence, in the terminology of the previous two sections. The aim is to find a mixture (convex sum) of distributions:

$$\sigma \coloneqq r_1 \cdot \sigma_1 + \cdots + r_N \cdot \sigma_N \qquad \text{with minimal} \qquad D_{KL}(Flrn(\psi), \sigma).$$

Implicitly, we have $\sigma_i \in \mathcal{D}(Y)$ and $r_i \in [0, 1]$ with $\sum_i r_i = 1$. The aim of expectation maximisation is thus to approximate the data distribution $Flrn(\psi)$ via a weighted sum of multiple distributions. In these situations the number N is typically fixed in advance.

To make things more concrete, see Figure 8.6 on the left, for a histogram of data points. It

looks like a combination of two binomial distributions. The goal of EM is to find in such a situation both an appropriate mixture and appropriate biases of the two binomial distributions. Appropriate here means: with low divergence between the data distribution and the prediction.

This situation that we just described with the convex sum $\sum_i r_i \cdot \sigma_i$ can be massaged into channel form. Take a finite set $X := \{1, \ldots, N\}$ with the indices as elements. The mixture probabilities r_i form a distribution $\omega := \sum_i r_i |i\rangle \in \mathcal{D}(X)$ on this set X. The distributions σ_i form a channel $c \colon X \to Y$, given by $c(i) = \sigma_i$. The above mixture $\sigma = \sum_i r_i \cdot \sigma_i$ is then the same as the prediction $c_*(\omega)$. The aim of expectation maximisation can thus be reformulated as:

$$\text{find } \omega \in \mathcal{D}(X),\ c \colon X \to Y \quad \text{with minimal} \quad D_{KL}(Flrn(\psi),\ c_*(\omega)).$$

This looks very much like a minimisation goal for Jeffrey's update rule, see Theorem 7.9.3 (2) or Theorem 8.4.6 (2). Indeed, Jeffrey's rule plays a central role in expectation maximisation. But there is more. Within the setting that we just described, Jeffrey's update rule allows us to find a 'better' distribution $\omega' \in \mathcal{D}(X)$, giving a lower divergence. This is the E-step of expectation maximisation. In the M-step one also determines a better channel $c' \colon X \to Y$, so that ω' and c' together give a lower divergence. In the expectation maximisation algorithm these two E and M steps are alternated and iterated, as in E-M-E-M-E-M-$\cdots$, until some (divergence) fixed point is reached sufficiently closely.

The first general formulation of expectation maximisation occurs in Dempster et al. (1977), but it was used in more specialised forms before, see (McLachlan and Krishnan, 2007, 1.8) for historical details. We elaborate several examples below, using the above channel-based description. Quite often, what the EM algorithm precisely achieves is poorly elaborated in the literature. One may find general claims, for instance, (Russell and Norvig, 2003, §20.3), that "EM increases the log-likelihood of the data at every iteration".

As sketeched above, we emphasise the divergence reduction with respect to a given multiset of data on a set Y, involving both a mixture distribution on a set X and a channel $X \to Y$. Such divergence reduction corresponds to an increase of (Jeffrey) validity, see the first two items of Theorem 8.4.6. This valdity increase may be expressed also as an increase of log-likelihood, see Remark 8.3.4 or Theorem 3.4.2. The domain and codomain X, Y of the channel are finite and fixed in advance. The elements of X are often seen as features that are unvealed in the data by the expectation maximisation algorithm. The mixture distribution tells what the relative weigths are of these features in the data. Therefore, the set X, and also the mixture distribution on X, are sometimes called 'latent' or 'hidden'.

The distribution $\omega \in \mathcal{D}(X)$ and the channel $c \colon X \to Y$ can be combined into a joint 'graph' distribution $\tau = \langle id, c\rangle_*(\omega) \in \mathcal{D}(X \times Y)$. Conversely, given such a joint distribution on $X \times Y$, we can decompose it into a distribution on X, via marginalisation, and into a channel $X \to Y$, via disintegration (extraction), see Section 7.2. By applying what we already know about finding better distributions to such joint distributions on $X \times Y$, we can learn the better distribution on X and channel $X \to Y$ implicitly, in one go. This is what we shall look at first.

Theorem 8.5.1 *Let $\omega \in \mathcal{D}(X)$ be given with a channel $c \colon X \to Y$ and with an evidence multiset $\psi \in \mathcal{N}(Fact(Y))$. We form the joint graph distribution $\tau := \langle id, c\rangle_*(\omega) \in \mathcal{D}(X \times Y)$. We consider the Jeffrey update of the joint distribution τ with the transferred evidence $(\pi_2)^{**}(\psi) \in$*

$\mathcal{N}(Fact(X \times Y))$ *giving:*

$$\tau' \coloneqq \tau\big|_{(\pi_2)^{**}(\psi)} \overset{(8.9)}{=} \sum_{q \in supp(\psi)} Flrn(\psi)(q) \cdot \tau|_{\mathbf{1} \otimes q} \in \mathcal{D}(X \times Y). \tag{8.20}$$

We disintegrate this Jeffrey update as $\tau' = \langle id, c' \rangle_*(\omega')$. *Then:*

1 *The new distribution* $\omega' = \tau'[1,0] \in \mathcal{D}(X)$ *that we obtain in this way is the Jeffrey update* $\omega' = \omega\big|_{c^{**}(\psi)}$ *of* ω, *with evidence transferred along the channel* c.

2 *The new channel* $c' = \tau'[0,1 \,\big|\, 1,0] \colon X \rightsquigarrow Y$ *can be described on* $x \in X$ *as convex sum of a channel updates* $c|_q$, *as in Definition 6.1.1 (3),*

$$c'(x) = \sum_{q \in supp(\psi)} \frac{Flrn(\psi)(q) \cdot \omega|_{c^*(q)}(x)}{\omega\big|_{c^{**}(\psi)}(x)} \cdot c|_q(x).$$

Then:

$$c'_*(\omega') = c_*(\omega)\big|_\psi.$$

3 *The new pair* ω', c' *gives an increase of Jeffrey validity:*

$$\omega' \models_J c'^{**}(\psi) \geq \omega \models_J c^{**}(\psi).$$

Proof. 1 For an element $x \in X$,

$$\begin{aligned}
\omega'(x) = \sum_{y \in Y} \tau'(x,y) &\overset{(8.20)}{=} \sum_{y \in Y} \sum_{q \in supp(\psi)} Flrn(\psi)(q) \cdot \frac{\tau(x,y) \cdot q(y)}{\tau \models \mathbf{1} \otimes q} \\
&= \sum_{q \in supp(\psi)} Flrn(\psi)(q) \cdot \sum_{y \in Y} \frac{\omega(x) \cdot c(x)(y) \cdot q(y)}{\sum_{x,y} \omega(x) \cdot c(x)(y) \cdot q(y)} \\
&= \sum_{q \in supp(\psi)} Flrn(\psi)(q) \cdot \frac{\omega(x) \cdot c^*(q)(x)}{\omega \models c^*(q)} \\
&= \sum_{q \in supp(\psi)} Flrn(\psi)(q) \cdot \omega|_{c^*(q)}(x) \overset{(8.9)}{=} \omega\big|_{c^{**}(\psi)}(x).
\end{aligned}$$

2 For elements $x \in X$ and $y \in Y$,

$$\begin{aligned}
c'(x)(y) = \tau'[0,1 \,\big|\, 1,0](x)(y) &= \frac{\tau'(x,y)}{\tau'[1,0](x)} \\
&= \sum_{q \in supp(\psi)} Flrn(\psi)(q) \cdot \frac{\tau(x,y) \cdot q(y)}{(\tau \models \mathbf{1} \otimes q) \cdot \omega'(x)} \\
&= \sum_{q \in supp(\psi)} Flrn(\psi)(q) \cdot \frac{\omega(x) \cdot c(x)(y) \cdot q(y)}{(\omega \models c^*(q)) \cdot \omega\big|_{c^{**}(\psi)}(x)} \\
&= \sum_{q \in supp(\psi)} Flrn(\psi)(q) \cdot \frac{\omega(x) \cdot c^*(q)(x)}{(\omega \models c^*(q)) \cdot \omega\big|_{c^{**}(\psi)}(x)} \cdot \frac{c(x)(y) \cdot q(y)}{c(x) \models q} \\
&= \sum_{q \in supp(\psi)} \frac{Flrn(\psi)(q) \cdot \omega|_{c^*(q)}(x)}{\omega\big|_{c^{**}(\psi)}(x)} \cdot c(x)|_q(y).
\end{aligned}$$

Since, $\tau' = \langle id, c'\rangle_*(\omega')$, by definition of ω' and c' we get:

$$\begin{aligned} c'_*(\omega') &= (\pi_2)_*(\tau') \overset{(8.20)}{=} \sum_{q\in supp(\psi)} Flrn(\psi)(q)\cdot(\pi_2)_*(\tau|_{\mathbf{1}\otimes q}) \\ &= \sum_{q\in supp(\psi)} Flrn(\psi)(q)\cdot(\pi_2)_*\big(\langle id, c\rangle_*(\omega)|_{\mathbf{1}\otimes q}\big) \\ &= \sum_{q\in supp(\psi)} Flrn(\psi)(q)\cdot c_*(\omega)|_q \quad \text{by Exercise 6.3.2} \\ &= c_*(\omega)\big|_{\psi}. \end{aligned}$$

3 We use Exercise 8.3.3 (1) multiple times in:

$$\begin{aligned} \omega' \models_J c'^{**}(\psi) = c'_*(\omega') \models_J \psi = (\pi_2)_*(\tau') \models_J \psi &= \tau' \models_J (\pi_2)^{**}(\psi) \\ &\geq \tau \models_J (\pi_2)^{**}(\psi) \quad \text{by Theorem 8.4.6 (1)} \\ &= (\pi_2)_*(\tau) \models_J \psi \\ &= c_*(\omega) \models_J \psi = \omega \models_J c^{**}(\psi). \end{aligned}$$

❑

Specialisation of the above result to point evidence yields formulations that are useful for expectation maximisation.

Theorem 8.5.2 *Consider the setting of Theorem 8.5.1 with mixture distribution $\omega \in \mathcal{D}(X)$ and $c\colon X \rightsquigarrow Y$, but now with* point *evidence $\psi \in \mathcal{N}(Y)$. The updates ω' and c' and increases that we saw there now take the following form.*

1 The updated mixture distribution is the Jeffrey update: $\omega' = \big(c^\dagger_\omega\big)_*(Flrn(\psi))$.
2 The updated channel is a 'double dagger'

$$c' = \big(c^\dagger_\omega\big)^\dagger_{Flrn(\psi)} \qquad \textit{with a perfect match} \qquad c'_*(\omega') = Flrn(\psi).$$

The Jeffrey validity $c'_*(\omega') \models_J \psi = Flrn(\psi) \models_J \psi$ is maximal for ψ, see Lemma 8.3.3 (2). Correspondingly, there is a minimal divergence:

$$\begin{aligned} D_{KL}(Flrn(\psi), c'_*(\omega')) = 0 &\leq D_{KL}(Flrn(\psi), c_*(\omega')) \\ &\leq D_{KL}(Flrn(\psi), c_*(\omega)). \end{aligned} \tag{8.21}$$

The divergence decreases in (8.21) differ from earlier such decreases, like in Theorem 7.9.3 (2) and in Theorem 8.4.6 (2), since not only the prior distribution ω is updated, but also the channel c.

Proof. 1 By Theorem 8.5.1 (1):

$$\omega' = \omega\big|_{c^{**}(\psi)} = \sum_{y\in Y} Flrn(\psi)(y)\cdot\omega|_{c^*(\mathbf{1}_y)} = \sum_{y\in Y} Flrn(\psi)(y)\cdot c^\dagger_\omega(y) = \big(c^\dagger_\omega\big)_*(Flrn(\psi)).$$

2 We use that a channel update of the form $c|_{\mathbf{1}_y}$ is the constant channel $z \mapsto 1|y\rangle$. Hence, by Theorem 8.5.1 (2) and the previous point we get that the updated channel c' can be described, at $x \in X$ as:

$$\begin{aligned} c'(x) = \sum_{y\in Y} \frac{Flrn(\psi)(y)\cdot\omega|_{c^*(\mathbf{1}_y)}(x)}{\omega\big|_{c^{**}(\psi)}(x)}\cdot c|_{\mathbf{1}_y}(x) &= \sum_{y\in Y} \frac{Flrn(\psi)(y)\cdot c^\dagger_\omega(y)(x)}{(c^\dagger_\omega)_*(Flrn(\psi))(x)}\,|y\rangle \\ &= \sum_{y\in Y} \big(c^\dagger_\omega\big)^\dagger_{Flrn(\psi)}(x)(y)\,|y\rangle = \big(c^\dagger_\omega\big)^\dagger_{Flrn(\psi)}(x). \end{aligned}$$

Then, by writing $d := c^{\dagger}_{\omega}$ for the single dagger, we get for the double dagger $c' = d^{\dagger}_{Flrn(\psi)}$,

$$c'_{*}(\omega') = \left(d^{\dagger}_{Flrn(\psi)}\right)_{*}(d_{*}(Flrn(\psi))) \overset{(7.4)}{=} Flrn(\psi). \qquad \square$$

In the sequel we describe several version of the EM algorithm, in pseudo code. The previous theorem translates into what we call the 'ideal' EM algorithm that reaches a perfect match with prediction (divergence error zero) in one go, as in (8.21). It uses point evidence ψ in the following `idealEM` function.

```
def idealEM (dist, chan):
   dagger = chan†_dist
   # E-step, as Jeffrey update
   new_dist = dagger_*(Flrn(ψ))
   # M-step, as double dagger
   new_chan = dagger†_Flrn(ψ)
   return (new_dist, new_chan)
```
(8.22)

This algorithm takes a pair of a distribution and a channel as input and produces a new pair. Hence it it can be iterated. But it reaches a fixed point after one run, with divergence zero between the (normalised) evidence and the prediction, as in (8.21).

In the next two sections we shall see variations on this ideal version (8.22) which do not reach a fixed point after one iteration. We shall describe these variations with similar code fragments, each time with a Jeffrey update of the mixture distribution and with a new channel that is obtained via the double dagger.

8.6 Expectation maximisation without parameters

These section elaborates three examples of expectation maximisation (EM) in separate subsections. The examples are variations of the ideal version of EM described in the previous version. These variations require separate arguments to prove correctness, in terms of divergence reduction. The next section concentrates on distributions from a fixed class, involving parameters, like binomial distributions with a bias. Here we do not assume that the distributions at hand are parameterised.

8.6.1 Naive expectation maximisation

The version of EM that we consider first is called 'naive' because it involves steps that are similar to the ones that we have seen in naive Bayesian classification, in Subsection 7.5.1. The similarity (and naivety) lies in the deliberate splitting up of a joint distribution via marginalisation, in a situation where dependencies (entwinedness between the product components) are known to exist but are ignored. We elaborate this in an earlier candy bag illustration

Recall Example 8.4.2 (from Russell and Norvig (2003)), where one has three different kinds of candies in two bags, labeled 0 and 1. There is a bag distribution $\rho \in \mathcal{D}(\mathbf{2})$ and a three channels combined in a tuple $\langle f, w, h\rangle \colon \mathbf{2} \rightsquigarrow \{C, L\}\times\{R, G\}\times\{H, H^{\perp}\}$ with point evidence ψ on its codomain. Earlier we used Jeffrey updating to obtain a new bag distribution ρ'. We now see it as the E-step of expectation maximisation. At this stage we add an M-step to obtain new channels.

The data multiset $\psi \in \mathcal{N}(\{C, L\} \times \{R, G\} \times \{H, H^\perp\})$ is as in (8.14). We form the (single) dagger $d := \langle f, w, h \rangle^\dagger_\rho \colon \{C, L\} \times \{R, G\} \times \{H, H^\perp\} \rightsquigarrow \{0, 1\}$ of this tuple channel, also as before, and use it for the Jeffrey update $\rho' := d_*(\mathit{Flrn}(\psi)) \in \mathcal{D}(\{0, 1\})$, see Example 8.4.2 for the outcome. At this stage we perform an extra step and take the double dagger $e := d^\dagger_{\mathit{Flrn}(\psi)} \colon \{0, 1\} \rightsquigarrow \{C, L\} \times \{R, G\} \times \{H, H^\perp\}$.

Now there is a twist. In the ideal EM algorithm 8.22 we would now return the Jeffrey update ρ' together with the double dagger e. Here however, we like to have three separate channels as output, of the same types as the three input channels. Therefore we take the marginals via projections, giving three new channels of the right type:

$$\begin{aligned} f' &:= \pi_1 \circ e = e[1, 0, 0] \ : \ \{0, 1\} \rightsquigarrow \{C, L\} \\ w' &:= \pi_2 \circ e = e[0, 1, 0] \ : \ \{0, 1\} \rightsquigarrow \{R, G\} \\ h' &:= \pi_3 \circ e = e[0, 0, 1] \ : \ \{0, 1\} \rightsquigarrow \{H, H^\perp\}. \end{aligned}$$

This manner of obtaining multiple channels from a channel producing a joint distribution is reminiscent of naive Bayesian classification, see Remark 7.5.1. For that reason, this version is called naive EM.

The new channels f', w', h' obtained after one EM round precisely match the calculations in Russell and Norvig (2003). Explicitly:

$$\begin{aligned} f'(0) &= 0.668|C\rangle + 0.332|L\rangle & f'(1) &= 0.389|C\rangle + 0.611|L\rangle \\ w'(0) &= 0.648|R\rangle + 0.352|G\rangle & w'(1) &= 0.382|R\rangle + 0.618|G\rangle \\ h'(0) &= 0.656|H\rangle + 0.344|H^\perp\rangle & h'(1) &= 0.383|H\rangle + 0.617|H^\perp\rangle. \end{aligned}$$

Hence, what is called an instance of expectation maximisation in Russell and Norvig (2003) involves (implicitly) the marginals of the double dagger of the original channel.

This updating of both the distribution (mixture) and the channels is describe by the following pseudo code. Notice that the input has the same type as the output so that the algorithm can be iterated.

```
def candybagEM (dist, f_chan, w_chan, h_chan):
   tuple_chan = ⟨ f_chan, w_chan, h_chan ⟩
   dagger = tuple_chan†dist
   # E-step, as Jeffrey update
   new_dist = dagger*(Flrn(ψ))
   # M-step, via channel marginalisations
   double_dagger = dagger†Flrn(ψ)
   new_f_chan = proj1 ⊙ double_dagger
   new_w_chan = proj2 ⊙ double_dagger
   new_h_chan = proj3 ⊙ double_dagger
   return (new_dist, new_f_chan, new_w_chan, new_h_chan)
```
(8.23)

When we iterate this algorithm we can keep track of the (decrease of the) divergences and validities. The table in Figure 8.4 shows successive runs of the EM algorithm (8.23), starting from the (prior) ρ, f, w, h. The middle column contains the prediction errors, as divergences $D_{KL}(\mathit{Flrn}(\psi), \langle f^{(n)}, w^{(n)}, h^{(n)} \rangle_*(\rho^{(n)}))$. The column on the right contains the log-values of the Jeffrey validities $\rho^{(n)} \models_{\mathrm{JL}} \langle f^{(n)}, w^{(n)}, h^{(n)} \rangle^{**}(\psi)$ for the n-th iteration of the algorithm, as in Re-

iteration	prediction error	log-validity
0	0.0649	-2044.26
10	0.0027	-1982.02
20	0.0013	-1980.7
30	0.0007	-1980.05
40	0.0004	-1979.72
50	0.0002	-1979.55

Figure 8.4 Results of the first 50 iterations of the EM algorithm (8.23).

mark 8.3.4. We can see that they stabilise, but their stabilisation point is less informative than the almost-zero divergence in the middle column.

The first log-validity -2044.26 and the one after 10 rounds -1982.02 in Figure 8.4 are close to the values reported in Russell and Norvig (2003), namely about -2044 and -1982.214. After these runs the log-validities and prediction errors are close to a fixed point. The associated final mixture of bags, after 50 runs, is $0.458|0\rangle + 0.542|1\rangle$. It shows a markable difference with the prior $\rho = 0.6|0\rangle + 0.4|1\rangle$.

The three channels $f^{(50)}$, $w^{(50)}$ and $h^{(50)}$ obtained by running the above EM algorithm (8.23) 50 times, with the original prior mixture ρ, are of the form:

$$\begin{aligned} f^{(50)}(0) &= 0.869|C\rangle + 0.131|L\rangle & f^{(50)}(1) &= 0.299|C\rangle + 0.0.701|L\rangle \\ w^{(50)}(0) &= 0.781|R\rangle + 0.219|G\rangle & w^{(50)}(1) &= 0.346|R\rangle + 0.654|G\rangle \\ h^{(50)}(0) &= 0.817|H\rangle + 0.183|H^{\perp}\rangle & h^{(50)}(1) &= 0.325|H\rangle + 0.675|H^{\perp}\rangle. \end{aligned}$$

They differ from their originals, as described in Example 8.4.2.

We have experimentally shown that this naive EM algorithm (8.23) leads to divergence decreases, but we would like to prove this in general. Theorem 8.5.2 does not apply since the double dagger channel e used there is split up into three channel in (8.23), via marginalisation. It turns out that also in this naive case there are divergence decreases, both in the M-step and in the E-step. We shall handle the binary case, with two marginalisations, instead of three in (8.23). This is not a fundamental difference. The proof works via Proposition 7.8.5.

Theorem 8.6.1 *Consider a distribution $\omega \in \mathcal{D}(X)$ with a pair of channels $c_1 \colon X \rightsquigarrow Y_1$, $c_2 \colon X \rightsquigarrow Y_2$ and with point evidence $\psi \in \mathcal{N}(Y_1 \times Y_2)$. Define:*

$$\begin{aligned} d &:= \langle c_1, c_2\rangle^{\dagger}_{\omega} \;:\; Y_1 \times Y_2 \rightsquigarrow X & c'_1 &:= c'[1,0] \;:\; X \rightsquigarrow Y_1 \\ c' &:= d^{\dagger}_{\mathrm{Flrn}(\psi)} \;:\; X \rightsquigarrow Y_1 \times Y_2 & c'_2 &:= c'[0,1] \;:\; X \rightsquigarrow Y_2 \end{aligned} \tag{8.24}$$

and $\omega' := d_*(\mathrm{Flrn}(\psi))$. Then:

1 There is an 'E-step' decrease, involving a change of distribution from ω to ω',

$$D_{KL}\big(\mathrm{Flrn}(\psi), \langle c_1, c_2\rangle_*(\omega')\big) + D_{KL}\big(\omega', \omega\big) \;\leq\; D_{KL}\big(\mathrm{Flrn}(\psi), \langle c_1, c_2\rangle_*(\omega)\big).$$

2 There is also a 'M-step' decrease, with stronger guarantees, involving an additional change

of channels, from c_1, c_2 to c'_1, c'_2 in:

$$D_{KL}\big(Flrn(\psi),\ \langle c'_1, c'_2\rangle_*(\omega')\big) \ +\ D_{KL}\big(\omega', \omega\big) \ +\ \omega' \models D_{KL}\big(c'_1(-), c_1(-)\big) + D_{KL}\big(c'_2(-), c_2(-)\big)$$
$$\leq\ D_{KL}\big(Flrn(\psi),\ \langle c_1, c_2\rangle_*(\omega)\big).$$

This last decrease yields the correctness of the naive EM algorithm, in the form of a decrease after updates of the distribution and the channels:

$$D_{KL}\big(Flrn(\psi),\ \langle c'_1, c'_2\rangle_*(\omega')\big) \ \leq\ D_{KL}\big(Flrn(\psi),\ \langle c_1, c_2\rangle_*(\omega)\big).$$

Proof. 1 The first step is simply the Jeffrey decrease of Theorem 7.9.3 (2).

2 The double dagger channel c' defined above (8.24) satisfies $c'_*(\omega') = Flrn(\psi)$, see Theorem 8.5.2 (2). We use Proposition 2.10.6 (3) for the inequality below.

$$\begin{aligned}
D_{KL}\big(Flrn(\psi),\ \langle c'_1, c'_2\rangle_*(\omega')\big) &= D_{KL}\big((\pi_2)_*(\langle id, c'\rangle_*(\omega')),\ (\pi_2)_*(\langle id, \langle c'_1, c'_2\rangle\rangle_*(\omega'))\big) \\
&\leq D_{KL}\big(\langle id, c'\rangle_*(\omega'),\ \langle id, \langle c'_1, c'_2\rangle\rangle_*(\omega')\big) \\
&= D_{KL}\big(\langle id, c'\rangle_*(\omega'),\ \langle id, \langle c_1, c_2\rangle\rangle_*(\omega')\big) \\
&\qquad - D_{KL}\big(\langle id, \langle c'_1, c'_2\rangle\rangle_*(\omega'),\ \langle id, \langle c_1, c_2\rangle\rangle_*(\omega')\big).
\end{aligned} \tag{8.25}$$

This last equation is from Proposition 7.8.5 (2). We elaborate the two terms in this last line of (8.25) separately.

- We use the dagger definition (7.1) for c' and d in:

$$\begin{aligned}
&D_{KL}\big(\langle id, c'\rangle_*(\omega'),\ \langle id, \langle c_1, c_2\rangle\rangle_*(\omega')\big) \\
&= \sum_{x, y_1, y_2} \omega'(x) \cdot c'(x)(y_1, y_2) \cdot \ln\left(\frac{\omega'(x) \cdot c'(x)(y_1, y_2)}{\omega'(x) \cdot c_1(x)(y_1) \cdot c_2(x)(y_2)}\right) \\
&\overset{(8.24)}{=} \sum_{x, y_1, y_2} \omega'(x) \cdot c'(x)(y_1, y_2) \cdot \ln\left(\frac{Flrn(\psi)(y_1, y_2) \cdot d(y_1, y_2)(x)}{d_*(Flrn(\psi))(x) \cdot c_1(x)(y_1) \cdot c_2(x)(y_2)}\right) \\
&\overset{(8.24)}{=} \sum_{x, y_1, y_2} \omega'(x) \cdot c'(x)(y_1, y_2) \cdot \ln\left(\frac{Flrn(\psi)(y_1, y_2) \cdot \omega(x)}{\omega'(x) \cdot \langle c_1, c_2\rangle_*(\omega)(y_1, y_2)}\right) \\
&= \sum_{x, y_1, y_2} \omega'(x) \cdot c'(x)(y_1, y_2) \cdot \ln\left(\frac{Flrn(\psi)(y_1, y_2)}{\langle c_1, c_2\rangle_*(\omega)(y_1, y_2)}\right) \\
&\qquad - \sum_{x, y_1, y_2} \omega'(x) \cdot c'(x)(y_1, y_2) \cdot \ln\left(\frac{\omega'(x)}{\omega(x)}\right) \\
&= \sum_{y_1, y_2} c'_*(\omega')(y_1, y_2) \cdot \ln\left(\frac{Flrn(\psi)(y_1, y_2)}{\langle c_1, c_2\rangle_*(\omega)(y_1, y_2)}\right) - \sum_{x} \omega'(x) \cdot \ln\left(\frac{\omega'(x)}{\omega(x)}\right) \\
&= \sum_{y_1, y_2} Flrn(\psi)(y_1, y_2) \cdot \ln\left(\frac{Flrn(\psi)(y_1, y_2)}{\langle c_1, c_2\rangle_*(\omega)(y_1, y_2)}\right) - D_{KL}(\omega', \omega) \\
&= D_{KL}\big(Flrn(\psi),\ \langle c_1, c_2\rangle_*(\omega)\big) - D_{KL}(\omega', \omega).
\end{aligned}$$

- For the second term in (8.25) we get:

$$
\begin{aligned}
& D_{KL}\Big(\langle id, \langle c_1', c_2'\rangle\rangle_*(\omega'),\ \langle id, \langle c_1, c_2\rangle\rangle_*(\omega')\Big) \\
&= \sum_{x,y_1,y_2} \omega'(x)\cdot c_1'(x)(y_1)\cdot c_2'(x)(y_2)\cdot \ln\left(\frac{\omega'(x)\cdot c_1'(x)(y_1)\cdot c_2'(x)(y_2)}{\omega'(x)\cdot c_1(x)(y_1)\cdot c_2(x)(y_2)}\right) \\
&= \sum_{x} \omega'(x)\cdot\left[\sum_{y_1} c_1'(x)(y_1)\cdot \ln\left(\frac{c_1'(x)(y_1)}{c_1(x)(y_1)}\right) + \sum_{y_2} c_2'(x)(y_2)\cdot \ln\left(\frac{c_2'(x)(y_2)}{c_2(x)(y_2)}\right)\right] \\
&= \omega' \models D_{KL}\Big(c_1'(-), c_1(-)\Big) + D_{KL}\Big(c_2'(-), c_2(-)\Big).
\end{aligned}
$$

Combination of theses bullets with (8.25) yields the inequality in item (2) and completes the proof. ❑

8.6.2 Batch expectation maximisation

An often cited article Do and Batzoglou (2008) in biotechnology explains expectation maximisation via an illustration. We will reconstruct the outcomes in principled manner, via double daggers. We have already seen the set-up in Example 8.4.8. It involves a channel $c\colon \mathbf{2} \rightsquigarrow \{H, T\}$ for two coins and a uniform prior state υ on $\mathbf{2} = \{0, 1\}$ over the coin indices. There are five pieces of point evidence $\psi_1, \ldots, \psi_5 \in \mathcal{N}[10](\{H, T\})$, giving rise to five newly learned distributions on $\mathbf{2}$ via separate Pearl updates along c, see the table (8.15). The five batches of evidence can be combined in a single multiset of multisets $\Psi = \sum_i 1|\psi_i\rangle \in \mathcal{N}(\mathcal{N}[10](\{H, T\}))$.

When we apply frequentist learning to this 'big' multiset Ψ we get a distribution $\mathit{Flrn}(\Psi) \in \mathcal{D}(\mathcal{N}[10](\{H, T\}))$. This is the codomain of the multinomial distribution, with draw-size 10. Hence it makes sense to use as channel $mn[10](c) = mn[10] \circ c\colon \{0, 1\} \rightsquigarrow \mathcal{N}[10](\{H, T\})$.

The dagger now takes the form $d := mn[10](c)^{\dagger}_{\omega}$, with type $\mathcal{N}[10](\{H, T\}) \rightsquigarrow \{0, 1\}$. Hence we can form in the usual way the Jeffrey update $\omega' = d_*(\mathit{Flrn}(\Psi)) \in \mathcal{D}(\{0, 1\})$, via pushforward along the dagger. Moreover, we can form the double dagger $e := d^{\dagger}_{\mathit{Flrn}(\Psi)}\colon \{0, 1\} \rightsquigarrow \mathcal{N}[10](\{H, T\})$. In naive EM one would return this e, but here we need to adapt it, since it needs to have the same type $\{0, 1\} \rightsquigarrow \{H, T\}$ as the original channel c. Thus we take as adapted channel $c' := \mathit{Flrn} \circ e$, which has the right type.

Below we describe the EM algorithm for this approach, in pseudocode, like in (8.23). We assume that the evidence $\Psi \in \mathcal{N}^2(\{H, T\})$ is fixed.

```
def coinsEM (dist, chan):
   def multinomial_chan(x): return mn[10](chan(x))
   dagger = multinomial_chan^†_dist
   # E-step, as Jeffrey update
   new_dist = dagger_*(Flrn(ψ))
   # M-part, as channel composition of frequentist
   #    learning and double dagger
   double_dagger = dagger^†_Flrn(ψ)
   new_chan = Flrn ⊙ double_dagger
   return (new_dist, new_chan)
```
(8.26)

Neither a multinomial channel, nor frequentist learning, nor a (double) dagger is mentioned in Do and Batzoglou (2008), but nevertheless this abstract description captures what is happening there.

iteration	prediction error	log-validity
0	0.6547	-11.32
1	0.406	-10.08
2	0.383	-9.96
3	0.3644	-9.87
4	0.355	-9.82
5	0.3517	-9.81
6	0.3506	-9.8
7	0.3502	-9.8
8	0.35	-9.8
9	0.3499	-9.8
10	0.3498	-9.8

Figure 8.5 Results of the first ten iterations of the EM algorithm (8.26).

Results about running this algorithm 10 times are summarised in Figure 8.5. The same outcomes as in Do and Batzoglou (2008) are obtained, under the special circumstance that we run this EM algorithm without updating the mixture — that is, each time with the original (uniform) prior — while the channel does get updated. If we do so we get as coin channel $\mathbf{2} \multimapsto \{H, T\}$, after 10 iterations:

$$0 \longmapsto 0.797|H\rangle + 0.203|T\rangle \qquad 1 \longmapsto 0.52|H\rangle + 0.48|T\rangle.$$

These probabilities for head, 0.797 rounded to 0.80 and 0.52, can be found in (Do and Batzoglou, 2008, Fig. 1, lower-right corner). When we run the EM algorithm (8.26) 10 times while we do update the mixture distribution, we get a final mixture $0.538|0\rangle + 0.462|1\rangle \in \mathcal{D}(\mathbf{2})$ with a slightly different channel:

$$0 \longmapsto 0.79|H\rangle + 0.21|T\rangle \qquad 1 \longmapsto 0.509|H\rangle + 0.491|T\rangle.$$

For the correctness of the naive EM algorithm in Theorem 8.6.1 we used the projection to a particular shape from Proposition 7.8.5. Now, for the batch EM version we need a similar result, but for a different shape.

Definition 8.6.2 A channel $c \colon X \multimapsto \mathcal{N}[K](Y)$ is called of multinomial shape if it can be written as:

$$c \;=\; mn[K](Flrn \circledcirc d) \;=\; mn[K] \circ (Flrn \circledcirc d),$$

for some channel $d \colon X \multimapsto \mathcal{N}[K](Y)$. In that case, $Flrn \circledcirc c = Flrn \circledcirc d$, see Diagram (3.19). Every channel $mn[K](e) = mn[K] \circ e$ is of multinomial shape.

Proposition 8.6.3 *We write* *for the graph string diagram shape where the channel is of multinomial shape. Let $\omega \in \mathcal{D}(X \times \mathcal{N}[K](Y))$ be a distribution, with first marginal $\sigma = \omega[1, 0]$ and channel $c = \omega[0, 1 \,|\, 1, 0] \colon X \multimapsto \mathcal{N}[K](Y)$. Define:*

$$\omega' \;\coloneqq\; \langle id, mn[K](Flrn \circledcirc c)\rangle_{*}(\sigma) \qquad \textit{with} \qquad \omega' \models\!\approx$$

Then, for all $\rho \in \mathcal{D}(X \times \mathcal{N}[K](Y))$,

$$\rho \models\!\approx \quad \Longrightarrow \quad D_{KL}(\omega, \rho) \;=\; D_{KL}(\omega, \omega') + D_{KL}(\omega', \rho).$$

Proof. Let distribution $\rho \in \mathcal{D}(X \times \mathcal{N}[K](Y))$ have the above shape, so that it can be written as $\rho = \langle id, mn[K](Flrn \circ d)\rangle_*(\tau)$, for some $\tau \in \mathcal{D}(X)$ and $d \colon X \rightsquigarrow \mathcal{N}[K](Y)$. Then:

$$\begin{aligned}
D_{KL}(\omega, \rho) - D_{KL}(\omega, \omega') &= \sum_{x \in X,\, \varphi \in \mathcal{N}[K](Y)} \omega(x, \varphi) \cdot \ln\left(\frac{\omega'(x, \varphi)}{\rho(x, \varphi)}\right) \\
&= \sum_{x \in X,\, \varphi \in \mathcal{N}[K](Y)} \sigma(x) \cdot c(x)(\varphi) \cdot \ln\left(\frac{\sigma(x) \cdot mn[K](Flrn_*(c(x)))(y)(\varphi)}{\tau(x) \cdot mn[K](Flrn_*(d(x))(y))(\varphi)}\right) \\
&= \sum_{x \in X,\, \varphi \in \mathcal{N}[K](Y)} \sigma(x) \cdot c(x)(\varphi) \cdot \ln\left(\frac{\sigma(x)}{\tau(x)}\right) \\
&\qquad + \sum_{x \in X,\, \varphi \in \mathcal{N}[K](Y)} \sigma(x) \cdot c(x)(\varphi) \cdot \ln\left(\frac{(\varphi) \cdot \prod_y (Flrn_*(c(x)))(y)^{\varphi(y)}}{(\varphi) \cdot \prod_y (Flrn_*(d(x))(y))^{\varphi(y)}}\right) \\
&= \sum_{x \in X} \sigma(x) \cdot \ln\left(\frac{\sigma(x)}{\tau(x)}\right) \\
&\qquad + \sum_{x \in X,\, \varphi \in \mathcal{N}[K](Y),\, y \in Y} \sigma(x) \cdot c(x)(\varphi) \cdot \varphi(y) \cdot \ln\left(\frac{Flrn_*(c(x))(y)}{Flrn_*(d(x))(y)}\right) \\
&= D_{KL}(\sigma, \tau) \;+\; K \cdot \sum_{x \in X,\, y \in Y} \sigma(x) \cdot Flrn_*(c(x))(y) \cdot \ln\left(\frac{Flrn_*(c(x))(y)}{Flrn_*(d(x))(y)}\right) \\
&= D_{KL}(\sigma, \tau) \;+\; K \cdot \Big(\sigma \models D_{KL}(Flrn_*(c(-)),\, Flrn_*(d(-)))\Big).
\end{aligned}$$

We show that this same out arises as follows, via Lemma 3.3.3.

$$\begin{aligned}
D_{KL}(\omega', \rho) &= \sum_{x \in X,\, \varphi \in \mathcal{N}[K](Y)} \sigma(x) \cdot mn[K](Flrn_*(c(x)))(\varphi) \cdot \ln\left(\frac{\sigma(x) \cdot mn[K](Flrn_*(c(x)))(\varphi)}{\tau(x) \cdot mn[K](Flrn_*(d(x)))(\varphi)}\right) \\
&= \sum_{x \in X} \sigma(x) \cdot \ln\left(\frac{\sigma(x)}{\tau(x)}\right) + \sum_{x \in X,\, \varphi \in \mathcal{N}[K](Y),\, y \in Y} \\
&\qquad \cdot \sigma(x) \cdot mn[K](Flrn_*(c(x)))(\varphi) \cdot \varphi(y) \cdot \ln\left(\frac{Flrn_*(c(x))(y)}{Flrn_*(d(x))(y)}\right) \\
&= D_{KL}(\sigma, \tau) \;+\; K \cdot \sum_{x \in X,\, y \in Y} \sigma(x) \cdot Flrn_*(c(x))(y) \cdot \ln\left(\frac{Flrn_*(c(x))(y)}{Flrn_*(d(x))(y)}\right) \\
&= D_{KL}(\sigma, \tau) \;+\; K \cdot \Big(\sigma \models D_{KL}(Flrn_*(c(-)),\, Flrn_*(d(-)))\Big).
\end{aligned}$$

❑

We can now prove correctness of batch EM, as used in (8.26).

Theorem 8.6.4 *Let a distribution* $\omega \in \mathcal{D}(X)$ *and a channel* $c \colon X \rightsquigarrow Y$ *be given, with point evidence* $\psi \in \mathcal{N}(\mathcal{N}[K](Y))$ *consisting of multisets of size* K. *Take:*

$$d \coloneqq mn[K](c)^{\dagger}_{\omega} \colon \mathcal{N}[K](Y) \rightsquigarrow X \qquad e \coloneqq d^{\dagger}_{Flrn(\psi)} \colon X \rightsquigarrow \mathcal{N}[K](Y) \qquad c' \coloneqq Flrn \circ e \colon X \rightsquigarrow Y,$$

with $\omega' \coloneqq d_*(Flrn(\psi)) \in \mathcal{D}(X)$, *so that* $e_*(\omega') = Flrn(\psi)$ *by* (7.4). *Then:*

1 There is an 'E-step' decrease:

$$D_{KL}\Big(Flrn(\psi),\, mn[K](c)_*(\omega')\Big) \;+\; D_{KL}\Big(\omega', \omega\Big) \;\leq\; D_{KL}\Big(Flrn(\psi),\, mn[K](c)_*(\omega)\Big).$$

2 *There is also a 'M-step' decrease:*

$$D_{KL}\Big(Flrn(\psi),\ mn[K](c')_*(\omega')\Big) \ +\ D_{KL}\Big(\omega', \omega\Big) \ +\ K \cdot \Big(\omega' \models D_{KL}\Big(c'(-), c(-)\Big)\Big)$$
$$\leq\ D_{KL}\Big(Flrn(\psi),\ mn[K](c)_*(\omega)\Big).$$

Thus, there is also a batch EM correctness result stating that by adapting both the distribution (from ω to ω') and the channel (from c to c') one gets a decrease of prediction divergence:

$$D_{KL}\Big(Flrn(\psi),\ mn[K](c')_*(\omega')\Big) \ \leq\ D_{KL}\Big(Flrn(\psi),\ mn[K](c)_*(\omega)\Big).$$

Proof. 1 This is the Jeffrey decrease of Theorem 7.9.3 (2).

2 We proceed along the same lines as in the proof of Theorem 8.6.1 (2), but now using Proposition 8.6.3 for the last equation below.

$$\begin{aligned}
&D_{KL}\Big(Flrn(\psi),\ mn[K](c')_*(\omega')\Big)\\
&= D_{KL}\Big((\pi_2)_*(\langle d, id\rangle_*(\omega')),\ (\pi_2)_*(\langle id, mn[K](c')\rangle_*(\omega'))\Big)\\
&\leq D_{KL}\Big(\langle id, e\rangle_*(\omega'),\ \langle id, mn[K](c')\rangle_*(\omega')\Big)\\
&= D_{KL}\Big(\langle id, e\rangle_*(\omega'),\ \langle id, mn[K](c)\rangle_*(\omega')\Big)\\
&\qquad - D_{KL}\Big(\langle id, mn[K](c')\rangle_*(\omega'),\ \langle id, mn[K](c)\rangle_*(\omega')\Big).
\end{aligned} \tag{8.27}$$

These last two expressions are elaborated separately.

- First:

$$\begin{aligned}
&D_{KL}\Big(\langle id, e\rangle_*(\omega'),\ \langle id, mn[K](c)\rangle_*(\omega')\Big)\\
&= \text{`}\sum_{x,\varphi} \omega'(x) \cdot e(x)(\varphi) \cdot \ln\left(\frac{\omega'(x) \cdot e(x)(\varphi)}{\omega'(x) \cdot mn[K](c(x))(\varphi)}\right)\\
&= \sum_{x,\varphi} \omega'(x) \cdot e(x)(\varphi) \cdot \ln\left(\frac{Flrn(\psi)(\varphi) \cdot d(\varphi)(x)}{d_*(Flrn(\psi))(x) \cdot mn[K](c(x))(\varphi)}\right)\\
&= \sum_{x,\varphi} \omega'(x) \cdot e(x)(\varphi) \cdot \ln\left(\frac{Flrn(\psi)(\varphi) \cdot \omega(x)}{\omega'(x) \cdot mn[K](c)_*(\omega)(\varphi)}\right)\\
&= \sum_{x,\varphi} \omega'(x) \cdot e(x)(\varphi) \cdot \ln\left(\frac{Flrn(\psi)(\varphi)}{mn[K](c)_*(\omega)(\varphi)}\right) - \sum_{x,\varphi} \omega'(x) \cdot e(x)(\varphi) \cdot \ln\left(\frac{\omega'(x)}{\omega(x)}\right)\\
&= \sum_{\varphi} e_*(\omega')(\varphi) \cdot \ln\left(\frac{Flrn(\psi)(\varphi)}{mn[K](c)_*(\omega)(\varphi)}\right) - \sum_{x} \omega'(x) \cdot \ln\left(\frac{\omega'(x)}{\omega(x)}\right)\\
&= \sum_{\varphi} Flrn(\psi)(\varphi) \cdot \ln\left(\frac{Flrn(\psi)(\varphi)}{mn[K](c)_*(\omega)(\varphi)}\right) - D_{KL}(\omega', \omega)\\
&= D_{KL}\Big(Flrn(\psi),\ mn[K](c)_*(\omega)\Big) - D_{KL}(\omega', \omega).
\end{aligned}$$

- For the second term in (8.27) we get:

$$
\begin{aligned}
& D_{KL}\Big(\langle id, mn[K](c')\rangle_*(\omega'),\ \langle id, mn[K](c)\rangle_*(\omega')\Big) \\
&= \sum_{x,\varphi} \omega'(x) \cdot mn[K](c'(x))(\varphi) \cdot \ln\left(\frac{\omega'(x) \cdot mn[K](c'(x))(\varphi)}{\omega'(x) \cdot mn[K](c(x))(\varphi)}\right) \\
&= \sum_{x,\varphi} \omega'(x) \cdot mn[K](c'(x))(\varphi) \cdot \ln\left(\frac{(\varphi) \cdot \prod_y c'(x)(y)^{\varphi(y)}}{(\varphi) \cdot \prod_y c(x)(y)^{\varphi(y)}}\right) \\
&= \sum_{x,\varphi,y} \omega'(x) \cdot mn[K](c'(x))(\varphi) \cdot \varphi(y) \cdot \ln\left(\frac{c'(x)(y)}{c(x)(y)}\right) \\
&= K \cdot \sum_{x,y} \omega'(x) \cdot c'(x)(y) \cdot \ln\left(\frac{c'(x)(y)}{c(x)(y)}\right) \qquad \text{by Lemma 3.3.3} \\
&= K \cdot \Big(\omega' \models D_{KL}\Big(c'(-), c(-)\Big)\Big)
\end{aligned}
$$

❑

8.6.3 Topic discovery in documents

Finding topics in a documents is another challenge that can be solved via EM. There is a standard algorithm called Latent Dirichlet Allocation (LDA) Blei et al. (2003). As the name suggests, it involves the (continuous) Dirichlet distribution. Here we present a simplied version, without such Dirichlet distributions — corresponding to the uniform case in LDA. In Jacobs (2023a) the general form is described, in parameterised form — as in the next section. The correctness of LDA, in the form of decrease of divergence, is shown there. Here we give only a sketch, without correctness proof, in order to convey the idea behind LDA.

One starts from (finite) sets W of words and T of topics. Using the bag-of-words model a document is a multiset over words. We assume a finite collection $\psi_i \in \mathcal{N}(W)$ for $i \in I$, representing documents as multisets of words. We seek topic distributions $\tau_i \in \mathcal{D}(T)$ decribing the topics of document ψ_i and a channel $c\colon T \twoheadrightarrow W$ describing for each topic $t \in T$ the word distribution $c(t)$. The goal is to minimise the sum of divergences:

$$
\sum_{i \in I} D_{KL}\Big(Flrn(\psi_i),\ c_*(\tau_i)\Big). \tag{8.28}
$$

This may be described as a single divergence via parallel products, see Proposition 2.10.2 (1).

Starting from given topic distributions τ_i and a topic-to-word channel c, we do the Jeffrey updates per document. This gives $\tau'_i := (d_i)_*(Flrn(\psi_i))$, using dagger channels $d_i := c^{\dagger}_{\tau_i}\colon W \twoheadrightarrow T$. We do not (explicitly) use double daggers in this case but define the new channel $c'\colon T \twoheadrightarrow W$ via normalisation as:

$$
c'(t) := Flrn\left(\sum_{w \in W}\left(\sum_{i \in I} d_i(w)(t) \cdot \psi_i(w)\right)\big|w\rangle\right). \tag{8.29}
$$

What we just described is formulated as code fragment below, where we loosely use an index set I, as above, for the documents ψ_i and topic distributions τ_i. The latter are assumed to be in an

input list `dist_list`.

```
def topicEM (dist_list, chan):
   dags = [ chan†dist_i for i in I ]
   # E-steps, as Jeffrey updates
   new_dist_list = [ dags[i]*(Flrn(ψi)) for i in I ]
   # M-step, as convex sum of double daggers
   def new_chan(t): return flrn(lambda w:
      sum([ dags[i](w)(t) * ψi(w) for i in I ]))
   return (new_dist_list, new_chan)
```
(8.30)

The particular form of the new channel is a simplified version of the formula derived in Jacobs (2023a), obtained by omitting the Dirichlet distributions. It can be described in terms of double daggers, see Exercise 8.6.1. The derivation used there guarantees that the divergence (8.28) decreases with every run.

We elaborate a small example[1] with a set of Harry Potter words $W = \{\text{Harry, Hermione, Malfoy, magic, wand, robe, spell}\}$ and with four documents, as multisets over W, namely:

$$\begin{aligned}
\psi_0 &= 3\big|\text{Harry}\rangle + 1\big|\text{magic}\rangle + 2\big|\text{wand}\rangle \\
\psi_1 &= 3\big|\text{Hermione}\rangle + 1\big|\text{magic}\rangle + 2\big|\text{robe}\rangle \\
\psi_2 &= 3\big|\text{Malfoy}\rangle + 1\big|\text{magic}\rangle + 2\big|\text{spell}\rangle \\
\psi_3 &= 2\big|\text{Harry}\rangle + 2\big|\text{Hermione}\rangle + 2\big|\text{Malfoy}\rangle.
\end{aligned}$$

We do a first experiment with $T = \{A, B\}$ as set of topics. Notice that these topics are meaningless letters. Running the algorithm (8.30) with random topic distributions τ_i for document ψ_i and with a random topic-to-word channel 10 times leads (for instance) to the following outcomes. The topic distribution associated with the documents are:

$$\begin{array}{rr}
1\big|B\rangle \text{ for } \psi_0 & 1\big|A\rangle \text{ for } \psi_2 \\
1\big|A\rangle \text{ for } \psi_1 & 0.667\big|A\rangle + 0.333\big|B\rangle \text{ for } \psi_3.
\end{array}$$

The channel $c\colon T \twoheadrightarrow \mathcal{D}(W)$ assigns the following word distributions to the topics.

$$\begin{aligned}
c(A) &= 0.312\big|\text{Hermione}\rangle + 0.312\big|\text{Malfoy}\rangle + 0.125\big|\text{magic}\rangle + 0.125\big|\text{robe}\rangle + 0.125\big|\text{spell}\rangle \\
c(B) &= 0.625\big|\text{Harry}\rangle + 0.125\big|\text{magic}\rangle + 0.25\big|\text{wand}\rangle.
\end{aligned}$$

Clearly, topic B is about Harry with his magic wand, whereas Hermione and Malfoy are combined into topic A. When we run the algorithm multiple times, the outcome is often like this, but sometimes the same topic is assigned to documents ψ_0 and ψ_2. The final divergence (8.28) is rather stable, in the order of 1.7.

In a second experiment we run the algorithm (8.30) with a set of three topics, say $T = \{A, B, C\}$. The classification that then often emerges is, as follows — possibly with a permutation of topics.

$$\begin{array}{rr}
1\big|B\rangle \text{ for } \psi_0 & 1\big|A\rangle \text{ for } \psi_2 \\
1\big|C\rangle \text{ for } \psi_1 & 0.333\big|A\rangle + 0.333\big|B\rangle + 0.333\big|C\rangle \text{ for } \psi_3.
\end{array}$$

[1] Adapted from tidytextmining.com/topicmodeling

The channel $c\colon T \twoheadrightarrow \mathcal{D}(W)$ now assigns the following word distributions to the topics.

$$\begin{aligned} c(A) &= 0.625\big|\text{Malfoy}\rangle + 0.125\big|\text{magic}\rangle + 0.25\big|\text{spell}\rangle \\ c(B) &= 0.625\big|\text{Harry}\rangle + 0.125\big|\text{magic}\rangle + 0.25\big|\text{wand}\rangle \\ c(C) &= 0.625\big|\text{Hermione}\rangle + 0.125\big|\text{magic}\rangle + 0.25\big|\text{robe}\rangle. \end{aligned}$$

This classification makes perfect sense. The associated divergence is around 0.57, so much lower than with only two topics.

Exercises

8.6.1 Consider the situation in Subsection 8.6.3. Let's write $K_i := \|\psi_i\|$ for the number of words in document i, and for the double daggers:

$$c_i := \big(d_i\big)^{\dagger}_{Flrn(\psi_i)} \qquad \text{where} \qquad d_i := c^{\dagger}_{\tau_i}.$$

Show that the updated channel c' in Equation (8.29) can be written as convex sum of double dagger channels:

$$c'(t) = \sum_{i \in I} \frac{K_i \cdot \tau'_i(t)}{\sum_j K_j \cdot \tau'_j(t)} \cdot c_i(t) \qquad \text{where} \qquad \tau'_i := d_{i*}(Flrn(\psi_i)).$$

8.7 Expectation maximisation with parameters

This section concentrates on a quite common form of expectation maximisation (EM) involving parameterisation and updating via a change of the parameters. This is a bit like for conjugate priors, see Proposition 6.9.3 (2) or Subsection **??** later on. Such parameters occur for instance in flip, binomial or Poisson distributions. In continuous probability theory such parameters are more common. The EM updating applies to both a (mixture) distribution and a channel. The parameter is usually indicated as θ. It may be a single number, a list of numbers, or even a matrix, of some dimension. We distinguish two situations, where the channel is parametrised — written as $c[\theta]$ — and where the mixture distribution is parametrised — written as $\omega[\theta]$. In both cases one starts from an initial (often random) parameter value $\theta^{(0)}$ and new values $\theta^{(1)}, \theta^{(2)}, \ldots$ are obtained in the M-step. Interestingly, in this situation double daggers also show up, in the examples.

We shall first do the case where the channel is parametrised, since this occurs most often. The description follows Jacobs (2023a).

Theorem 8.7.1 *Consider an initial 'mixture' distribution $\omega^{(0)} \in \mathcal{D}(X)$ and point evidence $\psi \in \mathcal{N}(Y)$. We assume a family of channels $c[\theta]\colon X \twoheadrightarrow Y$, with parameter θ, having an initial value $\theta^{(0)}$.*

Consider the following two steps at stage $n \in \mathbb{N}$, to produce new distributions and channels, assuming that we already have a distribution $\omega^{(n)} \in \mathcal{D}(X)$ and channel $c^{(n)} := c[\theta^{(n)}]\colon X \twoheadrightarrow Y$.

E-step *Use Jeffrey's update rule, from Definition 7.9.1 (2), to obtain a next mixture distribution as:*

$$\omega^{(n+1)} := \big(d^{(n)}\big)_*(Flrn(\psi)) \in \mathcal{D}(X) \qquad \textit{where} \qquad d^{(n)} := c[\theta^{(n)}]^{\dagger}_{\omega^{(n)}}.$$

M-step *Pick as next parameter value the one with minimal Kullback-Leibler divergence in:*

$$\theta^{(n+1)} \in \operatorname*{argmin}_{\theta} D_{KL}\Big(Flrn(\psi),\ c[\theta]_*(\omega^{(n)})\Big). \tag{8.31}$$

Take $c^{(n+1)} := c[\theta^{(n+1)}]$ *as next channel.*

One can then observe the following.

1 Each iteration yields a decrease of Kullback-Leibler divergence:

$$D_{KL}\Big(Flrn(\psi),\ (c^{(n+1)})_*(\omega^{(n+1)})\Big) \leq D_{KL}\Big(Flrn(\psi),\ (c^{(n)})_*(\omega^{(n)})\Big). \tag{8.32}$$

2 A next parameter $\theta^{(n+1)}$ *in* (8.31) *can (often) be found as solution* θ *to the equation:*

$$\sum_{x \in X,\, y \in Y} Flrn(\psi)(y) \cdot d^{(n)}(y)(x) \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big(c[\theta](x)(y)\Big) = 0. \tag{8.33}$$

This solution need not be the minimal one in (8.31)*, but it still yields the relevant decrease of divergence in* (8.32)*.*

The subtle point in (8.33) is that it uses the channel $c^{(n)} = c[\theta^{(n)}]$, with the already fixed parameter value $\theta^{(n)}$, and not with the variable θ with respect to which we take the derivative $\frac{\mathrm{d}}{\mathrm{d}\theta}$. In this way one does not obtain a next parameter with the actual minimum, but, as will be shown below, the choice of next parameter is good enough to obtain the required decrease of divergence. Moreover, by using $c^{(n)}$ one often obtains an equation that can actually be solved. This will be illustrated in several examples below.

Proof. 1 The claimed decrease of divergence arises as follows.

$$\begin{aligned}
&D_{KL}\Big(Flrn(\psi),\ (c^{(n+1)})_*(\omega^{(n+1)})\Big) \\
&= D_{KL}\Big(Flrn(\psi),\ c[\theta^{(n+1)}]_*(\omega^{(n+1)})\Big) \\
&\leq D_{KL}\Big(Flrn(\psi),\ c[\theta^{(n)}]_*(\omega^{(n+1)})\Big) && \text{since } \theta^{(n+1)} \text{ is argmin} \\
&\leq D_{KL}\Big(Flrn(\psi),\ (c^{(n)})_*((d^{(n)})_*(Flrn(\psi)))\Big) && \text{by definition of } \omega^{(n+1)} \\
&\leq D_{KL}\Big(Flrn(\psi),\ (c^{(n)})_*(\omega^{(n)})\Big) && \text{by Theorem 7.9.3 (2).}
\end{aligned}$$

2 The minimum value (8.31) of θ in the expression $D_{KL}\Big(Flrn(\psi),\ c[\theta]_*(\omega^{(n)})\Big)$ can be found where

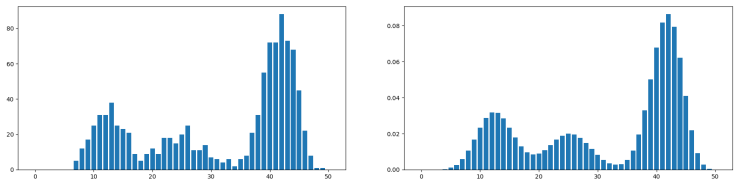


Figure 8.6 One thousand data points on the left, from which the mixture of three binomial distributions is learned, on the right, see Example 8.7.2 for details.

the derivative $\frac{\mathrm{d}}{\mathrm{d}\theta}$ of this expression is zero. This gives:

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta} D_{KL}\Big(Flrn(\psi),\, c[\theta]_*(\omega^{(n)})\Big) \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{y\in Y} Flrn(\psi)(y) \cdot \ln\left(\frac{Flrn(\psi)(y)}{c[\theta]_*(\omega^{(n)})(y)}\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{y\in Y} Flrn(\psi)(y) \cdot \ln\Big(Flrn(\psi)(y)\Big) - \sum_{y\in Y} Flrn(\psi)(y) \cdot \ln\Big(c[\theta]_*(\omega^{(n)})(y)\Big) \\
&= -\sum_{y\in Y} Flrn(\psi)(y) \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big(c[\theta]_*(\omega^{(n)})(y)\Big) \\
&= -\sum_{y\in Y} \frac{Flrn(\psi)(y)}{c[\theta]_*(\omega^{(n)})(y)} \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} c[\theta]_*(\omega^{(n)})(y) \\
&= -\sum_{y\in Y} \frac{Flrn(\psi)(y)}{c[\theta]_*(\omega^{(n)})(y)} \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{z\in Z} c[\theta](z)(y) \cdot \omega^{(n)}(z) \\
&= -\sum_{x\in X,\, y\in Y} \frac{Flrn(\psi)(y) \cdot \omega^{(n)}(x)}{c[\theta]_*(\omega^{(n)})(y)} \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} c[\theta](x)(y) \\
&= -\sum_{x\in X,\, y\in Y} \frac{Flrn(\psi)(y) \cdot \omega^{(n)}(x) \cdot c[\theta](x)(y)}{c[\theta]_*(\omega^{(n)})(y)} \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big(c[\theta](x)(y)\Big) \\
&= -\sum_{x\in X,\, y\in Y} Flrn(\psi)(y) \cdot c[\theta]^{\dagger}_{\omega^{(n)}}(y)(x) \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big(c[\theta](x)(y)\Big). \qquad (*)
\end{aligned}
$$

At this stage we need two more observations to see why it suffices to solve the equation (8.33).

(a) The minus sign $-$ can be dropped from the above last line $(*)$ when we seek a solution via setting it to zero; then we are not looking for a minimum, but for a maximum.

(b) The first θ in the dagger expression $c[\theta]^{\dagger}_{\omega^{(n)}}$ in the last line $(*)$ can be replaced by $\theta^{(n)}$, which turns $c[\theta]^{\dagger}_{\omega^{(n)}}$ in to $c[\theta^{(n)}]^{\dagger}_{\omega^{(n)}} = (c_n)^{\dagger}_{\omega^{(n)}}$, as in (8.33). This is a subtle point. As we can see in the previous item we only need that the solution $\theta^{(n+1)}$ yields a divergence that is less than the divergence for $\theta^{(n)}$. Hence if one of the θ's in $(*)$ equals $\theta^{(n)}$, we do not get the real minimum divergence for $\theta^{(n+1)}$, but we still get a divergence that is less than the one for $\theta^{(n)}$.

❑

Example 8.7.2 In this illustration of EM learning we will see how well it reconstructs a situation that we already know. Our starting point is a mixture of three binomial distributions, of the

following form, for $K = 50$.

$$\tau \coloneqq \tfrac{1}{4} \cdot bn[K](\tfrac{1}{4}) \,+\, \tfrac{1}{6} \cdot bn[K](\tfrac{1}{2}) \,+\, \tfrac{7}{12} \cdot bn[K](\tfrac{5}{6}) \,\in\, \mathcal{D}(\{0, \ldots, K\}).$$

This distribution can be obtained as state transformation:

$$\tau \,=\, c_*\Big(\tfrac{1}{4}|0\rangle + \tfrac{1}{6}|1\rangle + \tfrac{7}{12}|2\rangle\Big) \qquad \text{where} \qquad \begin{array}{rcl} \{0,1,2\} & \xrightarrow{c} & \{0,1,\ldots,K\} \\ 0 & \longmapsto & bn[K](\frac{1}{4}) \\ 1 & \longmapsto & bn[K](\frac{1}{2}) \\ 2 & \longmapsto & bn[K](\frac{5}{6}) \end{array} \tag{8.34}$$

We draw $N = 1000$ samples from the distribution τ, giving a data multiset, which we write as $\psi \in \mathcal{N}[1000](\{0, \ldots, K\})$. For this illustration it is:

$$\psi \,=\, 5|7\rangle + 12|8\rangle + 17|9\rangle + \cdots + 8|47\rangle + 1|48\rangle + 1|49\rangle.$$

The diagram on the left in Figure 8.6 is the histogram of ψ, clearly with three separate groups of data points. We set ourselves as goal to reconstruct the mixture probabilities $\frac{1}{4}$, $\frac{1}{6}$ and $\frac{7}{12}$ and also the binomial biases $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{5}{6}$ in (8.34) via EM learning.

We start from an arbitrary distribution $\omega^{(0)} \in \mathcal{D}(\mathbf{3})$, where $\mathbf{3} = \{0, 1, 2\}$. In this case the parameter θ is a triple of biases, $\theta = (\theta_0, \theta_1, \theta_2) \in [0, 1]^3$. It determines the biases of the three binomials in the channel $c[\theta] \colon \mathbf{3} \rightsquigarrow \{0, \ldots, K\}$, where:

$$c[\theta](0) \,=\, bn[K](\theta_0) \qquad c[\theta](1) \,=\, bn[K](\theta_1) \qquad c[\theta](2) \,=\, bn[K](\theta_2).$$

We now elaborate Equation (8.33), with partial derivatives $\frac{\partial}{\partial \theta_i}$, for $i \in \mathbf{3}$. As before we abbreviate $c^{(n)} = c[\theta^{(n)}]$ with associated n-stage dagger $d^{(n)} \coloneqq (c^{(n)})^{\dagger}_{\omega^{(n)}}$.

$$\begin{aligned} 0 \,&=\, \sum_{x \in \mathbf{3}} \sum_{y \in \{0,\ldots,K\}} Flrn(\psi)(y) \cdot (c^{(n)})^{\dagger}_{\omega^{(n)}}(y)(x) \cdot \frac{\partial}{\partial \theta_i} \ln\Big(c[\theta](x)(y)\Big) \\ &=\, \sum_{x \in \mathbf{3}} \sum_{y \in \{0,\ldots,K\}} Flrn(\psi)(y) \cdot d^{(n)}(y)(x) \cdot \frac{\partial}{\partial \theta_i} \ln\left(\binom{K}{y} \cdot \theta_x^y \cdot (1 - \theta_x)^{K-y}\right) \\ &=\, \sum_{y \in \{0,\ldots,K\}} Flrn(\psi)(y) \cdot d^{(n)}(y)(i) \cdot \left(\frac{y}{\theta_i} - \frac{K-y}{1-\theta_i}\right) \,=\, \frac{A_i}{\theta_i} - \frac{K \cdot B_i - A_i}{1 - \theta_i}, \end{aligned}$$

where we abbreviate:

$$\begin{aligned} A_i \,&=\, \sum_{y \in \{0,\ldots,K\}} Flrn(\psi)(y) \cdot d^{(n)}(y)(i) \cdot y \\ B_i \,&=\, \sum_{y \in \{0,\ldots,K\}} Flrn(\psi)(y) \cdot d^{(n)}(y)(i) \,=\, \Big(d^{(n)}\Big)_*(Flrn(\psi))(i) \,=\, \omega^{(n+1)}(i). \end{aligned}$$

Solving the above equation $0 = \cdots$ yields new triple of parameters $\theta^{(n+1)} = (\theta_0^{(n+1)}, \theta_1^{(n+1)}, \theta_2^{(n+1)}) \in [0, 1]^3$, where:

$$\begin{aligned} \theta_i^{(n+1)} \,=\, \frac{A_i}{K \cdot B_i} \,&=\, \sum_{y \in \{0,\ldots,K\}} \frac{Flrn(\psi)(y) \cdot d^{(n)}(y)(i) \cdot y}{K \cdot (d^{(n)})_*(Flrn(\psi))(i)} \\ &\overset{(7.1)}{=} \sum_{y \in \{0,\ldots,K\}} \frac{(d^{(n)})^{\dagger}_{Flrn(\psi)}(i)(y) \cdot y}{K} \,=\, \tfrac{1}{K} \cdot Mean\Big((d^{(n)})^{\dagger}_{Flrn(\psi)}(i)\Big). \end{aligned} \tag{8.35}$$

Interestingly, we see the double dagger appear in expectation maximisation, like in the previous

section. This allows us to use a similar EM algorithm for this mixture of binomials, starting from the point evidence $\psi \in \mathcal{N}(\{0, \ldots, K\})$

```
def BinomialMixEM (dist, chan):
   dagger = chan†dist
   # E-step, as Jeffrey update
   new_dist = dagger*(Flrn(ψ))
   # M-part, via means of double dagger
   double_dagger = dagger†Flrn(ψ)
   def new_chan(x) = bn[K]( mean(double_dagger(x)) / K )
   return (new_dist, new_chan)
```
(8.36)

The algorithm assumes that the input channel `chan` is given by a triple of binomials $\mathbf{3} \rightsquigarrow \{0, 1 \ldots, K\}$. The output channel `new_chan` is constructed to be also of this form. We implicitly use the equation $Mean(bn[K](r)) = K \cdot r$ from Example 4.2.2 (2). The number K is fixed in this situation. So if we compute the mean and divide by K, as in the one-but-last line of (8.36), we get a bias r that makes sense for the next channel.

We iteratively run this EM algorithm (8.36) with a uniform mixture $\omega^{(0)} = \frac{1}{3}|0\rangle + \frac{1}{3}|1\rangle + \frac{1}{3}|2\rangle \in \mathcal{D}(\mathbf{3})$ as initial distribution and with a triple of binomial distribution parameters $\theta^{(0)} = (0.51, 0, 50, 0.49)$, giving $c^{(0)} \colon \mathbf{3} \rightsquigarrow \{0, 1, \ldots, K\}$ with $c^{(0)}(0) = bn[K](0.51)$, $c^{(0)}(1) = bn[K](0.50)$ and $c^{(0)}(2) = bn[K](0.49)$. They give an initial divergence $D_{KL}(Flrn(\psi), (c^{(0)})_*(\omega^{(0)}))$ that is approximately 10.739.

The following table contains the outcomes that emerge in successive steps, where '**div**' stands for the divergence $D_{KL}(Flrn(\psi), (c^{(n)})_*(\omega^{(n)}))$ between the (normalised) data and the prediction, at step n. We see that these values stabilise rather quickly, namely after five iterations (using two decimals).

step n	$\omega^{(n)}$	$\theta_0^{(n)}$	$\theta_1^{(n)}$	$\theta_2^{(n)}$	**div**
0	$0.333\|0\rangle + 0.333\|1\rangle + 0.333\|2\rangle$	0.51	0.5	0.49	10.739
2	$0.564\|0\rangle + 0.0682\|1\rangle + 0.368\|2\rangle$	0.831	0.614	0.326	0.301
4	$0.569\|0\rangle + 0.137\|1\rangle + 0.294\|2\rangle$	0.833	0.539	0.28	0.067
6	$0.573\|0\rangle + 0.17\|1\rangle + 0.257\|2\rangle$	0.832	0.512	0.259	0.027
8	$0.574\|0\rangle + 0.177\|1\rangle + 0.249\|2\rangle$	0.832	0.506	0.254	0.025
10	$0.574\|0\rangle + 0.178\|1\rangle + 0.248\|2\rangle$	0.831	0.505	0.254	0.025

The learned mixture distribution $\omega \in \mathcal{D}(\mathbf{3})$ after 10 steps is $\omega = 0.574|0\rangle + 0.178|1\rangle + 0.248|2\rangle \approx \frac{7}{12}|0\rangle + \frac{1}{6}|1\rangle + \frac{1}{4}|2\rangle$. When we look at the originals (8.34) from which we sampled, the probability $\frac{7}{12}$ is associated with the point $2 \in \mathbf{3}$. In the learned distribution it belongs to $0 \in \mathbf{3}$. Apparently, the EM algorithm (8.36) does not take these orders into account: it is invariant under substitution. Correspondingly, the learned biases are 0.831, 0.505 and 0.254 approximate the biases $\frac{5}{6}$, $\frac{1}{2}$ and $\frac{1}{4}$ in the orginal set-up (8.34). They are also swapped, correspondingly.

We started from three binomial parameters 0.51, 0.50 and 0.49, differing slightly from the middle value $0.5 \in [0, 1]$. This difference is deliberate, since if these parameter are equal, all subsequent learned values will be equal too. For different initial parameters, the same values appear in the learning process, but they may be swapped.

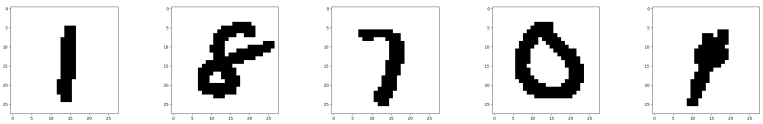


Figure 8.7 First five pictures from the MNIST library of handwritten digits that is frequently used for training purposes in image processing.

The learned values are pretty good, but not perfect. The imperfectness comes mainly from the relatively poor quality data that we have, since we took only $N = 1000$ samples. If we take $N = 10000$ samples instead, learned values will be closer to the originals.

In the above example we see a good match between the data and the statistical model, namely the mixture of binomials. The next example illustrates that if one does not have the right model, the learning outcome may be poor.

Example 8.7.3 The MNIST dataset of handwritten digits is commonly used for learning purposes. It contains thousands of pictures of such digits, each represented as a binary 28×28 matrix, see Figure 8.7. Such a matrix corresponds to an element of the product $\mathbf{2}^{784}$, where $\mathbf{2} = \{0, 1\}$ and $784 = 28^2$. In (Murphy, 2012, 1.2.1.3 and 11.2.3) this dataset is used as EM learning illustration, via parallel Bernoulli (or flip) distributions. Here we elaborate this example in channel-based form, making the role of Jeffrey's update rule and of parameter optimisation explicit — corresponding to the E- and M-steps in Theorem 8.7.1. It turns out that also in this situation we encounter a double dagger.

The channel that we use has the form $c[\theta]\colon \mathbf{10} \rightsquigarrow \mathbf{2}^{784}$, where θ is a 10×784 matrix with entries $\theta_{i,j} \in [0, 1]$, for $0 \leq i < 10$ and $0 \leq j < 784$. The channel is defined on a digit $i \in \mathbf{10} = \{0, 1, \ldots, 9\}$ via many parallel flips:

$$c[\theta](i) \coloneqq \mathit{flip}(\theta_{i,0}) \otimes \cdots \otimes \mathit{flip}(\theta_{i,783}) \in \mathcal{D}\big(\mathbf{2}^{784}\big).$$

This representation is not convenient from a computational perspective, since the state space $\mathbf{2}^{784}$ is huge. We use it for a mathematical analysis.

For an arbitrary mixture distribution $\omega \in \mathcal{D}(\mathbf{10})$ we have for a picture $y = (y_j) \in \mathbf{2}^{784}$ and a digit $i \in \mathbf{10}$, a dagger channel probability:

$$c[\theta]^{\dagger}_{\omega}(y)(i) = \frac{\omega(i) \cdot c[\theta](i)(y)}{c[\theta]_*(\omega)(y)} = \prod_{0 \leq j < 784} \frac{\omega(i) \cdot (\theta_{i,j})^{y_j} \cdot (1 - \theta_{i,j})^{1-y_j}}{c[\theta]_*(\omega)(y)}.$$

We use the first $M = 1000$ handwritten digits from the MNIST database as data $\psi = \sum_{1 \leq m \leq M} 1|y_m\rangle$, where $y_m = (y_{m,j}) \in \mathbf{2}^{784}$. By iteratively applying (parametrised) EM we get at stage n the channel $c^{(n)} = c[\theta^{(n)}]\colon \mathbf{10} \rightsquigarrow \mathbf{2}^{784}$ with dagger $d^{(n)} = (c^{(n)})^{\dagger}_{\omega^{(n)}}\colon \mathbf{2}^{784} \rightsquigarrow \mathbf{10}$. The next distribution $\omega^{(n+1)} = (d^{(n)})_*(\mathit{Flrn}(\psi)) \in \mathcal{D}(\mathbf{10})$ is obtained via Jeffrey's rule.

We solve Equation (8.33) with the partial derivatives $\frac{\partial}{\partial \theta_{i,j}}$.

$$\begin{aligned}
0 &= \sum_{0 \leq x < 10} \sum_{y \in \mathbf{2}^{784}} \mathit{Flrn}(\psi)(y) \cdot (c^{(n)})^{\dagger}_{\omega^{(n)}}(y)(x) \cdot \frac{\partial}{\partial \theta_{i,j}} \ln\Big(c[\theta](x)(y)\Big) \\
&= \sum_{0 \leq x < 10} \sum_{y \in \mathbf{2}^{784}} \mathit{Flrn}(\psi)(y) \cdot d^{(n)}(y)(x) \cdot \frac{\partial}{\partial \theta_{i,j}} \ln\left(\prod_{0 \leq k < 784} \mathit{flip}(\theta_{x,k})(y_k)\right) \\
&= \sum_{0 \leq x < 10} \sum_{y \in \mathbf{2}^{784}} \mathit{Flrn}(\psi)(y) \cdot d^{(n)}(y)(i) \cdot \frac{\partial}{\partial \theta_{i,j}} \sum_{0 \leq k < 784} \ln\Big((\theta_{x,k})^{y_k} \cdot (1 - \theta_{x,k})^{1-y_k}\Big) \\
&= \sum_{y \in \mathbf{2}^{784}} \mathit{Flrn}(\psi)(y) \cdot d^{(n)}(y)(i) \cdot \left(\frac{y_j}{\theta_{i,j}} - \frac{1 - y_j}{1 - \theta_{i,j}}\right).
\end{aligned}$$

Solving this equation yields, like in Example 8.7.2, a new parameter matrix $\theta^{(n+1)}$ with entries given by the double dagger:

$$\theta^{(n+1)}_{i,j} = \sum_{y \in \mathbf{2}^{784}} \frac{\mathit{Flrn}(\psi)(y) \cdot d^{(n)}(y)(i) \cdot y_j}{(d^{(n)})_*(\mathit{Flrn}(\psi))(i)} = \sum_{y \in \mathbf{2}^{784}} (d^{(n)})^{\dagger}_{\mathit{Flrn}(\psi)}(i)(y) \cdot y_j = \mathit{Mean}\Big((d^{(n)})^{\dagger}_{\mathit{Flrn}(\psi)}(i)\Big)_j.$$

We can now describe a single EM step as pseudocode:

```
def FlipsMixEM  (dist, chan):
   dagger = chan^†_dist
   # E-step, as Jeffrey update
   new_dist = dagger_*(Flrn(ψ))
   # M-part, via means of double dagger
   double_dagger = dagger^†_Flrn(ψ)
   def theta(i,j) = mean( double_dagger(i) )(j)
   def new_chan(i) = tensor( [ flip(theta(i,j)) for j in range(784) ] )
   return (new_dist, new_chan)
```
(8.37)

We can take an arbitrary initial mixture distribution and matrix θ and run ten EM iterations. This yields a mixture, say $\omega \in \mathcal{D}(\mathbf{10})$, and a 10×784 matrix, say θ. For each digit $i \in \mathbf{10}$ we get a sequence of coin biases $\theta_i \in [0,1]^{784}$, where θ_i is the i-th row in the matrix θ. These numbers θ_i can themselves be interpreted as a 28×28 matrix and can be rendered as a picture with grey scales. This is done in Figure 8.8 for a learned mixture probability $\omega(i)$ and corresponding matrix rows θ_i.

Interestingly, we recognise several digits. But "overall the results aren't great" ((Murphy, 2012, p. 342)). We cannot expect a relation between the digit i occurring as $|i\rangle$ in the learned mixture and the actual digit i, since these labels i are meaningless. But what is problematic is that we do not see all digits, and that some of them occur multiple times. Hence the learning algorithm does not recognise and distinguish all ten digits. At this stage we used these handwritten digits only to provide another illustration of the EM learning.

As mentioned in the beginning of this section, there is also a version of EM learning where the mixture distribution is parameterised, and not the channel (like in Theorem 8.7.1). This works for deterministic channels; they are not updated and remain the same throughout. That looks limited,

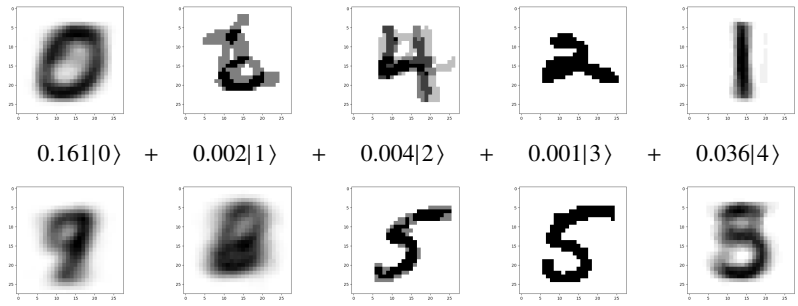


Figure 8.8 Parameter θ learned in Example 8.7.3, where each of the 10 columns in $[0,1]^{784}$ of θ is represented as a picture, with the corresponding learned mixture distribution on **10** aligned below.

but it does cover an EM example that plays an important role in the literature. We formulate a general description and then elaborate the example.

Theorem 8.7.4 *Let $\omega[\theta] \in \mathcal{D}(X)$ be a distribution with a parameter θ, for which there is an initial value $\theta^{(0)}$ with associated initial distribution $\omega^{(0)} := \omega[\theta^{(0)}]$. Further, let $f\colon X \to Y$ be a function, considered as a deterministic channel, and let $\psi \in \mathcal{N}(Y)$ be a multiset of data items. Now:*

E-step *Use Jeffrey's update rule to define:*

$$\rho_n \coloneqq \Big(f^{\dagger}_{\omega^{(n)}}\Big)_*(\mathrm{Flrn}(\psi)) \in \mathcal{D}(X). \tag{8.38}$$

M-step *Choose as next parameter value the one with minimal Kullback-Leibler divergence in:*

$$\theta^{(n+1)} \in \operatorname*{argmin}_{\theta} D_{KL}\Big(\rho_n,\, \omega[\theta]\Big), \tag{8.39}$$

and take $\omega^{(n+1)} := \omega[\theta^{(n+1)}]$ as next mixture distribution.

Now the following holds.

1 Each iteration yields a decrease of Kullback-Leibler divergence:

$$D_{KL}\Big(\mathrm{Flrn}(\psi),\, f_*(\omega^{(n+1)})\Big) \leq D_{KL}\Big(\mathrm{Flrn}(\psi),\, f_*(\omega^{(n)})\Big). \tag{8.40}$$

2 A next parameter $\theta^{(n+1)}$ can be found as solution to the equation, with ρ_n as defined in (8.38).

$$\sum_{x \in X} \rho_n(x) \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big(\omega[\theta](x)\Big) = 0. \tag{8.41}$$

Proof. 1 The decrease of divergence arises in the following way, where the first equation uses

the fact that the channel is deterministic.

$$\begin{aligned}
D_{KL}\Big(Flrn(\psi),\ f_*(\omega^{(n+1)})\Big) &= D_{KL}\Big(f_*\Big((f^{\dagger}_{\omega^{(n)}})_*(Flrn(\psi))\Big),\ f_*(\omega^{(n+1)})\Big) && \text{by (7.2)}\\
&\leq D_{KL}\Big((f^{\dagger}_{\omega^{(n)}})_*(Flrn(\psi)),\ \omega^{(n+1)}\Big) && \text{by Proposition 2.10.6 (3)}\\
&\leq D_{KL}\Big((f^{\dagger}_{\omega^{(n)}})_*(Flrn(\psi)),\ \omega^{(n)}\Big) && \text{via the argmin in (8.39)}\\
&\leq D_{KL}\Big(Flrn(\psi),\ f_*(\omega^{(n)})\Big) && \text{by Lemma 7.9.11.}
\end{aligned}$$

2 We can find the minimal divergence via the derivative:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\theta} D_{KL}\Big(Flrn(\psi),\ f_*(\omega[\theta])\Big)\\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\sum_{y\in Y} Flrn(\psi)(y)\cdot\ln\Big(Flrn(\psi)(y)\Big) - \sum_{y\in Y} Flrn(\psi)(y)\cdot\ln\Big(f_*(\omega[\theta])(y)\Big)\\
&= -\sum_{y\in Y}\frac{Flrn(\psi)(y)}{f_*(\omega[\theta])(y)}\cdot\frac{\mathrm{d}}{\mathrm{d}\theta}\sum_{x\in f^{-1}(y)}\omega[\theta](x)\\
&= -\sum_{y\in Y,\,x\in f^{-1}(y)}\frac{Flrn(\psi)(y)}{f_*(\omega[\theta])(y)}\cdot\frac{\mathrm{d}}{\mathrm{d}\theta}\omega[\theta](x)\\
&= -\sum_{y\in Y,\,x\in f^{-1}(y)}\frac{Flrn(\psi)(y)\cdot\omega[\theta](z)}{f_*(\omega[\theta])(y)}\cdot\frac{\mathrm{d}}{\mathrm{d}\theta}\ln\Big(\omega[\theta](z)\Big)\\
&= -\sum_{y\in Y,\,x\in X} Flrn(\psi)(y)\cdot f^{\dagger}_{\omega[\theta]}(y)(x)\cdot\frac{\mathrm{d}}{\mathrm{d}\theta}\ln\Big(\omega[\theta](x)\Big)\\
&= -\sum_{x\in X}(f^{\dagger}_{\omega[\theta]})_*(Flrn(\psi))(x)\cdot\frac{\mathrm{d}}{\mathrm{d}\theta}\ln\Big(\omega[\theta](x)\Big).
\end{aligned}$$

At this stage we reason as in point (b) at the end of the proof of Theorem 8.7.1: in the last line we can replace the first occurrence of $\omega[\theta]$ by $\omega^{(n)} = \omega[\theta^{(n)}]$, so that $(f^{\dagger}_{\omega^{(n)}})_*(Flrn(\psi)) = \rho_n$. This replacement means that solving equation (8.41) does not give the actual minimum divergence, but a divergence that is smaller than the one for $\theta^{(n)}$. This suffices for the reasoning in the previous point. ❑

One could argue whether the name E-step in the above theorem is appropriate, since it is used for computing the successor mixture distribution. In the previous section we have used the E-step to compute the successor channel, via a double dagger. In this case the channel is deterministic and remains the same throughout. We have already seen in Exercise 7.1.3 that taking a double dagger of a deterministic channel always yields the original channel, no matter what the priors are.

The following example comes originally from Rao (1955) and occurs also in (Rao, 1973, p.368) in a genetic context. It is used to motivate the EM algorithm in Dempster et al. (1977), and also in (McLachlan and Krishnan, 2007, 1.4.2).

Example 8.7.5 We use $\mathbf{5} = \{0, 1, 2, 3, 4\}$ and $\mathbf{4} = \{0, 1, 2, 3\}$ with function $f\colon \mathbf{5} \to \mathbf{4}$ given by:

$$f(0) = f(1) = 0 \qquad f(2) = f(1) \qquad f(3) = f(2) \qquad f(4) = f(3).$$

There is a multiset of data of size 197:

$$\psi = 125|0\rangle + 18|1\rangle + 20|2\rangle + 34|3\rangle \in \mathcal{M}(\mathbf{4}).$$

The mixture distribution $\omega[\theta] \in \mathcal{D}(\mathbf{5})$ is parameterised by $\theta \in [0,1]$ via:

$$\omega[\theta] \coloneqq \tfrac{1}{2}\big|0\rangle + \tfrac{\theta}{4}\big|1\rangle + \tfrac{1}{4}(1-\theta)\big|2\rangle + \tfrac{1}{4}(1-\theta)\big|3\rangle + \tfrac{\theta}{4}\big|4\rangle.$$

The aim is to find the value $\theta \in [0,1]$ that best fits the data ψ, along the deterministic channel f, that is, with minimal divergence $D_{KL}(\mathit{Flrn}(\psi), f_*(\omega[\theta]))$.

We first calculate the pushforward $f_*(\omega[\theta]) = \mathcal{D}(f)(\omega[\theta]) \in \mathcal{D}(\mathbf{4})$ and the dagger channel $f^\dagger_{\omega[\theta]} \colon \mathbf{4} \rightsquigarrow \mathbf{5}$ in general.

$$\begin{aligned}
f_*(\omega[\theta]) &= \sum_{j\in\mathbf{5}} \omega[\theta](j)\,\big|f(j)\rangle = (\frac{1}{2} + \frac{\theta}{4})\big|0\rangle + \frac{1}{4}(1-\theta)\big|1\rangle + \frac{1}{4}(1-\theta)\big|2\rangle + \frac{\theta}{4}\big|3\rangle \\
f^\dagger_{\omega[\theta]}(0) &= \sum_{j\in f^{-1}(0)} \frac{\omega[\theta](j)}{f_*(\omega[\theta])(0)}\,\big|j\rangle = \frac{1/2}{1/2 + \theta/4}\big|0\rangle + \frac{\theta/4}{1/2+\theta/4}\big|1\rangle = \tfrac{2}{2+\theta}\big|0\rangle + \tfrac{\theta}{2+\theta}\big|1\rangle \\
f^\dagger_{\omega[\theta]}(i) &= 1\big|i+1\rangle \qquad \text{for } i = 1,2,3.
\end{aligned}$$

Thus:

$$\begin{aligned}
\big(f^\dagger_{\omega[\theta]}\big)_*(\mathit{Flrn}(\psi)) &= \tfrac{125}{197}\Big(\tfrac{2}{2+\theta}\big|0\rangle + \tfrac{\theta}{2+\theta}\big|1\rangle\Big) + \tfrac{18}{197}\big|2\rangle + \tfrac{20}{197}\big|3\rangle + \tfrac{34}{197}\big|4\rangle \\
&= \tfrac{250}{394+197\cdot\theta}\big|0\rangle + \tfrac{125\cdot\theta}{394+197\cdot\theta}\big|1\rangle + \tfrac{18}{197}\big|2\rangle + \tfrac{20}{197}\big|3\rangle + \tfrac{34}{197}\big|4\rangle.
\end{aligned}$$

When we already have a established the n-th approximation $\theta^{(n)}$ of the parameter we set $\omega^{(n)} = \omega[\theta^{(n)}]$ and $\rho_n = (f^\dagger_{\omega^{(n)}})_*(\mathit{Flrn}(\psi))$. Equation (8.41) now amounts to:

$$0 = \sum_{j\in\mathbf{5}} \rho_n(j) \cdot \frac{\mathrm{d}}{\mathrm{d}\theta} \ln\Big(\omega[\theta](j)\Big) = \frac{\rho_n(1)+\rho_n(4)}{\theta} - \frac{\rho_n(2)+\rho_n(3)}{1-\theta}.$$

The solution is:

$$\theta = \frac{\rho_n(1)+\rho_n(4)}{\rho_n(1)+\rho_n(2)+\rho_n(3)+\rho_n(4)} \in [0,1].$$

It will be used as next iteration $\theta^{(n+1)}$. When we start from $\theta^{(0)} = \frac{1}{2}$, the successive mixture distributions $\omega^{(n)} = \omega[\theta^{(n)}]$ come close to a fixed point after four iterations, which is then of the form:

$$0.5\big|0\rangle + 0.157\big|1\rangle + 0.0933\big|2\rangle + 0.0933\big|3\rangle + 0.157\big|4\rangle.$$

The divergence $D_{KL}(\mathit{Flrn}(\psi),\ f_*(\omega^{(n)}))$ becomes 0.00153.

8.8 Learning Markov chains

The current and next section continue the study of Markov chains and hidden Markov models (HMMs) from Section 6.7. Learning for Markov chains will be done first, here, and the next section extends the approach to hidden Markov models.

We recall that a Markov chain is a pair (σ, t) where $\sigma \in \mathcal{D}(X)$ is the initial state and $t \colon X \rightsquigarrow X$ is the transition channel. It can be seen as a special case of an HMM, where the emission channel is the identity. Thus we will use validity for HMMs also for Markov chains, where $(\sigma, t) \models \vec{p}$ should be understood as $(\sigma, t, \mathit{id}) \models \vec{p}$, in the sense of Definition 6.7.2.

Below we first look at the simple situation where we like to learn a Markov chain directly from (point) observations, in the style of frequentist learning *Flrn*. Subsequently we look at the more difficult situation of increasing the validity of a given sequence of observations. This will

be done via the sum-increase approach of Section 8.2. It leads to a channel-based description of the Baum-Welch algorithm Baum et al. (1970) (for Markov chains).

8.8.1 Learning by counting elements, without priors

We consider an example. Let $X = \{A, C, G, T\}$ contain the letters for the four DNA nucleobase elements adenine (A), cytosine (C), guanine (G) and thymine (T). Suppose we are given two 8-letter sequences $\alpha_1, \alpha_2 \in \mathcal{L}(X)$, namely:

$$\alpha_1 = [C, G, A, A, T, T, G, C] \qquad \alpha_2 = [A, C, T, C, C, A, T, G].$$

What can we learn from this? In particular, is there a Markov chain that (best) fits these data?

For the initial state we only have the first entries in the two sequences, namely C and A. This gives a multiset $1|A\rangle + 1|C\rangle$, which yields the initial state $\sigma = \frac{1}{2}|A\rangle + \frac{1}{2}|C\rangle$, via frequentist learning *Flrn*.

For the transition channel we turn the above two sequences of letters into sequences of consecutive pairs of letters, as in:

$$\begin{aligned}&[(C,G),(G,A),(A,A),(A,T),(T,T),(T,G),(G,C)]\\&[(A,C),(C,T),(T,C),(C,C),(C,A),(A,T),(T,G)]\end{aligned}$$

We combine these two lists and turn them into a multiset $\psi \in \mathcal{M}(X \times X)$ of pairs:

$$\begin{aligned}\psi = {}& 1|A,A\rangle + 1|A,C\rangle + 2|A,T\rangle + 1|C,A\rangle + 1|C,C\rangle + 1|C,G\rangle + 1|C,T\rangle\\&+ 1|G,A\rangle + 1|G,C\rangle + 1|T,C\rangle + 2|T,G\rangle + 1|T,T\rangle.\end{aligned}$$

Applying frequentist learning to ψ gives a joint state $\mathit{Flrn}(\psi) \in \mathcal{D}(X \times X)$, from which we can extract the transition channel $t\colon X \rightsquigarrow X$ by disintegration. This yields:

$$\begin{aligned}t(A) &= \tfrac{1}{4}|A\rangle + \tfrac{1}{4}|C\rangle + \tfrac{1}{2}|T\rangle & t(G) &= \tfrac{1}{2}|A\rangle + \tfrac{1}{2}|C\rangle\\ t(C) &= \tfrac{1}{4}|A\rangle + \tfrac{1}{4}|C\rangle + \tfrac{1}{4}|G\rangle + \tfrac{1}{4}|T\rangle & t(T) &= \tfrac{1}{4}|C\rangle + \tfrac{1}{2}|G\rangle + \tfrac{1}{4}|T\rangle.\end{aligned}$$

We then obtain validities:

$$(\sigma, t) \models \alpha_1 = \tfrac{1}{2048} \approx 0.000488 \qquad (\sigma, t) \models \alpha_2 = \tfrac{1}{4096} \approx 0.000244.$$

8.8.2 Finding a better Markov chain

We are going to apply the sum increase techniques from Section 8.2 to learn a new, better, Markov chain, starting form a given (prior) one. This given chain will be written as (σ, t), with initial distribution $\sigma \in \mathcal{D}(X)$ and transition channel $t\colon X \rightsquigarrow X$. We further assume that a list of ℓ factors $\vec{p} = p_1, \dots p_\ell \in \mathit{Fact}(X)$ is given, as consecutive observations, from which we like to learn. Our aim is to find a better Markov chain (σ', t') with:

$$(\sigma', t') \models \vec{p} \;\geq\; (\sigma, t) \models \vec{p}.$$

We recall from Definition 6.7.4 the filtered sequence of (ℓ+1)-many distributions $\sigma_1, \dots, \sigma_{\ell+1} \in \mathcal{D}(X)$, determined as $\sigma_1 := \sigma$ and $\sigma_{i+1} := t_*(\sigma_i|_{p_i})$. We introduce three new sequences, of factors

$q_1, \ldots, q_{\ell+1} \in Fact(X)$, of distributions $\gamma_1, \ldots, \gamma_{\ell+1} \in \mathcal{D}(X)$, and of channels $t|_{q_i} \colon X \rightsquigarrow X$. First, the factors are defined, 'backwards', as in:

$$q_{\ell+1} \coloneqq \mathbf{1} \qquad \text{and} \qquad q_{i-1} \coloneqq p_{i-1} \mathbin{\&} t^*(q_i). \tag{8.42}$$

We then define the distributions γ_i, for $1 \leq i \leq \ell$, via entry-wise updating, simply as:

$$\gamma_i \coloneqq \sigma_i\big|_{q_i}. \tag{8.43}$$

We can now formulate the main result about updating Markov chains.

Theorem 8.8.1 *In the above setting, starting from a Markov chain (σ, t) we define the* Baum-Welch *improvement $BW(\sigma, t, \vec{p})$ of (σ, t) with factors $\vec{p} = p_1, \ldots, p_\ell$ as the new Markov chain:*

$$BW(\sigma, t, \vec{p}) \coloneqq (\sigma', t') \qquad \textit{where} \qquad \begin{cases} \sigma' \coloneqq \gamma_1 \\ t'(x) \coloneqq Flrn\Big(\displaystyle\sum_{1 \leq i < \ell} \gamma_i(x) \cdot t|_{q_{i+1}}(x)\Big). \end{cases}$$

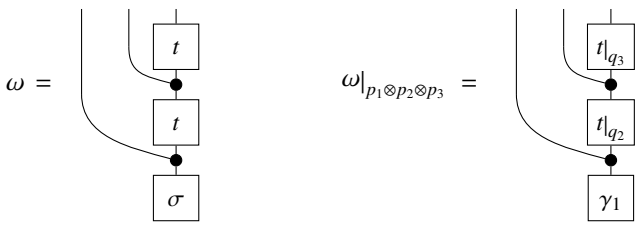
Then: $BW(\sigma, t, \vec{p}) \models \vec{p} \geq (\sigma, t) \models \vec{p}$.

This result describes just one improvement step. In Figure 8.9 in the next section we shall describe it in pseudo code, for hidden Markov models (with an emission channel). What is called the *Baum-Welch algorithm* for Markov chains involves repeating this step multiple times, with the same factors $\vec{p}$. The number of repetitions may depend on some precision requirement, like the increase in validity being smaller than some epsilon value.

We start with an explanation of what happens in the above theorem, for $\ell = 3$ factors p_1, p_2, p_3. We then have:

$$q_1 = p_1 \mathbin{\&} t^*(p_2 \mathbin{\&} t^*(p_3)) \qquad q_2 = p_2 \mathbin{\&} t^*(p_3) \qquad q_3 = p_3 \qquad q_4 = \mathbf{1}.$$

We write $\omega \in \mathcal{D}(X \times X \times X)$ for the joint distribution as given by the string diagram below on the left. One can show that the updated distribution $\omega|_{p_1 \otimes p_2 \otimes p_3}$ has a shape as on the right.



(8.44)

The validity of the factors p_1, p_2, p_3 in the Markov chain (σ, t) is given by the validity $\omega \models p_1 \otimes p_2 \otimes p_3$, see Definition (6.7.2) (with identity emission channel). In order to increase this validity, one can switch to the updated distribution $\omega|_{p_1 \otimes p_2 \otimes p_3}$. We show that this update indeed

has the shape given by the string diagram on the right.

$$
\begin{aligned}
&\sum_{x_1,x_2,x_3} \gamma_1(x_1)\cdot t|_{q_2}(x_1)(x_2)\cdot t|_{q_3}(x_2)(x_3)\,\big|x_1,x_2,x_3\big\rangle \\
&= \sum_{x_1,x_2,x_3} (\sigma_1|_{q_1})(x_1)\cdot t(x_1)|_{q_2}(x_2)\cdot t(x_2)|_{q_3}(x_3)\,\big|x_1,x_2,x_3\big\rangle \\
&= \sum_{x_1,x_2,x_3} \frac{\sigma(x_1)\cdot q_1(x_1)}{\sigma\models q_1}\cdot\frac{t(x_1)(x_2)\cdot q_2(x_2)}{t(x_1)\models q_2}\cdot\frac{t(x_2)(x_3)\cdot q_3(x_3)}{t(x_2)\models q_3}\,\big|x_1,x_2,x_3\big\rangle \\
&= \sum_{x_1,x_2,x_3} \frac{\sigma(x_1)\cdot p_1(x_1)\cdot t^*(q_2)(x_1)}{\sigma\models p_1\ \&\ t^*(p_2\ \&\ t^*(p_3))}\cdot\frac{t(x_1)(x_2)\cdot p_2(x_2)\cdot t^*(q_3)(x_2)}{t^*(q_2)(x_1)} \\
&\qquad\qquad \cdot\frac{t(x_2)(x_3)\cdot p_3(x_3)}{t^*(q_3)(x_2)}\,\big|x_1,x_2,x_3\big\rangle \\
&= \sum_{x_1,x_2,x_3} \frac{\sigma(x_1)\cdot p_1(x_1)\cdot t(x_1)(x_2)\cdot p_2(x_2)\cdot t(x_2)(x_3)\cdot p_3(x_3)}{\sigma\models\langle id,\langle id,t\rangle\circ t\rangle^*(p_1\otimes p_2\otimes p_3)}\,\big|x_1,x_2,x_3\big\rangle \\
&= \sum_{x_1,x_2,x_3} \frac{\omega(x_1,x_2,x_3)\cdot(p_1\otimes p_2\otimes p_3)(x_1,x_2,x_3)}{\omega\models p_1\otimes p_2\otimes p_3}\,\big|x_1,x_2,x_3\big\rangle \\
&= \omega|_{p_1\otimes p_2\otimes p_3}.
\end{aligned}
$$

Additionally, the following equations hold for the marginals of the update.

$$
\omega|_{p_1\otimes p_2\otimes p_3}[1,0,0] = \gamma_1 \qquad \omega|_{p_1\otimes p_2\otimes p_3}[0,1,0] = \gamma_2 \qquad \omega|_{p_1\otimes p_2\otimes p_3}[0,0,1] = \gamma_3.
$$

The first equation obviously holds. The second and third equation follow from general facts described in Exercise 8.8.1. A more general formulation occurs in Exercise 8.8.3. We now see that the distributions γ_i and channels $t|_{q_i}$, used in Theorem 8.8.1, appear naturally in this setting.

The updated joint distribution $\omega|_{p_1\otimes p_2\otimes p_3}$ on the right in (8.44) involves two *different* channels $t|_{q_2}$ and $t|_{q_3}$. For an (updated) Markov chain we need to have *equal* channels in these places, as on the left in (8.44). What Baum-Welch does in the formulation of the updated channel t' in Definition 8.8.1 is to take a suitable (normalised) weigthed sum of these updated channels $t|_{q_2}$ and $t|_{q_3}$. In this situation, with three factors, we get the better channel t' given on $x\in X$ as:

$$
\begin{aligned}
t'(x) &= \mathit{Flrn}\Big(\gamma_1(x)\cdot t\big|_{q_2}(x) + \gamma_2(x)\cdot t\big|_{q_3}(x)\Big) \\
&= \mathit{Flrn}\Big(\sigma\big|_{p_1\&t^*(p_2\&t^*(p_3))}(x)\cdot t\big|_{p_2\&t^*(p_3)}(x) + t_*(\sigma|_{p_1})\big|_{p_2\&t^*(p_3)}(x)\cdot t\big|_{p_3}(x)\Big).
\end{aligned}
$$

The latter expanded formulation uses only the original Markov chain (σ,t) and the factors p_1,p_2,p_3. The posterior validity of these factors in the resulting, updated Markov chain will be more than in ω, and thus in the original chain. This gives the required increase of validity.

The rest of this section will be devoted to the proof of Theorem 8.8.1. It works via the sum increase from Lemma 8.2.1 that has been useful before. We start with several technical properties of the q's and γ's in definitions (8.42) and (8.43). They are collected below, for laters use.

Lemma 8.8.2 *In the above situation, described before Definition 8.8.1,*

1 *The (prior) validity* $(\sigma,t)\models\vec{p}$ *equals* $\sigma_1\models q_1$*;*

2 $\sigma_{i+1}\models q_{i+1} = \dfrac{\sigma_i\models q_i}{\sigma_i\models p_i}$ *and thus* $\sigma_i\models q_i = \dfrac{\sigma\models q_1}{\prod_{j<i}\sigma_j\models p_j}$*;*

3 $\gamma_i(x) = \dfrac{(\prod_{j<i}\sigma_j\models p_j)\cdot\sigma_i(x)\cdot p_i(x)\cdot t^*(q_{i+1})(x)}{\sigma\models q_1}$.

Proof. 1 Directly by Proposition 6.7.3.

2 By unravelling the definitions and using Bayes' product rule we get:

$$\sigma_{i+1} \models q_{i+1} \;=\; t_*(\sigma_i|_{p_i}) \models q_{i+1} \;=\; \sigma_i|_{p_i} \models t^*(q_{i+1}) \;=\; \frac{\sigma_i \models p_i \;\&\; t^*(q_{i+1})}{\sigma_i \models p_i} \;=\; \frac{\sigma_i \models q_i}{\sigma_i \models p_i}.$$

The second part of point (2) follows by induction.

3 We have:

$$\begin{aligned}\gamma_i(x) \;=\; \sigma_i|_{q_i}(x) \;=\; \frac{\sigma_i(x)\cdot q_i(x)}{\sigma_i \models q_i} \;&=\; \frac{\sigma_i(x)\cdot p_i(x)\cdot t^*(q_{i+1})(x)}{\sigma_1 \models q_1 / \prod_{j<i} \sigma_j \models p_j} \qquad \text{by (2)}\\ &=\; \frac{(\prod_{j<i} \sigma_j \models p_j)\cdot \sigma_i(x)\cdot p_i(x)\cdot t^*(q_{i+1})(x)}{\sigma_1 \models q_1}.\end{aligned}$$

❑

We now turn to the sum-increase setting of Lemma 8.2.1, in order to get a better version of both σ and t, we take $F\colon X \times \mathcal{D}(X)^X \times X^\ell \to \mathbb{R}_{\geq 0}$, to be:

$$\begin{aligned}F(\sigma, t, \vec{x}) \;&:=\; \sigma(x_1)\cdot p_1(x_1)\cdot t(x_1)(x_2)\cdot p_2(x_2)\;\cdots\; t(x_{\ell-1})(x_\ell)\cdot p_\ell(x_\ell)\\ &=\; \sigma(x_1)\cdot p_1(x_1)\cdot \prod\nolimits_{1\leq k<\ell} t(x_k)(x_{k+1})\cdot p_{k+1}(x_{k+1}).\end{aligned} \tag{8.45}$$

We also collect some results about the function F.

Lemma 8.8.3 *1* $F_1(\sigma, t) := \sum_{\vec{x}} F(\sigma, t, \vec{x}) = (\sigma, t) \models \vec{p}$*;*

2 For $a, b \in X$ *and* $1 \leq k < \ell$ *one has:*

$$\sum_{x_1,\ldots,x_{k-1}} \sum_{x_{k+2},\ldots,x_\ell} F(\sigma, t, x_1, \ldots, x_{k-1}, a, b, x_{k+2}, \ldots, x_\ell) \;=\; (\sigma \models q_1)\cdot \gamma_k(a)\cdot t|_{q_{k+1}}(a)(b).$$

Proof. 1 Since, using Lemma 8.8.2 (1),

$$F_1(\sigma, t) \;=\; \sum_{\vec{x}} F(\sigma, t, \vec{x}) \;=\; \sigma \models q_1 \;=\; (\sigma, t) \models \vec{p}.$$

2 By carefully keeping track of the indices:

$$\begin{aligned}&\sum_{x_1,\ldots,x_{k-1}} \sum_{x_{k+2},\ldots,x_\ell} F(\sigma, t, x_1, \ldots, x_{k-1}, a, b, x_{k+2}, \ldots, x_\ell)\\ &= \left(\sum_{x_1,\ldots,x_{k-1}} \sigma(x_1)\cdot p_1(x_1)\cdot t(x_1)(x_2)\;\cdots\; p_{k-1}(x_{k-1})\cdot t(x_{k-1})(a)\right)\\ &\qquad \cdot p_k(a)\cdot t(a)(b)\cdot p_{k+1}(b)\cdot \left(\sum_{x_{k+2},\ldots,x_\ell} t(b)(x_{k+2})\;\cdots\; p_\ell(x_\ell)\right)\\ &= (\sigma_1 \models p_1)\;\cdots\;(\sigma_{k-1} \models p_{k-1})\cdot \sigma_k(a)\\ &\qquad \cdot p_k(a)\cdot t(a)(b)\cdot p_{k+1}(b)\cdot t^*(q_{k+2})(b)\\ &= \left(\prod_{j<k} \sigma_j \models p_j\right)\cdot \sigma_k(a)\cdot p_k(a)\cdot t(a)(b)\cdot q_{k+1}(b)\\ &= (\sigma \models q_1)\cdot \gamma_k(a)\cdot \frac{t(a)(b)\cdot q_{k+1}(b)}{t^*(q_{k+1})(a)} \qquad \text{by Lemma 8.8.2 (3)}\\ &= (\sigma \models q_1)\cdot \gamma_k(a)\cdot t|_{q_{k+1}}(a)(b), \qquad \text{since } t^*(q_{k+1})(a) \;=\; t(a) \models q_{k+1}.\end{aligned}$$

❑

We now turn to the actual proof of the main result of this section.

Proof. (of Theorem 8.8.1) According to Lemma 8.2.1 we seek a better distribution and channel σ', t' as maximum of the function $G(\sigma, t, -, -)$, where:

$$G(\sigma, t, \sigma', t') \;=\; \sum_{\vec{x}} F(\sigma, t, \vec{x}) \cdot \ln\Big(F(\sigma', t', \vec{x})\Big).$$

Let the set X have L elements, say $X = \{x_1, \ldots, x_L\}$. Recall that we update with ℓ factors, $p_1, \ldots, p_\ell$. We write w_i as variable for $\sigma'(x_i)$ and z_{ij} for $t'(x_i)(x_j)$. We use additional variables λ and $\vec{\kappa}$, as in Example 8.2.4, in a new function H, for the Lagrange multiplier method, where:

$$\begin{aligned} &H(\vec{w}, \vec{z}, \lambda, \vec{\kappa}) \\ &:= \sum_{i_1, \ldots, i_\ell} F(\sigma, t, x_{i_1}, \ldots x_{i_\ell}) \cdot \ln\left(w_{i_1} \cdot p(x_{i_1}) \cdot \prod_{1 \leq j < \ell} z_{i_j i_{j+1}} \cdot p_{j+1}(x_{i_{j+1}})\right) \\ &\qquad + \lambda \cdot \left(1 - \sum_{1 \leq i \leq L} w_i\right) + \sum_{1 \leq i \leq L} \kappa_i \cdot \left(1 - \sum_{1 \leq j \leq L} z_{ij}\right). \end{aligned}$$

The partial derivatives of H are as follows. First,

$$\frac{\partial H}{\partial w_k}(\vec{w}, \vec{z}, \lambda, \vec{\kappa}) \;=\; \sum\nolimits_{\vec{i}, i_1 = k} \frac{F(\sigma, t, x_{\vec{i}})}{w_{i_1} \cdot p_1(x_{i_1})} \cdot p_1(x_{i_1}) \;-\; \lambda \;=\; \frac{\sigma(x_k) \cdot q_1(x_k)}{w_k} \;-\; \lambda.$$

The partial derivatives w.r.t. the auxiliary variable λ is easy:

$$\frac{\partial H}{\partial \lambda}(\vec{w}, \vec{z}, \lambda, \vec{\kappa}) \;=\; 1 - \sum_{1 \leq i \leq L} w_i.$$

By setting the previous two expressions to zero we get:

$$1 \;=\; \sum_{1 \leq k \leq L} w_k \;=\; \sum_{1 \leq k \leq L} \frac{\sigma(x_k) \cdot q_1(x_k)}{\lambda} \;=\; \frac{\sigma \models q_1}{\lambda}.$$

Hence $\lambda = \sigma \models q_1$ and thus:

$$w_k \;=\; \frac{\sigma(x_k) \cdot q_1(x_k)}{\lambda} \;=\; \frac{\sigma(x_k) \cdot q_1(x_k)}{\sigma \models q_1} \;=\; \sigma|_{q_1}(x_k) \;=\; \gamma_1(x_k).$$

Thus we conclude that the better initial state σ' equals γ_1, as claimed in Theorem 8.8.1.

We turn to getting a better channel. The easy part are the derivatives:

$$\frac{\partial H}{\partial \kappa_i}(\vec{w}, \vec{z}, \lambda, \vec{\kappa}) \;=\; 1 - \sum_{1 \leq j \leq L} z_{ij}.$$

The z-derivatives of the function H are more complicated, since these z variables may occur multiple times inside the $\ln(\,-\,)$ part of the function H. Hence we use the following subsets of vectors in which two elements $a, b \in X$ occur n times in consecutive positions:

$$vec_n(a, b) \;:=\; \{\vec{i} \in \{1, \ldots, L\}^\ell \;\big|\; \big|\{j \mid x_{i_j} = a, x_{i_{j+1}} = b\}\big| = n\}.$$

We now proceed as follows.

$$\begin{aligned} \frac{\partial H}{\partial z_{km}}(\vec{w}, \vec{z}, \lambda, \vec{\kappa}) \;&=\; \sum_{1 \leq n < \ell} \sum_{\vec{i} \in vec_n(x_k, x_m)} F(\sigma, t, x_{\vec{i}}) \cdot \frac{\partial}{\partial v_{km}} \ln\Big(\cdots (z_{k,m})^n \cdots\Big) \;-\; \kappa_k \\ &=\; \sum_{1 \leq n < \ell} \sum_{\vec{i} \in vec_n(x_k, x_m)} \frac{n \cdot F(\sigma, t, x_{\vec{i}})}{v_{km}} \;-\; \kappa_k. \end{aligned}$$

When we set all these to zero, we get for each k,

$$1 = \sum_{1\leq m\leq L} v_{km} = \sum_{1\leq m\leq L} \sum_{1\leq n<\ell} \sum_{\vec{i}\in vec_n(x_k,x_m)} \frac{n\cdot F(\sigma,t,x_{\vec{i}})}{\kappa_k}.$$

Thus we can use κ_k as normalisation term. Ignoring this normalisation for a moment, we get as better channel t',

$$\begin{aligned} t'(x_k)(x_m) = v_{km} &\sim \sum_{1\leq n<\ell} \sum_{\vec{i}\in vec_n(x_k,x_m)} n\cdot F(\sigma,t,x_{\vec{i}}) \\ &\overset{(*)}{=} \sum_{1\leq n<\ell} \sum_{y_1,\dots,y_{n-1}} \sum_{y_{n+2},\dots,y_\ell} \\ &\qquad F(\sigma,t,y_1,\dots,y_{n-1},x_k,x_m,y_{n+2},\dots,y_\ell) \\ &= \sum_{1\leq n<\ell} (\sigma\models q_1)\cdot \gamma_n(x_k)\cdot t|_{q_{n+1}}(x_k)(x_m). \end{aligned}$$

The latter step follows from Lemma 8.8.3 (2). When we re-introduce normalisation, via frequentist learning *Flrn*, we can ignore the common validity $\sigma\models q_1$ in the last line and obtain the formulation claimed in Theorem 8.8.1, namely:

$$t'(x) = Flrn\left(\sum_{y\in X}\sum_{1\leq n<\ell} \gamma_n(x)\cdot t|_{q_{n+1}}(x)(y)\,\big|y\rangle\right) = Flrn\left(\sum_{1\leq n<\ell} \gamma_n(x)\cdot t|_{q_{n+1}}(x)\right).$$

What we skipped is the marked equation $\overset{(*)}{=}$. Its validity involves a lot of book-keeping details, so we present an exemplaric proof, for a set X with $L=4$ elements $X=\{x_1,x_2,x_3,x_4\}$ and $\ell=3$ factors p_1,p_2,p_3. We first look at a case, say $k=3, m=2$, involving two different positions. Then:

$$\begin{aligned} &\sum_{1\leq n<\ell} \sum_{\vec{i}\in vec_n(x_3,x_2)} n\cdot F(\sigma,t,x_{\vec{i}}) \\ &= \sum_{(i,j)\neq(3,2)} 1\cdot F(\sigma,t,x_3,x_2,x_i,x_j) + 1\cdot F(\sigma,t,x_i,x_j,x_3,x_2) \\ &\qquad + \sum_{i,j} 1\cdot F(\sigma,t,x_i,x_3,x_2,x_j) \;+\; 2\cdot F(\sigma,t,x_3,x_2,x_3,x_2) \\ &= \sum_{i,j} F(\sigma,t,x_3,x_2,x_i,x_j) + F(\sigma,t,x_i,x_3,x_2,x_j) + F(\sigma,t,x_i,x_j,x_3,x_2). \end{aligned}$$

We also consider the case with equal positions, say $k=m=2$. Then:

$$\begin{aligned} &\sum_{1\leq n<\ell} \sum_{\vec{i}\in vec_n(x_2,x_2)} n\cdot F(\sigma,t,x_{\vec{i}}) \\ &= \sum_{i\neq 2,j} 1\cdot F(\sigma,t,x_2,x_2,x_i,x_j) + 1\cdot F(\sigma,t,x_j,x_i,x_2,x_2) \\ &\qquad + \sum_{i\neq 2, j\neq 2} 1\cdot F(\sigma,t,x_i,x_2,x_2,x_j) \\ &\qquad + \sum_{i\neq 2} 2\cdot F(\sigma,t,x_i,x_2,x_2,x_2) + \sum_{j\neq 2} 2\cdot F(\sigma,t,x_2,x_2,x_2,x_j) \\ &\qquad + 3\cdot F(\sigma,t,x_2,x_2,x_2,x_2) \\ &= \sum_{i,j} F(\sigma,t,x_2,x_2,x_i,x_j) + F(\sigma,t,x_i,x_2,x_2,x_j) + F(\sigma,t,x_i,x_j,x_2,x_2). \end{aligned}$$

Now that we see the pattern, we can consider the equation $\overset{(*)}{=}$ proven. This completes the proof of Theorem 8.8.1. ❑

We shall briefly discuss one example now. The next section contains more.

Example 8.8.4 The Wikipedia entry on Bauwm-Welch[2] contains a simple illustration with daily observations about chickens laying eggs. Chickens are assumed to be in two possible (hidden) configurations / positions, so $X = \{x_1, x_2\}$. The initial distribution is $\sigma = \frac{1}{5}|x_1\rangle + \frac{4}{5}|x_2\rangle$, with transition channel $t\colon X \rightsquigarrow X$ given by:

$$t(x_1) \;=\; \tfrac{1}{2}|x_1\rangle + \tfrac{1}{2}|x_2\rangle \qquad\qquad t(x_2) \;=\; \tfrac{3}{10}|x_1\rangle + \tfrac{7}{10}|x_2\rangle.$$

We shall use the predicate $p = \frac{7}{10} \cdot \mathbf{1}_{x_1} + \frac{1}{5} \cdot \mathbf{1}_{x_2}$, together with its orthocomplement $p^\perp = \frac{3}{10} \cdot \mathbf{1}_{x_1} + \frac{4}{5} \cdot \mathbf{1}_{x_2}$.

(Actually, the Wikipedia example involves a hidden Markov model with an emission channel $e\colon X \rightsquigarrow \{E, E^\perp\}$ describing the egg and no-egg probability $e(x_1) = \frac{7}{10}|E\rangle + \frac{3}{10}|E^\perp\rangle$ and $e(x_2) = \frac{1}{5}|E\rangle + \frac{4}{5}|E^\perp\rangle$. Here we use the resulting predicates $p = e^*(\mathbf{1}_E)$ and thus $p^\perp = e^*(\mathbf{1}_{E^\perp})$ only, ignore the emission channel e, and concentrate on learning the next initial distribution and transition channel.)

The example involves a list of 10 observations, which, in the form of predicates, takes the form:

$$obs \;=\; [p^\perp, p^\perp, p^\perp, p^\perp, p^\perp, p, p, p^\perp, p^\perp, p^\perp].$$

The prior validity $(\sigma, t) \models obs$ is around 0.00398.

We can in this case also form a joint distribution $\omega \in \mathcal{D}(X^{10})$, as on the left in (8.44), but now with many more branches — and 9 occurrences of the channel t. The validity $\omega \models \bigotimes obs$ is then also, by definition, equal to the prior validity 0.00398. We can do the Bayesian update, as on the right in (8.44). It gives a much higher validity — by Theorem 6.1.6:

$$\omega\big|_{\bigotimes obs} \models \bigotimes obs \;\approx\; 0.0237. \tag{8.46}$$

As we have discussed in the paragraphs before Lemma 8.8.2, we use the different channels in the updated joint distribution $\omega\big|_{\bigotimes obs}$ to form a weighted sum of channels $t|_{q_i}$ to get a new (single) channel $t'\colon X \rightsquigarrow X$. Updating with the above list of predicates *obs* yields, by Theorem 8.8.1, as new distribution and channel:

$$\sigma' \;=\; 0.072|x_1\rangle + 0.928|x_2\rangle \qquad\qquad \begin{aligned} t'(x_1) \;&=\; 0.439|x_1\rangle + 0.561|x_2\rangle \\ t'(x_2) \;&=\; 0.214|x_1\rangle + 0.786|x_2\rangle. \end{aligned}$$

The posterior validity $(\sigma', t') \models obs$ is 0.00864, more than twice the prior validity — but a lot less than the validity 0.0237 after the Bayesian update (8.46).

The same σ', t' appear if we do a single run of the Python implementation `hmmlearn` of Bauwm-Welch in this situation, see Subsection 8.9.3 for more about `hmmlearn`. The Wikipedia page on Baum-Welch however, describes a different new channel t^*. It is obtained via a different procedure, by counting pairs, a bit like in Subsection 8.8.1, but taking priors into account. Actually, this counting of pairs also happens, more deeply hidden, in our approach, via the subsets vec_n in the above proof. At Wikipedia the outcome is:

$$t^*(x_1) \;=\; 0.3973|x_1\rangle + 0.6027|x_2\rangle \qquad\qquad t^*(x_2) \;=\; 0.1833|x_1\rangle + 0.8167|x_2\rangle.$$

[2] Consulted on Feb. 4, 2025.

No new initial state is given on Wikipedia, but if we reuse σ' as defined above, we get $(\sigma', t^*) \models obs \approx 0.00593$. This validity is higher than the prior one, but less than the validity 0.00864 that is obtained via Theorem 8.8.1.

Exercises

8.8.1 Prove that in the situation described in the beginning of Subsection 8.8.2 one has:

$$(t|_{q_{i+1}})_*(\gamma_i) = \gamma_{i+1}.$$

Hint: Recall Exercise 6.3.10.

Conclude that there is a chain of morphisms in the category of kernels **Krn** of the form:

$$(X, \gamma_1) \xrightarrow{t|_{q_2}} (X, \gamma_3) \xrightarrow{t|_{q_3}} \cdots \xrightarrow{t|_{q_n}} (X, \gamma_{n+1}).$$

8.8.2 In the context of Theorem 8.8.1, consider a new sequence of joint states $\tau_1, \ldots, \tau_\ell \in \mathcal{D}(X \times X)$, via:

$$\begin{aligned} \tau_i \;:=\; \langle id, t\rangle_*(\sigma_i)\big|_{p_i \otimes q_{i+1}} &= \langle id, t|_{q_{i+1}}\rangle_*(\sigma_i|_{p_i \,\&\, t^*(q_{i+1})}) \quad \text{by Theorem 7.2.8 (2)} \\ &= \langle id, t|_{q_{i+1}}\rangle_*(\sigma_i|_{q_i}) \\ &= \langle id, t|_{q_{i+1}}\rangle_*(\gamma_i). \end{aligned}$$

Then define $\tau := \sum_{1 \le i < \ell} \tau_i$ and prove that the new channel t' in Theorem 8.8.1 can also be defined via disintegration as:

$$t' = \tau[0, 1 \,\big|\, 1, 0].$$

8.8.3 Write $\omega_n = \langle id, t\rangle_n \gg\!= \sigma \in \mathcal{D}(X^n \times X)$, for the joint state defined in (6.24) with $e = id$. Prove that the states $\gamma_i \in \mathcal{D}(X)$ in (8.43) arise as i-th marginal:

$$\gamma_i = \omega_\ell\big|_{p_1 \otimes \cdots \otimes p_\ell \otimes \mathbf{1}}[0, \ldots, 0, 1, 0, \ldots, 0],$$

with the (single) 1 in the mask on the i-th position, for $1 \le i \le \ell$, in the mask of length $\ell + 1$.

8.9 Learning hidden Markov models

In this section we extend the Baum-Welch algorithm from Markov chains to hidden Markov models (HMM). This means that we have to deal with an additional emission channel. We shall see that the improvement approach for Markov chains in the previous section carries over HMMs. It only involves a bit more book-keeping. We shall concentrate on what is new, namely the emission channel. As before, we use a channel-based formulation and allow arbitrary factors as observables, and not just point observations.

After stating an proving the main result, we elaborate an example from the literature.

8.9.1 Finding a better hidden Markov model

Throughout this subsection we shall use:

- a fixed Hidden Markov model $\mathcal{H} = (1 \overset{\sigma}{\rightsquigarrow} X \overset{t}{\rightsquigarrow} X \overset{e}{\rightsquigarrow} Y)$;
- a list $\vec{p} = p_1, \dots, p_\ell$ of factors on Y, from which we wish to learn;
- the filtered sequence of distributions $\sigma_1, \dots, \sigma_{n+1} \in \mathcal{D}(X)$ with $\sigma_1 := \sigma$ and $\sigma_{i+1} := t_*(\sigma|_{e^*(p_i)})$, as introduced in Definition 6.7.4;
- adaptations of the factors $q_1, \dots q_{\ell+1}$ on X, defined in (8.42) to accomodate the emission channel $e\colon X \rightsquigarrow Y$, namely:

$$q_{\ell+1} := \mathbf{1} \qquad \text{and} \qquad q_{i-1} := e^*(p_{i-1}) \mathbin{\&} t^*(q_i). \tag{8.47}$$

- Definitions of distributions $\gamma_1 \dots, \gamma_{\ell+1} \in \mathcal{D}(X)$, defined as $\gamma_i := \sigma_i\big|_{q_i}$, like in (8.43), but now with the factors q_i as defined in the previous bullet.

Our aim is to define a newly learned HMM (σ', t', e').

Theorem 8.9.1 *In the above setting, starting from a Markov chain (σ, t, e) we define the* Baum-Welch *improvement of (σ, t, e) with factors $\vec{p} = p_1, \dots, p_\ell$ as:*

$$BW(\sigma, t, e, \vec{p}) := (\sigma', t', e'),$$

where $\sigma' := \gamma_1$ and:

$$t'(x) := \mathit{Flrn}\left(\sum_{1 \le i < \ell} \gamma_i(x) \cdot t|_{q_{i+1}}(x)\right) \qquad e'(x) := \mathit{Flrn}\left(\sum_{1 \le i \le \ell} \gamma_i(x) \cdot e|_{p_i}(x)\right).$$

Then: $BW(\sigma, t, e, \vec{p}) \models \vec{p} \ge (\sigma, t, e) \models \vec{p}$.

A subtle point is that the new transition channel t' is defined via a normalised sum of $\ell - 1$ updated channels — via sum over $1 \le i < \ell$ — whereas the new emmission channel e' involves ℓ summands, with $1 \le i \le \ell$, now using $\le \ell$ instead of $< \ell$. This makes sense, since e can be updated via all ℓ factors p_i. Updating the transition channel t proceeds via the pairs (p_i, p_{i+1}), of which there are only $\ell - 1$.

The definitions in Theorem 8.9.1 are reformulated as pseudo code in Figure 8.9, in the style that we used earlier for expectation maximisation, *e.g.* in Sections 8.5.

For convenience, we overload the operation BW and use it both for Markov chains (with three arguments) and for HMMs (with four arguments). It is not hard to see that the above newly learned state and transition channel σ', t' satisfy:

$$(\sigma', t') = BW\Big(\sigma, t, \overrightarrow{e^*(p_i)}\Big),$$

with BW as in Theorem 8.8.1. This is in fact what happened in Example 8.8.4.

For the proof of Theorem 8.9.1 we use Lemma 8.2.1 once more, now with function $F\colon \mathcal{D}(X) \times \mathcal{D}(X)^X \times \mathcal{D}(Y)^X \times X^\ell \times Y^\ell \to \mathbb{R}_{\ge 0}$ defined by:

$$\begin{aligned} F(\sigma, t, e, \vec{x}, \vec{y}) &:= \sigma(x_1) \cdot e(x_1)(y_1) \cdot p_1(y_1) \cdot t(x_1)(x_2) \cdot e(x_2)(y_2) \cdot p_2(y_2) \;\cdots \\ &\qquad \cdots\; t(x_{\ell-1})(x_\ell) \cdot e(x_\ell)(y_\ell) \cdot p_\ell(y_\ell) \\ &= \sigma(x_1) \cdot e(x_1)(y_1) \cdot p_1(y_1) \cdot \textstyle\prod_{1 \le k < \ell} t(x_k)(x_{k+1}) \cdot e(x_{k+1})(y_{k+1}) \cdot p_{k+1}(y_{k+1}). \end{aligned}$$

```
def BaumWelch ( init, trans, emit, facts ):
   emit_updates = [ emit|_p for p in facts ]
   filters = [ init ]
   for p in facts:
      filters = filters + [ trans_*( filters[-1]|_emit*(p) ) ]
   pulls = [ truth ]
   for p in reversed(facts):
      pulls = [ emit*(p) & trans*(pulls[0]) ] + pulls
   new_init = pulls[0]
   updates = [ filters[i]|_pulls[i] for i < len(facts) ]
   def new_emit(x):
      return Flrn( sum( [ updates[i](x) * emit_updates[i](x)
                          for i < len(facts) ] ) )
   trans_updates = [ trans|_q for q in pulls ]
   def new_trans(x):
      return Flrn( sum([ updates[i](x) * trans_updates[i+1](x)
                         for i < len(facts) - 1 ]) )
   return ( new_init, new_trans, new_emit )
```

Figure 8.9 The mathematical constructions from Theorem 8.9.1 written in pseudo code, for updating a hidden Markov model given by an initial distribution `init`, a transition channel `chan`, and an emission channel `emit`, with a list of factors `facts`. The `filters` is the list of σ distributions and the `pulls` is the list of q factors, and `updates` contains the γ distributions, as described in the beginning of this subsection. The list index `-1` in `filters[-1]` is used to access the last element of the list. In this code fragment the index `i` starts at 0, and not at 1 as in Theorem 8.9.1.

Then:

$$\begin{aligned}\sum_{\vec{x},\vec{y}} F(\sigma, t, e, \vec{x}, \vec{y}) &= \sum_{\vec{x}} \sigma(x_1) \cdot t(x_1)(x_2) \cdot e^*(p_1)(x_1) \;\cdots\; t(x_{\ell-1})(x_\ell) \cdot e^*(p_\ell)(x_\ell) \\ &= \sigma \models q_1 \;=\; (\sigma, t, e) \models \vec{p}.\end{aligned}$$

We collect the main facts about the above function F that we need in the setting of HMMs, as analogues of the earlier Lemmas 8.8.2 and 8.8.3 for Markov chains. Proofs are very similar and are left to the interested reader.

Lemma 8.9.2 *In the above situation,*

1 $\gamma_i(x) \;=\; \dfrac{(\prod_{j<i} \sigma_j \models e^*(p_j)) \cdot \sigma_i(x) \cdot e^*(p_i)(x) \cdot t^*(q_{i+1})(x)}{\sigma_1 \models q_1}.$

2 For each $1 \leq k < \ell$ *one has:*

$$\sum_{x_1,\ldots,x_{k-1}} \sum_{x_{k+2},\ldots,x_\ell} \sum_{\vec{y}} F\big(\sigma, t, e, x_1, \ldots, x_{k-1}, a, b, x_{k+2}, \ldots, x_\ell, \vec{y}\big) \;=\; (\sigma_1 \models q_1) \cdot \gamma_k(a) \cdot t|_{q_{k+1}}(a)(b).$$

3 For each $1 \leq k < \ell$ *one also has:*

$$\sum_{x_1,\ldots,x_{k-1}} \sum_{x_{k+1},\ldots,x_\ell} \sum_{y_1,\ldots,y_{k-1}} \sum_{y_{k+1},\ldots,y_\ell} F\big(\sigma, t, e, x_1, \ldots, x_{k-1}, a, x_{k+1}, \ldots, x_\ell, y_1, \ldots, y_{k-1}, b, y_{k+1}, \ldots, y_\ell\big)$$
$$= (\sigma_1 \models q_1) \cdot \gamma_k(a) \cdot e|_{p_k}(a)(b).$$

❑

Proof. (of Theorem 8.9.1) The spaces X, Y are assumed to be finite, say of the form $X = \{x_1, \ldots, x_L\}$ and $Y = \{y_1, \ldots, y_M\}$. We use (indexed) variables

$$\big(w_i\big)_{1\le i\le L} \qquad \big(z_{ii'}\big)_{1\le i\le L, 1\le i'\le L} \qquad \big(u_{ij}\big)_{1\le i\le L, 1\le j\le M}.$$

for, respectively, the values $\sigma'(x_i)$ of new distribution σ', $t'(x_i)(x_{i'})$ of the new transition channel t', and $e'(x_i)(y_j)$ of the new emmission channel e'. The function that we we use for the Lagrange method is:

$$\begin{aligned}
&H(\vec{w}, \vec{z}, \vec{u}, \lambda, \vec{\kappa}, \vec{\mu}) \\
&:= \sum_{\vec{i},\vec{j}} F(\sigma, t, e, x_{\vec{i}}, y_{\vec{j}}) \cdot \ln\Big(w_{i_1} \cdot u_{i_1 j_1} \cdot p_1(y_{j_1}) \cdot \textstyle\prod_{1\le k<\ell} z_{i_k i_{k+1}} \cdot u_{i_{k+1} j_{k+1}} \cdot p_{k+1}(y_{i_{k+1}})\Big) \\
&\qquad + \lambda \cdot \Big(1 - \sum_i w_i\Big) + \sum_i \kappa_i \cdot \Big(1 - \sum_{i'} z_{ii'}\Big) + \sum_i \mu_i \cdot \Big(1 - \sum_j u_{ij}\Big).
\end{aligned}$$

As before, we take the partial derivatives of H, set them to zero, and solve the resulting equations in order to obtain the maximum. We shall only do so for the variables u_{ij}, corresponding to the new emmission channel e'. The other cases work much as in the previous section. We use Lemma 8.9.2 (3) to obtain:

$$\begin{aligned}
&\frac{\partial H}{\partial u_{km}}(\vec{w}, \vec{z}, \vec{u}, \lambda, \vec{\kappa}, \vec{\mu}) \\
&= \sum_{\vec{i},\vec{j}, i_1=k,\, j_1=m} \frac{F(\sigma, t, e, x_{\vec{i}}, y_{\vec{j}})}{w_{i_1} \cdot u_{i_1 j_1} \cdot p_1(y_{j_1})} \cdot w_{i_1} \cdot p_1(y_{j_1}) \\
&\quad + \sum_{1\le n<\ell} \sum_{\vec{i},\vec{j}, i_{n+1}=i,\, j_{n+1}=j} \frac{F(\sigma, t, e, x_{\vec{i}}, y_{\vec{j}})}{z_{i_n i_{n+1}} \cdot u_{i_{n+1} j_{n+1}} \cdot p_{n+1}(y_{j_{n+1}})} \cdot z_{i_n i_{n+1}} \cdot p_{n+1}(y_{j_{n+1}}) - \kappa_k \\
&= \frac{(\sigma_1 \models q_1) \cdot \gamma_1(x_k) \cdot e|_{p_1}(x_k)(y_m)}{u_{mk}} + \frac{\sum_{1\le n<\ell} (\sigma_1 \models q_1) \cdot \gamma_{n+1}(x_i) \cdot e|_{p_{n+1}}(x_k)(y_m)}{u_{km}} - \kappa_k \\
&= \frac{(\sigma_1 \models q_1) \cdot \sum_{1\le n\le \ell} \gamma_n(x_k) \cdot e|_{p_n}(x_k)(y_m)}{u_{km}} - \kappa_k.
\end{aligned}$$

We then obtain, by setting these to zero, for each k,

$$1 = \sum_m u_{km} = \frac{\sum_m (\sigma_1 \models q_1) \cdot \sum_{1\le n\le \ell} \gamma_n(x_k) \cdot e|_{p_n}(x_k)(x_m)}{\kappa_k} = \frac{(\sigma_1 \models q_1) \cdot \sum_{1\le n\le \ell} \gamma_n(x_k)}{\kappa_k}.$$

Hence $\kappa_k = (\sigma_1 \models q_1) \cdot \sum_{1\le n\le \ell} \gamma(x_n)$, so that:

$$\begin{aligned}
e'(x_k)(x_m) = u_{km} &= \frac{(\sigma_1 \models q_1) \cdot \sum_{1\le n\le \ell} \gamma_n(x_k) \cdot e|_{p_n}(x_k)(x_m)}{\kappa_k} \\
&= \sum_{1\le n\le \ell} \frac{\gamma_k(x_k)}{\sum_{1\le n\le \ell} \gamma_n(x_k)} \cdot e|_{p_n}(x_k)(x_m). = \mathit{Flrn}\Big(\sum_{1\le n\le \ell} \gamma_n(x_k) \cdot e|_{p_n}(x_k)\Big)(x_m).
\end{aligned}$$

This is the formula for e' in Theorem 8.9.1. ❑

8.9.2 A soft drink example

We elaborate in detail the 'crazy soft drink machine' from (Manning and Schütze, 2003, §9.2). It is an HMM with two internal positions *cp* and *ip* for 'cola preference' and 'iced tea preference',

so $X = \{cp, ip\}$, with initial distribution:

$$\sigma = 1|cp\rangle.$$

The transition channel $t\colon X \rightsquigarrow X$ is:

$$t(cp) = \tfrac{7}{10}|cp\rangle + \tfrac{3}{10}|ip\rangle \qquad\qquad t(ip) = \tfrac{1}{2}|cp\rangle + \tfrac{1}{2}|ip\rangle.$$

There are three output elements of the machine, namely: c for 'cola', i for 'iced tea' and l for 'lemon', in a set $Y = \{c, i, l\}$. The machine is called crazy because it may give a lemon drink, with some probability, in both internal positions, see the transition channel $e\colon X \rightsquigarrow Y$,

$$e(cp) = \tfrac{3}{5}|c\rangle + \tfrac{1}{10}|i\rangle + \tfrac{3}{10}|l\rangle \qquad\qquad e(ip) = \tfrac{1}{10}|c\rangle + \tfrac{7}{10}|i\rangle + \tfrac{1}{5}|l\rangle.$$

The sequence of observations — visible elements from Y — that we wish to learn from is:

$$[l, i, c]. \tag{8.48}$$

This list of elements corresponds a sequence of point predicates $p_1 = \mathbf{1}_l$, $p_2 = \mathbf{1}_i$, $p_3 = \mathbf{1}_c$. The prior validity of these predicates in the HMM is:

$$(\sigma, t, e) \models [p_1, p_2, p_3] = \tfrac{63}{2000} = 0.0315.$$

We elaborate the details of (one step of) the Baum-Welch algorithm, for this example, as in Theorem 8.9.1. Pulling the predicates p_1, p_2, p_3 on Y along the emission channel $e\colon X \rightsquigarrow Y$ yields:

$$\begin{array}{lll} e^*(p_1)(cp) = \frac{3}{10} & e^*(p_2)(cp) = \frac{1}{10} & e^*(p_3)(cp) = \frac{6}{10} \\ e^*(p_1)(ip) = \frac{2}{10} & e^*(p_2)(ip) = \frac{7}{10} & e^*(p_3)(ip) = \frac{1}{10}. \end{array}$$

The sequence of filtered states starts with $\sigma_1 = \sigma = 1|cp\rangle$ and continues with:

$$\sigma_2 = t_*(\sigma_1|_{e^*(p_1)}) = \tfrac{7}{10}|cp\rangle + \tfrac{3}{10}|ip\rangle \qquad \sigma_3 = t_*(\sigma_2|_{e^*(p_2)}) = \tfrac{11}{20}|cp\rangle + \tfrac{9}{20}|ip\rangle.$$

The pulled-back predicates q_1, q_2, q_3, q_4 from (8.47) on X satisfy $q_4 = \mathbf{1}$ and:

$$q_3 = \tfrac{3}{5} \cdot \mathbf{1}_{cp} + \tfrac{1}{10} \cdot \mathbf{1}_{ip} \qquad q_2 = \tfrac{9}{200} \cdot \mathbf{1}_{cp} + \tfrac{49}{200} \cdot \mathbf{1}_{ip} \qquad q_1 = \tfrac{63}{200} \cdot \mathbf{1}_{cp} + \tfrac{29}{1000} \cdot \mathbf{1}_{ip}.$$

The distributions $\gamma_i = \sigma_i|_{q_i}$ on X are:

$$\gamma_1 = 1|cp\rangle \qquad \gamma_2 = \tfrac{3}{10}|cp\rangle + \tfrac{7}{10}|ip\rangle \qquad \gamma_3 = \tfrac{22}{25}|cp\rangle + \tfrac{3}{25}|ip\rangle.$$

Thus, the new initial distribution is the same as the original one:

$$\sigma' = \gamma_1 = 1|cp\rangle = \sigma.$$

The new transition channel $t'\colon X \rightsquigarrow X$ is a convex sum of *two* updated channels. We elaborate it

explicitly, for one element $cp \in X$.

$$\begin{aligned}
t'(cp) &= \mathit{Flrn}\Big(\gamma_1(cp)\cdot t|_{q_2}(cp) \;+\; \gamma_2(cp)\cdot t|_{q_3}(cp)\Big) \\
&= \frac{\gamma_1(cp)}{\gamma_1(cp)+\gamma_2(cp)}\cdot t(cp)|_{q_2} + \frac{\gamma_2(cp)}{\gamma_1(cp)+\gamma_2(cp)}\cdot t(cp)|_{q_3} \\
&= \frac{1}{1+{}^{3}/_{10}}\cdot\left(\frac{t(cp)(cp)\cdot q_2(cp)}{t(cp)\models q_2}\big|cp\rangle + \frac{t(cp)(ip)\cdot q_2(ip)}{t(cp)\models q_2}\big|ip\rangle\right) \\
&\qquad + \frac{{}^{3}/_{10}}{1+{}^{3}/_{10}}\cdot\left(\frac{t(cp)(cp)\cdot q_3(cp)}{t(cp)\models q_3}\big|cp\rangle + \frac{t(cp)(ip)\cdot q_3(ip)}{t(cp)\models q_3}\big|ip\rangle\right) \\
&= \tfrac{10}{13}\cdot\left(\frac{{}^{7}/_{10}\cdot{}^{9}/_{200}}{{}^{7}/_{10}\cdot{}^{9}/_{200}+{}^{3}/_{10}\cdot{}^{49}/_{200}}\big|cp\rangle + \frac{{}^{3}/_{10}\cdot{}^{49}/_{200}}{{}^{7}/_{10}\cdot{}^{9}/_{200}+{}^{3}/_{10}\cdot{}^{49}/_{200}}\big|ip\rangle\right) \\
&\qquad + \tfrac{3}{13}\cdot\left(\frac{{}^{7}/_{10}\cdot{}^{6}/_{10}}{{}^{7}/_{10}\cdot{}^{6}/_{10}+{}^{3}/_{10}\cdot{}^{1}/_{10}}\big|cp\rangle + \frac{{}^{3}/_{10}\cdot{}^{1}/_{10}}{{}^{7}/_{10}\cdot{}^{6}/_{10}+{}^{3}/_{10}\cdot{}^{1}/_{10}}\big|ip\rangle\right) \\
&= \tfrac{10}{13}\cdot\left(\tfrac{63}{210}\big|cp\rangle + \tfrac{147}{210}\big|ip\rangle\right) + \tfrac{3}{13}\cdot\left(\tfrac{42}{45}\big|cp\rangle + \tfrac{3}{45}\big|ip\rangle\right) \\
&= \left(\tfrac{3}{13}+\tfrac{14}{65}\right)\big|cp\rangle + \left(\tfrac{7}{13}+\tfrac{1}{65}\right)\big|ip\rangle \;=\; \tfrac{29}{65}|cp\rangle + \tfrac{36}{65}|ip\rangle \;\approx\; 0.446|cp\rangle + 0.554|ip\rangle.
\end{aligned}$$

In a similar way one gets:

$$t'(ip) \;=\; \tfrac{6}{7}|cp\rangle + \tfrac{1}{7}|ip\rangle \;\approx\; 0.857|cp\rangle + 0.143|ip\rangle.$$

The new emission channel $e' : X \rightsquigarrow Y$ is a convex sum of *three* updated channels:

$$\begin{aligned}
e'(cp) &= \mathit{Flrn}\Big(\gamma_1(cp)\cdot e|_{p_1}(cp) \;+\; \gamma_2(cp)\cdot e|_{p_2}(cp) \;+\; \gamma_3(cp)\cdot e|_{p_3}(cp)\Big) \\
&= \frac{\gamma_1(cp)}{\gamma_1(cp)+\gamma_2(cp)+\gamma_3(cp)}\cdot e(cp)|_{p_1} + \frac{\gamma_2(cp)}{\gamma_1(cp)+\gamma_2(cp)+\gamma_3(cp)}\cdot e(cp)|_{p_2} \\
&\qquad + \frac{\gamma_3(cp)}{\gamma_1(cp)+\gamma_2(cp)+\gamma_3(cp)}\cdot e(cp)|_{p_3} \\
&= \frac{1}{1+{}^{3}/_{10}+{}^{22}/_{25}}\cdot 1\big|l\rangle + \frac{{}^{3}/_{10}}{1+{}^{3}/_{10}+{}^{22}/_{25}}\cdot 1\big|i\rangle + \frac{{}^{22}/_{25}}{1+{}^{3}/_{10}+{}^{22}/_{25}}\cdot 1\big|c\rangle \\
&= \frac{1}{{}^{109}/_{50}}\big|l\rangle + \frac{{}^{3}/_{10}}{{}^{109}/_{50}}\big|i\rangle + \frac{{}^{22}/_{25}}{{}^{109}/_{50}}\big|c\rangle \\
&= \tfrac{44}{109}|c\rangle + \tfrac{15}{109}|i\rangle + \tfrac{50}{109}|l\rangle \;\approx\; 0.404|c\rangle + 0.138|i\rangle + 0.459|l\rangle.
\end{aligned}$$

In a similar way one obtains:

$$e'(ip) \;=\; \tfrac{6}{41}|c\rangle + \tfrac{35}{41}|i\rangle \;\approx\; 0.146|c\rangle + 0.854|i\rangle.$$

In the end the posterior HMM-validity as a result of the single Baum-Welch learning step is higher than the prior validity of 0.0315:

$$(\sigma', t', e') \models [p_1, p_2, p_3] \;=\; 0.0869.$$

Remark 8.9.3 The posterior distribution and channels and the resulting posterior validity that we compute above coincide with the outcomes of the Python implementation `hmmlearn`, see Subsection 8.9.3 below.

As mentioned in the beginning of this subsection, the soft drink example is taken from (Manning and Schütze, 2003, §9.3.3). The newly learned emission channel that is computed there coincides with the channel e' that we calculate above — except for what seems to be a typo in Manning and Schütze (2003), namely 0.1363 instead of 0.1463 in the distribution $e'(ip)$.

However, the new transition channel in Manning and Schütze (2003) differs in a more serious manner from the one that we obtain, written as t'. Recall from Theorem 8.9.1 that the update of the transition channel involves sum of $\ell - 1$ items, whereas the update of the emission channels uses a sum of ℓ; this ℓ is the number of predicates, which is 3 in the above soft drink example.

The update of the transition channel of Manning and Schütze (2003), written below as t^*, can be reconstructed as a convex sum of *three* updated channels, as in:

$$\begin{aligned} t^*(cp) &= Flrn\left(\sum_{1\leq i\leq 3} \gamma_i(cp)\cdot t|_{q_{i+1}}(cp)\right) = \tfrac{299}{545}|cp\rangle + \tfrac{246}{545}|ip\rangle \\ &\approx 0.5486|cp\rangle + 0.4514|ip\rangle \\ t^*(ip) &= Flrn\left(\sum_{1\leq i\leq 3} \gamma_i(ip)\cdot t|_{q_{i+1}}(ip)\right) = \tfrac{33}{41}|cp\rangle + \tfrac{8}{41}|ip\rangle \\ &\approx 0.8049|cp\rangle + 0.1951|ip\rangle. \end{aligned}$$

This does not fit the Baum-Welch steps as given in Theorem 8.9.1 — or actually Theorem 8.8.1 — which was derived in a systematic manner, via the sum-increase of Lemma 8.2.1. With this update t^* one gets a lower validity than with sum of two, as in the above calculation of t'. Explicitly:

$$(\sigma', t', e') \models [p_1, p_2, p_3] = 0.0869 \qquad\qquad (\sigma', t^*, e') \models [p_1, p_2, p_3] = 0.0724.$$

Below we shall see that the `hmmlearn` implementation computes 'our' t', and not t^* from Manning and Schütze (2003). We see that Baum-Welch is a complicated algorithm that is difficult to get right in all its details.

8.9.3 Baum-Welch in Python

In the end we take a brief look at one implementation of Baum-Welch, namely in the programming language Python. It is called `hmmlearn`, first released in 2015[3] and extended since then. It supports not only finite and infinite discrete distributions (including multinomials and Poisson) but also continuous distributions (like Gaussians). Here we give a brief impression of `hmmlearn`, first by redoing the above soft drink example, and then by taking a closer look at how `hmmlearn` handles batches. This turns out be very similar to what we have seen in Example 8.4.8 and in Subsection 8.6.2 for coin sequences via multinomial distributions.

The Python code below illustrates how a `soda` HMM object is defined in `hmmlearn`, corresponding to the soft drink example of Subsection 8.9.2. The lines with the `array` expressions correspond to the initial distribution, the translation channel and the emission channel that we defined above. It is not hard to see how they match. The `hmmlearn` representation is purely in terms of matrices and does not have labels. Thus, the above emission channel definition $e(cp) = \frac{3}{5}|c\rangle + \frac{1}{10}|i\rangle + \frac{3}{10}|l\rangle$ clearly associates the probability $\frac{3}{5} = 0.6$ with the 'cola' output $|c\rangle$. In `hmmlearn` one has to keep track oneself of the orders of such label c, i, l in the matrix. This is relevant for the sequent list of observations `obs`. Its three list elements `[ 0,0,1 ]`, `[ 0,1,0 ]`, `[ 1,0,0 ]` correspond to the sequence of three point observations l, i, c in (8.48), with l at the

[3] See pypi.org/project/hmmlearn/ for details.

third place, *etc*. As we shall see below, these three lists correspond to (singleton) multisets.

```
from hmmlearn import hmm
soda = hmm.MultinomialHMM ( n_components=2, n_trials=1,
                            n_iter=1, init_params='' )
soda.startprob_    = np.array( [1, 0] )
soda.transmat_     = np.array( [ [0.7, 0.3], [0.5, 0.5] ] )
soda.emissionprob_ = np.array( [ [0.6, 0.1, 0.3], [0.1, 0.7, 0.2] ] )
obs = [ [ 0,0,1 ], [ 0,1,0 ], [ 1,0,0 ] ]
print( "Prior varlidity:", math.exp( soda.score(obs) ) )
```

HMM-validity in `hmmlearn` is called score and is computed as (natural) logarithm — in order to better handle small numbers. It is printed as e-exponent in the last line of the above code fragement. It outputs 0.315, like in Subsection 8.9.2.

In the first line below, the actual updating of the `soda` HMM takes place, from the observations as described above. Subsequently the posterior validity is printed, in this case as number `0.08692962572522758`, corresponding to the (truncated) outcome 0.0869 that we computed via Theorem 8.9.1.

```
soda.fit(obs)
print( "Posterior validity:", math.exp( soda.score(obs) ) )
print( np.round(soda.startprob_,    3) )
print( np.round(soda.transmat_,     3) )
print( np.round(soda.emissionprob_, 3) )
```

In the last three lines the initial distribution, the transition channel and the emission channel are printed, as matrices. For readability, we round to 3 decimals. This gives:

```
[1. 0.]          [[0.446 0.554]          [[0.404 0.138 0.459]
                  [0.857 0.143]]          [0.146 0.854 0. ]]
```

These numbers correspond to the ones in the learned distribution σ' and channels t' and e' in the previous subsection.

It is reassuring to see that the outcomes of the abstract, mathematical formulation of Baum-Welch in Theorem 8.9.1 coincide with the outcomes of a (standard) implementation of Baum-Welch. This coincidence happens not only in this example. However, we have not compared the actual implementation of `hmmlearn` with the constructions in Theorem 8.9.1. The implementation in Python does precede the channel-based formulations of Theorem 8.9.1.

But there is more to say. As suggested above, the `hmmlearn` algorithm learns from a list of multisets, each represented via their list of multiplicities. We have seen such batch learning situations before, in Subsection 8.6.2, where we have described them in terms of the multinomial of a channel and frequentist learning. This approach matches what happens in `hmmlearn`, as we will illustrate via an example.

We take as sets $X = \{a, b, c\}$ and $Y = \{1, 2, 3, 4\}$ with initial distribution $\sigma = \frac{1}{2}|a\rangle + \frac{1}{3}|b\rangle + \frac{1}{6}|c\rangle$

```
def BaumWelchBatch ( init, trans, emit, multisets ):
   def mulnom_emit(x): return mn[K]( emit(x) )
   (new_init, new_trans, e) =
      BaumWelch( init, trans, mulnom_emit, multisets )
   def new_emit(x): return Flrn*( e(x) )
   return ( new_init, new_trans, new_emit )
```

Figure 8.10 Handling of a list of multisets, all assumed to be of size `K`, for updating a hidden Markov model. The emission channels are adapted twice, first by post-composition with the multinomial channel `mn[K]`. After running the ordinary `BaumWelch` function from Figure 8.9, the resulting new emission channel `e` is adapated by post-composition with frequentist learning `Flrn`. When this function `BaumWelch` is called, the multisets in the list `multisets` are implicitly turned into point predicates, on the space of multisets of size `K` over the visible space of the `emit` channel.

and with two channels $t\colon X \multimap X$ and $e\colon X \multimap Y$ given as:

$$\begin{array}{rcl} t(a) &=& \frac{5}{8}|a\rangle + \frac{1}{4}|b\rangle + \frac{1}{8}|c\rangle \\ t(b) &=& \frac{1}{3}|a\rangle + \frac{1}{3}|b\rangle + \frac{1}{3}|c\rangle \\ t(c) &=& \frac{1}{4}|a\rangle + \frac{3}{5}|b\rangle + \frac{3}{20}|c\rangle \end{array} \qquad \begin{array}{rcl} e(a) &=& \frac{1}{8}|1\rangle + \frac{1}{4}|1\rangle + \frac{1}{2}|3\rangle + \frac{1}{8}|4\rangle \\ e(a) &=& \frac{1}{4}|1\rangle + \frac{1}{4}|1\rangle + \frac{1}{4}|3\rangle + \frac{1}{4}|4\rangle \\ e(a) &=& \frac{1}{6}|1\rangle + \frac{1}{12}|1\rangle + \frac{1}{3}|3\rangle + \frac{5}{12}|4\rangle \end{array}$$

We use three (natural) multisets of size 10 over Y, namely

$$\begin{array}{rcl} \varphi_1 &=& 3|1\rangle + 4|2\rangle + 2|3\rangle + 1|4\rangle \\ \varphi_2 &=& 2|1\rangle + 3|2\rangle + 2|3\rangle + 3|4\rangle \\ \varphi_3 &=& 1|1\rangle + 4|2\rangle + 2|3\rangle + 3|4\rangle. \end{array} \qquad \text{in} \qquad obs \;=\; [\varphi_1, \varphi_2, \varphi_3].$$

We can deal with these multiset as points when we change the emission channel from $e\colon X \multimap Y$ to $mn[10](e) = mn[10] \circ e\colon X \multimap \mathcal{N}[10](Y)$. Then we can compute as prior validity of the multinomially adapted HMM:

$$\Big(\sigma, t, mn[10](e)\Big) \models obs \;=\; 2.93 \cdot 10^{-7}.$$

We can update with these observations, giving:

$$(\sigma', t', e') \;=\; BW\Big(\sigma, t, mn[10](e), obs\Big) \qquad \text{where} \qquad e'\colon X \multimap \mathcal{N}[10](Y).$$

This emission channel $e'\colon X \multimap \mathcal{N}[10](Y)$ is not of the right type, but we can turn it into $e^*\colon X \multimap$ via frequentist learning, namely as $e^*(x) := Flrn_*(e'(x))$. The posterior validity is then:

$$(\sigma', t', mn[10](e^*)) \models obs \;=\; 1.02 \cdot 10^{-5}.$$

It turns out that this same outcome is obtained via `hmmlearn`, via the following code fragment.

Notice how the above batch multisets φ_i appear in the list `obs`, via their multiplicities.

```
test = hmm.MultinomialHMM ( n_components=2, n_trials=10,
                            n_iter=1, init_params='' )
test.startprob_    = np.array([ 1/2, 1/3, 1/6 ])
test.transmat_     = np.array([ [ 5/8, 1/4, 1/8 ],
                                [ 1/3, 1/3, 1/3 ],
                                [ 1/4, 3/5, 3/20 ] ])
test.emissionprob_ = np.array([ [ 1/8, 1/4, 1/2, 1/8],
                                [ 1/4, 1/4, 1/4, 1/4],
                                [ 1/6, 1/12, 1/3, 5/12 ] ])
obs = [ [ 3,4,2,1 ], [ 2,3,2,3 ], [ 1,4,2,3 ] ]
print( "Prior varlidity:", math.exp( test.score(obs) ) )
test.fit(obs)
print( "Posterior varlidity:", math.exp( test.score(obs) ) )
```

This prints, as above, in rounded form:

```
2.93e-07            1.0216e-05.
```

Also the newly learned initial distributions and channels coincide. The handling of a list of multisets as observations in this illustration is described in general form in Figure 8.10. There is a clear resemblance to what happens in Theorem 3.4.4 and to the batch approach in Subsection 8.6.2. The fact that the outcomes of `hmmlearn` can be described at this abstract level, in terms of batches via multinomial distributions and frequentist learning, demonstrates the synthesising power of the channel-based approach.

Exercises

8.9.1 Prove in the general setting of Subsection 8.9.1:

$$(\sigma, t, e) \models \vec{p} \;=\; \sigma_1 \models q_1 \;\leq\; \sigma_1|_{q_1} \models q_1 \;=\; (\sigma', t, e) \models \vec{p}$$

The first equation follows directly from (6.27).

8.9.2 By repeatedly applying Baum-Welch to the crazy soda HMM $\mathcal{H}$ from Subsection 8.9.2 we obtain an HMM $\mathcal{H}^\infty = (\sigma^\infty, t^\infty, e^\infty)$ with $\sigma^\infty = 1|cp\rangle$ and with:

$$t^\infty(cp) \;=\; 1|ip\rangle \qquad t^\infty(ip) \;=\; 1|cp\rangle \qquad e^\infty(cp) \;=\; \tfrac{1}{2}|c\rangle + \tfrac{1}{2}|1\rangle \qquad e^\infty(ip) \;=\; 1|i\rangle.$$

Check that $\mathcal{H}^\infty \models [\mathbf{1}_1, \mathbf{1}_i, \mathbf{1}_c] = \frac{1}{4}$.

8.9.3 Compute the newly learned channel t' in Subsection 8.9.2 also via the formulas in Exercise 8.8.2, as disintegration of $\tau = \tau_1 + \tau_2 + \tau_3$ where $\tau_i = \langle id, t\rangle_*(\sigma_i)\big|_{e^*(p_i)\otimes q_{i+1}}$.

8.9.4 This exercise gives an alternative way to the describe the newly learned emission channel, in the style of Exercise 8.8.2. In the context described in Subsection 8.9.1, define for $1 \leq i \leq \ell$,

$$\rho_i \;:=\; \langle e, id\rangle_*(\sigma_i)\big|_{p_i \otimes t^*(q_{i+1})} \;\in\; \mathcal{D}(Y \times X).$$

1 Show that:

$$\rho_i \;=\; \langle e|_{p_i}, id\rangle_*(\gamma_i).$$

2 Take $\rho := \sum_i \rho_i$ and prove that the disintegration $\rho[1,0\,|\,0,1]$ equals the newly learned emission channel e'.

8.9.5 For the soft drink example from Subsection 8.9.2, consider the associated joint distribution $\omega := \langle id, \langle id, t\rangle \circ t\rangle_*(\sigma)$ on $X \times X \times X$.

1 Check first that:

$$\omega \;=\; \tfrac{49}{100}\big|cp, cp, cp\big\rangle + \tfrac{21}{100}\big|cp, cp, ip\big\rangle + \tfrac{3}{20}\big|cp, ip, cp\big\rangle + \tfrac{3}{20}\big|cp, ip, ip\big\rangle,$$

and also that the prior validity $\frac{63}{2000}$ of the point predicates p_1, p_2, p_3 in this example appears as:

$$\omega \models e^*(p_1) \otimes e^*(p_2) \otimes e^*(p_3) \;=\; \tfrac{63}{2000}.$$

2 Now calculate that updating like on the right in (8.44) yields:

$$\omega\big|_{e^*(p_1)\otimes e^*(p_2)\otimes e^*(p_3)} \;=\; \tfrac{7}{25}\big|cp, cp, cp\big\rangle + \tfrac{1}{50}\big|cp, cp, ip\big\rangle + \tfrac{3}{5}\big|cp, ip, cp\big\rangle + \tfrac{1}{10}\big|cp, ip, ip\big\rangle,$$

and that:

$$\omega\big|_{e^*(p_1)\otimes e^*(p_2)\otimes e^*(p_3)} \models e^*(p_1) \otimes e^*(p_2) \otimes e^*(p_3) \;=\; \tfrac{207}{2500} \;\approx\; 0.0828.$$

This outcome is lower than the posterior HMM-validity 0.0869.

9

Causal diagram surgery and counterfactual reasoning

In many situations that we considered so far we are interested in probabilistic influence — which we determine via inference: what happens to A if we have such-and-such evidence on B? For such *conditioning* questions it is not needed to know exactly what the underlying mechanism is that causes the probability of A to change if we know more about B. But in situations where we wish to *intervene* in order to change the probability of A in a certain direction, it is relevant to know if B causes A, or maybe something else does. Interventions are meant to change the situation at hand, for instance in order to cure patients, or to increase profits. Intervention also plays a role in cognition theory, as the action part of active inference, see *e.g.* Parr et al. (2022); Smith et al. (2022); Tull et al. (2023). This chapter is about the difference between conditioning and intervening. It will explained how intervention can be described via conditioning after modification of the graph (string diagram) involved, that is, after what is called diagram surgery. Subsequently, these techniques will be used in counterfactual reasoning.

There is a well known, fundamental difference between correlation and causation. One may find a (strong) probabilistic correlation between the occurrence of lung cancer in a group of individuals and the presence of ashtrays in their homes. Is this a coincidence? What can we conclude? That ashtrays cause lung cancer? Common sense tells us that we cannot conclude that there is a causal connection between ashtrays and cancer. The situation is better understood in terms of a third element playing the role of a 'confounder', namely smoking, which causes both lung cancer and the presence of ashtrays.

The difference between correlation and causality is highly relevant. In 2008 the editor in chief of the influential new-technology magazine *Wired*, Chris Anderson, published a provocative article Anderson (2008) with title *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. In it he wrote:

> The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

This quote is meant to sketch the wider perspective. Modern generative AI tools perform remarkably well but are based on correlations, not on a causal understanding of the world. Humans have a "hardwired habit of understanding the world in terms of causes and effects" (Stein (2018)). This tendency is called apophenia. Humans are thus inclined to read things into the output of such AI tools that are not there.

This chapter will give an impression of the techniques that are used to separate correlation and causation. The exposition given here builds especially on techniques that have been developed by Judea Pearl Pearl (2009), see also Pearl et al. (2016); Pearl and Mackenzie (2019) and Slo-

man (2005); Spirtes (2010) for more background information. This approach is based on 'causal diagrams'. To quote (Pearl and Mackenzie, 2019, p.95):

> a causal diagram is a Bayesian network in which every arrow signifies a causal direction, or at least the possibility of one, in the direction of that arrow. Not all Bayesian networks are causal, and in many applications it does not matter.

The basic idea is that a causal question can never be answered from data alone; it requires a graphical, causal diagram that underlies the data, for instance in the form of a shape, as described in Section 7.8.

Causal diagrams are thus special kinds of Bayesian networks. These causal diagrams are not singled out among Bayesian networks via some mathematical *property*, but via an *intention* of the user of the diagrams, namely an intention to address causal questions. A causal diagram is simply assumed to represent the causal connections in the data. How this assumption is justified is out of scope here (but see the work on causal discovery in *e.g.* (Koller and Friedman, 2009, §21.7) or Pearl and Verma (1991); Spirtes et al. (2000)). These causal diagrams are manipulated, via what is called diagram surgery Jacobs et al. (2021a), so that confounders are eliminated and the 'true' causal influence between wires can be determined. This will be illustrated in several examples.

It turns out that these same causal diagram surgery techniques can be used for counterfactual reasoning, that is, for reasoning about alternatives for events that have already occurred. An example is: had John been in time, the outcome of the meeting would have been different. The implicit assumption is that the presence of John at the meeting would have caused another decision than the one that was actually taken. Counterfactuals are common in human reasoning. We quote from (Walton, 2023, Chap. 13, p. 335): "One of modern history's great counterfactuals is whether, if he had lived longer, Kennedy would have escalated US involvement in Vietnam as Johnson did."

The core of this chapter is a string diagrammatic version of Pearl's method, based on Jacobs et al. (2021a); Lorenz and Tull (2023). It changes a string diagram in order to disentagle influences within that diagram. The method assumes that the original diagram is 'causal', representing assumed causal connections, where a box from A to B forms a direct causal link from A as cause to B as effect. Causal connections give rise to probabilistic correlations. In order to extract the direct effect of A on B we need to eliminate from the diagram the indirect links, which are typically formed by latent (hidden) variables working as confounders. Like in the ashtray example that we started with, confounders may happen for instance via a joint cause C, that is, via causal connections $A \leftarrow C \rightarrow B$. Diagram surgery can remove such indirect common causes C so that the relevant direct connections $A \rightarrow B$, if any, can be analysed. In this way, causal influence from A to B is disentangled from correlations between A and B that arise from other parts of the diagram.

This chapter will describe causal diagram surgery in terms of four steps, see Subsection 9.1.2. The input consists of a causal (string) diagram together with a joint distribution that has the same outgoing wires as the causal diagram. The output is an 'intervention' or 'do' diagram together with an interpretation (of this diagram) in terms of the joint distribution. The first step of this four-step surgery procedure involves 'punching' a suitably placed hole in the causal diagram and adding $\perp$ and $\bar{\top}$ to the dangling wires resulting from the punch. In this way the incoming connections / boxes are removed. In the second step this mutilated string diagram is simplified

via the discard equations. In the third step the intervention diagram that we seek is extracted via (ordinary) disintegration, performed as graphical operations via bending of wires and normalisation, as elaborated in Section 7.10. In the fourth and final step interpretations are sought for the boxes in this intervention diagram. These interpretations are obtained from the assumed joint distribution, via marginalisation and/or disintegration. This fourth step typically involves a certain amount of ingenuity — and may fail.

If all this works out, the resulting channel obtained from the interpreted intervention diagram is written by Pearl as $P(B \mid do(A))$. The *do* part distinguishes this channel from an ordinary channel $P(B \mid A)$, expressed in terms of conditional probabilities. This special channel $P(B \mid do(A))$ answers intervention queries, since it gives the probability distribution on B after forcing A to have a specific value $a \in A$, separate from other influences on B. We often call this channel the 'do' or 'intervention' channel.

This conditional do notation $P(B \mid do(A))$ is not ideal, and even somewhat confusing, since it leaves the ambient graph (string diagram) itself implicit. Indeed, what is considered a causal connection is determined by the assumed, surrounding graph. An ordinary channel is a concept that can be described on its own. A causal channel is a contextual notion: it only makes sense within the wider context of the ambient graph. This is what Pearl himself insists on. For this reason we avoid Pearl's 'do' notation for conditional probability channels

As already mentioned, this chapter is very much based on the work of Pearl Pearl (2009), but there are also some significant differences.

- First of all, we use string diagrams as (directed) graphical models, with their compositional interpretation via distributions and channels. Pearl uses causal diagrams only as a visual tool, without an explicit syntax, (compositional) semantics, or equations to rewrite diagrams into simpler forms. The chapter includes several illustrations of Pearl's causal diagram notation, and also the corresponding string diagrams.
- This usage of string diagrams means that we can distinguish syntax — the first three steps of the causal surgery procedure described below — from semantics — the fourth and final step that seeks an interpretation for the mutilated string diagram — for extraction of the causal channel via disintegration. This distinction between syntax and semantics is emphasised in Jacobs et al. (2021a).
- Our description is very much at the graphical level, and does not use Pearl's equations from the do-calculus Pearl (1995). These equations try to incorporate the relevant parts of the graph — via subscripts with overlines and underlines — but are not very intuitive or easy to use — in contrast to graphical surgery.
- Finally, our approach incorporates disintegration into the graphical formalism as in Jacobs et al. (2025), via the shaded normalisation boxes and the associated graphical rewrite rules, as introduced in Section 7.10.

One could say that by putting (string) diagrams first, and not the data, our approach is possibly more Pearlian than Pearl's own approach.

We like to remark that we are somewhat uncomfortable with the phrase 'causal' in this context, since actual mechanism-based analysis Druzdzel and Simon (1993) — such as the biomedical process that leads from inhaling tabacco to lung cancer — is not part of the analysis. We see the techniques described in this chapter primarily as methods for 'deconfonding', that is for eliminating influence via confounders in string diagrams (and their interpretations). This is important

for isolating direct influence. However, the word 'causal' has already been established in this setting and so we stick to it.

This chapter first introduces techniques for extracting a causal channel from an (interpreted) string diagram. It subsequently applies these techniques to counterfactual queries. There, the standard approach is to use two copies of the relevant string diagram, one for the 'actual' and one for the 'counterfactual' world. These two copies are connected via what are called exogenous distributions. They capture the randomness in probabilistic channels. This approach, which is also described in functional form Balke and Pearl (1994); Pearl and Verma (1991), is elaborated here in terms of deterministic channels plus 'exogenous' distributions. The description below involves a general randomness pushback result (due to Fritz Fritz (2020)), which allows a decomposition of an arbitrary probabilistic channel in terms of a deterministic channel and a probability distribution (state). The outcome of counterfactual reasoning should be independent of such decompositions. As the last section illustrates, this is not always the case.

The focus of this chapter is not on new mathematical results, but on procedures for extracting causal channels and for counterfactual reasoning. These procedures are best described via illustrations. That is why this chapter, even more than other chapters in this book, involves many examples. This chapter does not in any way cover everything that is going on in the quickly growing area of causality. Instead, its main goal is to illustrate how the theory of string diagrams, with its own compositional interpretation and equational calculus, can be used in this area.

9.1 Simpson's paradox

Simpson's paradox Simpson (1951) is a famous riddle in probability which is not only interesting in itself, but which also forms a motivating example to look for causal connections. That's why we start with it here. The paradox exists in many variations. We have seen one form with urns in Example 2.2.3, see also Exercise 9.1.2 below. Here we describe a version in a medical setting.

A medical study into kidney stones involves two kinds of stones, small and large, for which two treatments are considered (labeled A and B). There is data about the succesfulness of the treatment for 1000 patients, as given in the following two tables.

Large kidney stones

treatment	**success**	**failure**
A	280	100
B	80	35

Small kidney stones

treatment	**success**	**failure**
A	120	5
B	330	50

We formalise this situation via the following sample spaces for kidney stones, treatment and success:

$$K = \{\ell, s\} \qquad T = \{A, B\} \qquad S = \{y, n\}.$$

The above two tables can then be captured jointly via a single multiset in $\mathcal{N}(K \times T \times S)$. We consider its emperical distribution $\tau \in \mathcal{D}(K \times T \times S)$, obtained by frequentist learning (normalisation):

$$\begin{aligned}\tau = {} & \tfrac{280}{1000}\big|\ell, A, y\rangle + \tfrac{100}{1000}\big|\ell, A, n\rangle + \tfrac{80}{1000}\big|\ell, B, y\rangle + \tfrac{35}{1000}\big|\ell, B, n\rangle \\ & + \tfrac{120}{1000}\big|s, A, y\rangle + \tfrac{5}{1000}\big|s, A, n\rangle + \tfrac{330}{1000}\big|s, B, y\rangle + \tfrac{50}{1000}\big|s, B, n\rangle.\end{aligned} \tag{9.1}$$

The success of the treatments can be obtained via the extracted channel $c\colon T \twoheadrightarrow S$ given by:

$$c \coloneqq \tau[0,0,1\,|\,0,1,0] \quad \text{so that} \quad \begin{cases} c(A) = \frac{80}{101}|y\rangle + \frac{21}{101}|n\rangle \approx 0.7921|y\rangle + 0.2079|n\rangle \\ c(B) = \frac{82}{99}|y\rangle + \frac{17}{99}|n\rangle \approx 0.8283|y\rangle + 0.1717|n\rangle. \end{cases} \tag{9.2}$$

We see that treatment B has a slightly higher success probability, roughly 83% against 80%.

Next we wish to consider the success rates for large stones and for small stones separately. We can obtain them by first conditioning, with the point predicates $\mathbf{1}_\ell$ and $\mathbf{1}_s$ for large and small stones, suitably weakened. We then form the two channels $T \twoheadrightarrow S$ given by:

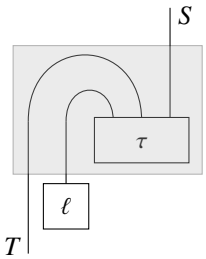
$$c_\ell \coloneqq (\tau|_{\mathbf{1}_\ell\otimes\mathbf{1}\otimes\mathbf{1}})[0,0,1\,|\,0,1,0] \qquad\qquad c_s \coloneqq (\tau|_{\mathbf{1}_s\otimes\mathbf{1}\otimes\mathbf{1}})[0,0,1\,|\,0,1,0].$$

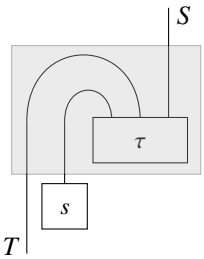
We then have:

$$\begin{aligned} c_\ell(A) &= \tfrac{14}{19}|y\rangle + \tfrac{5}{19}|n\rangle \\ &\approx 0.7368|y\rangle + 0.2632|n\rangle \\ c_\ell(B) &= \tfrac{16}{23}|y\rangle + \tfrac{7}{23}|n\rangle \\ &\approx 0.6957|y\rangle + 0.3043|n\rangle \end{aligned} \qquad\qquad \begin{aligned} c_s(A) &= \tfrac{24}{25}|y\rangle + \tfrac{1}{25}|n\rangle \\ &\approx 0.96|y\rangle + 0.04|n\rangle \\ c_s(B) &= \tfrac{33}{38}|y\rangle + \tfrac{5}{38}|n\rangle \\ &\approx 0.8684|y\rangle + 0.1316|n\rangle. \end{aligned}$$

We now see that both for large stones and for small stones treatment A gives better results. What is going on?

The string-diagrammatic descriptions of the above constructions might help. The two channels $c_\ell, c_s\colon T \twoheadrightarrow S$, with results for large and small stones separately, can be described as:

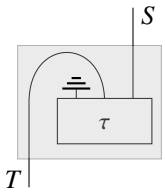
$c_\ell =$ 

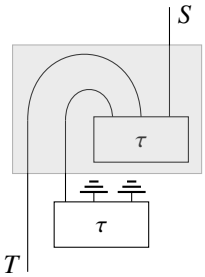
$c_s =$ 

(9.3)

Recall that we use the shaded box for the normalisation part of disintegration, see Section 7.10, and that the boxes labeled with ℓ and s represent the point distributions $1|\ell\rangle$ and $1|s\rangle$ in $\mathcal{D}(K)$ for large and small stones.

The channel c that we started from is described by the picture on the left:

$c =$ 

$d =$ 

(9.4)

The picture on the right describes an 'adjustment' channel $d\colon T \twoheadrightarrow S$, which in the terminology of Pearl Pearl (2009); Pearl et al. (2016); Pearl and Mackenzie (2019) is called a 'do' channel. It is the channel that arises from a causal analysis of this situation. Before getting into that, let's

first compute what it does. We have as first marginal:

$$\tau[1,0,0] \;=\; 0.495|\ell\rangle + 0.505|s\rangle \;=\; \tfrac{99}{200}|\ell\rangle + \tfrac{101}{200}|s\rangle.$$

Hence:

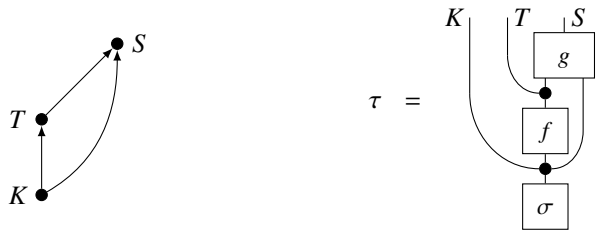
$$\begin{aligned} d(A) &= \tfrac{99}{200}\cdot c_\ell(A) + \tfrac{101}{200}\cdot c_s(A) \\ &= \tfrac{99}{200}\cdot\big(\tfrac{14}{19}|y\rangle + \tfrac{5}{19}|n\rangle\big) + \tfrac{101}{200}\cdot\big(\tfrac{24}{25}|y\rangle + \tfrac{1}{25}|n\rangle\big) \\ &= \tfrac{40353}{47500}|y\rangle + \tfrac{7147}{47500}|n\rangle \;\approx\; 0.8495|y\rangle + 0.1505|n\rangle \\ d(B) &= \tfrac{136851}{174800}|y\rangle + \tfrac{37949}{174800}|n\rangle \;\approx\; 0.7829|y\rangle + 0.2171|n\rangle. \end{aligned} \tag{9.5}$$

We see that this adjusted channel d also gives better results for treatment A, in line with the separate channels c_ℓ and c_s for large and small stones. This is not surprising, since d is a suitable convex combination of the channels c_ℓ and c_s, proportional to the convex combination $\tau[1,0,0] = \frac{99}{200}|\ell\rangle + \frac{101}{200}|s\rangle$ of large and small stones in the data.

9.1.1 Deleting backdoors via conditioning

We now start a more systematic analysis of Simpon's paradox, in causal terms, where we are looking for an intervention channel from treatment T to success S. A causal analysis always starts with a *causal diagram*. In the present setting, such a diagram is simply a string diagram whose arrows are understood as reflecting the causal relationships in the situation at hand. This means that for each box in the string diagram, the domain describes the causes and the codomain the effects. How this causal diagram comes about is usually left unspecified: it comes from some deeper understanding of the mechanics of the situation.

For the Simpson case we describe the assumed causal diagram informally on the left below — in the notation that is used in Pearl's tradition — and more formally as a string diagram on the right, for the joint distribution τ from (9.1).

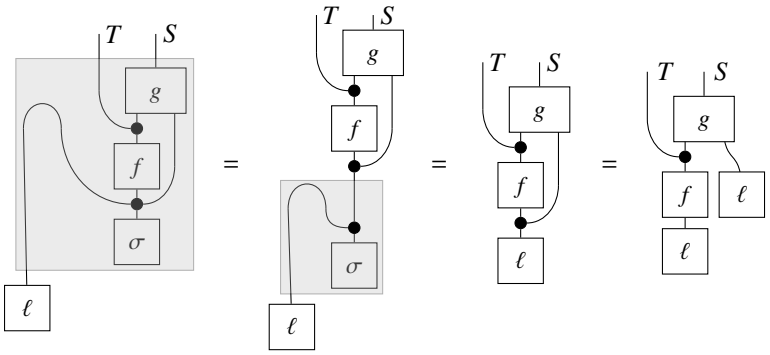


(9.6)

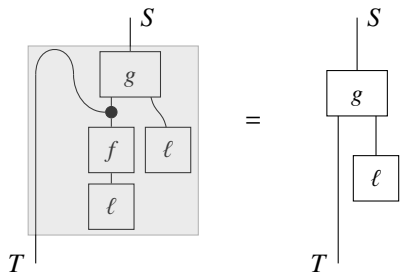
The graph on the left suggests that the kidney stone size K causes both the treatment (choice) T and the success S, and also that the treatment directly causes the success. Via these connections the kidney stone size creates a 'back door' connection between treatment and success. It forms a 'confounder'. This backdoor distorts our calculation of the connection between T and S, as represented by the channel c in (9.2). The backdoor occurs in the string diagram on the right in (9.6) as connection from T to S, downward via f, and then upward from σ to g. The aims is to get rid of this backdoor, so that the true causal influence of T on S can be extracted.

One way of deleting this backdoor is via conditioning with a point predicate. Applying it in graphical form (7.53) in the current situation yields the diagrams below. The equation in the middle uses the shaded box removal from Lemma 7.10.4 and the last step uses that point

distributions commute with copiers (2.47).



In the diagram on the right there are no indirect connections from T to S left. Hence, for the large stone case, we can obtain the direct influence of T on S, as a channel $T \twoheadrightarrow S$, via another disintegration:



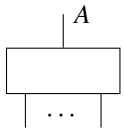
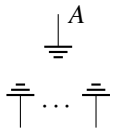
This is precisely what happns on the left in (9.3), without assuming any internal structure (9.6).

A more systematic technique for deleting back doors is via 'diagram surgery', which we shall describe more informally as: punching holes in string diagrams and cleaning up what's left. It will first be described in general form; then it will be applied in the Simpson case to obtain a channel that describes the direct, undistorted influence.

9.1.2 Causal surgery procedure in general

Suppose, quite generally, that we are interested in finding the direct influence of A on B, occurring as two types (of specific wires) in a 'causal' model. The latter model consists of a string diagram S with empty domain, together with a joint distribution ω on the codomain of S. This ω need not be an interpretation of the string diagram S, since we do not assume that we have interpretations of all the boxes in S. The causal surgery procedure involves trying to perform the following four steps.

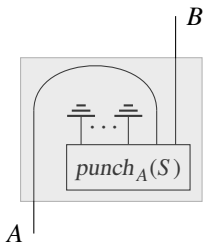
1 We punch holes in the string diagram S below the (specific) wire of type A:

each box  inside S gets replaced by  (9.7)

2 We clean up the punched diagram using the string diagram equations (for discard ⏚) from

Subsection 2.5.1. We shall write $punch_A(S)$ for the result. It contains no more 'backdoor' connections starting in A.

3 Next we discard the outgoing wires of $punch_A(S)$ that we don't need — and clean-up again, if possible, with the equations for $\overset{=}{\top}$. Now we can disintegrate and get a string diagram from A to B as in:

$do \;=$ 

(9.8)

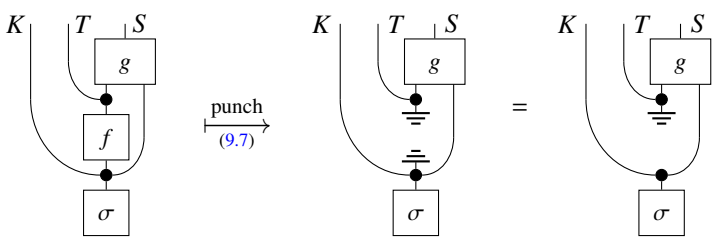
4 In the final step we seek interpretations for all the boxes which are left in $punch_A(S)$ in (9.8). These interpretations should be obtained from the joint distribution ω. If this succeeds, then we the direct channel $A \twoheadrightarrow B$ that we are interested. It captures the causal influence from A to B.

We notice that the first three of the above four steps take place at the syntactic level, dealing with string diagrams. Only the last is about the interpretation of the mutilated string diagram (9.8) as the causal channel. This last step may fail if we do not succeed to extract interpretations of the remaining channels from the original joint distribution ω.

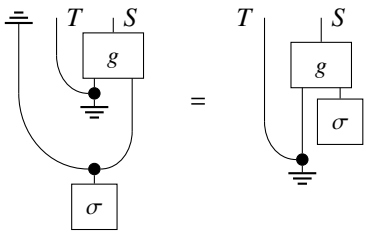
9.1.3 Causal surgery for Simpson

We elaborate the above four causal surgery steps for the kidney stone example from Simpon's paradox, where we are interested in extracting a causal channel $T \twoheadrightarrow S$. Our starting point is the assumed causal diagram structure on the right in (9.6) for the joint distribution $\tau \in \mathcal{D}(K \times T \times S)$ from (9.1). Such a diagram structure can always be found, see Exercise 7.3.6.

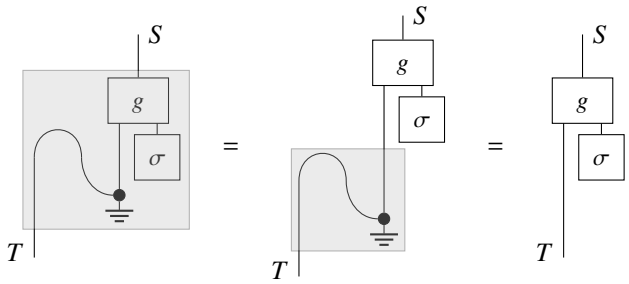
Step 1 involves punching a whole in the assumed causal diagram, below the T copy node $\bullet$, as on the left below. This diagram is subsequently rewritten to the one on the right below, following step 2.

$\tau \;=$ 

In step 3 we first discard the K-wire in the last diagram, giving:

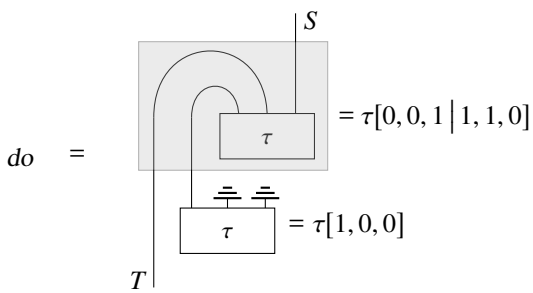


Then we disintegrate this last diagram and get the required intervention channel from T to S, as described abstractly in (9.8), via Lemma 7.10.3 (2) and Lemma 7.10.4.



(9.9)

Finally, in step 4 of the causal surgery procedure we extract from the given distribution $\tau \in \mathcal{D}(K \times T \times S)$, the interpretations for the boxes σ and g in (9.9), following the assumed causal diagram structure from (9.6). Clearly the box σ can be interpreted as the first marginal $\tau[1,0,0] \in \mathcal{D}(K)$. An interpretation of the box g can be obtained as disintegration $\tau[0,0,1 \,|\, 1,1,0] \colon T \times K \rightsquigarrow S$, see below for details. Combination as on the right in (9.9) yields the required intervention channel $do \colon T \rightsquigarrow S$ in:

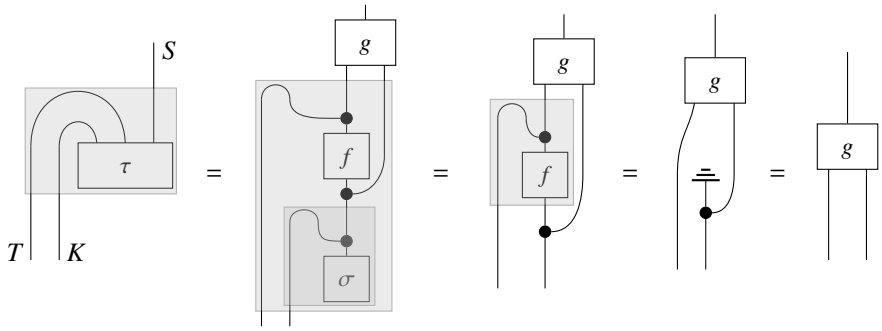


(9.10)

Hence the last step (9.9) of the general causal surgery procedure yields precisely the adjustment channel d that we have seen earlier in (9.4). Notice that the assumed causal structure in (9.6) has guided us to perform the right steps, but the end-result in (9.10) is obtained purely from the empirical distribution τ that we started from (9.1).

For completeness, the extraction of an interpretation for g is described below. The rewriting uses Lemma 7.10.3 (4) to introduce two separate normalisation boxes and uses box removal

twice, from Lemma 7.10.4.



Exercises

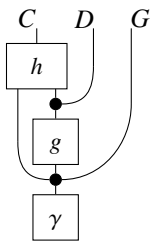
9.1.1 The following illustration is copied from (Koller and Friedman, 2009, Ex. 21.9 and 21.14). It involves sample spaces $G = \{m, f\}$ for gender, $D = \{d, d^\perp\}$ for taking a drug, and $C = \{c, c^\perp\}$ for being cured. There is a uniform gender distribution $\gamma = \frac{1}{2}|m\rangle + \frac{1}{2}|f\rangle$ and a channel $g\colon G \twoheadrightarrow D$ that describes the drug taking of of males and females:

$$g(m) = \tfrac{3}{4}|d\rangle + \tfrac{1}{4}|d^\perp\rangle \qquad\qquad g(f) = \tfrac{1}{4}|d\rangle + \tfrac{3}{4}|d^\perp\rangle.$$

There is also a cure channel $h\colon G \times D \twoheadrightarrow C$ given by:

$$\begin{aligned} h(m,d) &= \tfrac{7}{10}|c\rangle + \tfrac{3}{10}|c^\perp\rangle & \qquad h(f,d) &= \tfrac{1}{5}|c\rangle + \tfrac{4}{5}|c^\perp\rangle \\ h(m,d^\perp) &= \tfrac{4}{5}|c\rangle + \tfrac{1}{5}|c^\perp\rangle & \qquad h(f,d^\perp) &= \tfrac{2}{5}|c\rangle + \tfrac{3}{5}|c^\perp\rangle. \end{aligned}$$

1 Check that the joint distribution $\omega \in \mathcal{D}(C \times D \times G)$ obtained as interpretation of the string diagram



is

$$\begin{aligned} \omega = {} & \tfrac{21}{80}|c,d,m\rangle + \tfrac{1}{40}|c,d,f\rangle + \\ & \tfrac{1}{10}|c,d^\perp,m\rangle + \tfrac{3}{20}|c,d^\perp,f\rangle + \\ & \tfrac{9}{80}|c^\perp,d,m\rangle + \tfrac{1}{10}|c^\perp,d,f\rangle + \\ & \tfrac{1}{40}|c^\perp,d^\perp,m\rangle + \tfrac{9}{40}|c^\perp,d^\perp,f\rangle. \end{aligned}$$

2 Show that the cure marginal $\omega[1,0,0]$ is $\frac{43}{80}|c\rangle + \frac{37}{80}|c^\perp\rangle$.

3 Verify that the drug-to-cure extraction $\omega[1,0,0 \,|\, 0,1,0]\colon D \twoheadrightarrow C$ is:

$$d \mapsto \tfrac{23}{40}|c\rangle + \tfrac{17}{40}|c^\perp\rangle \qquad\qquad d^\perp \mapsto \tfrac{1}{2}|c\rangle + \tfrac{1}{2}|c^\perp\rangle.$$

Check that this same outcome can be obtained via the description in Exercise 7.3.6.

4 Perform the causal diagram surgery procedure on the above string diagram and deduce that the resulting intervention channel $D \twoheadrightarrow C$ is:

$$d \mapsto \tfrac{9}{20}|c\rangle + \tfrac{11}{20}|c^\perp\rangle \qquad\qquad d^\perp \mapsto \tfrac{3}{5}|c\rangle + \tfrac{2}{5}|c^\perp\rangle.$$

5 Does the drug work?

9.1.2 We recall the formulation of Simpon's paradox from Example 2.2.3, with two urn-games. We collect the data from that example in one multiset table $\varphi \in \mathcal{N}(\{1,2\} \times \{B,W\} \times \{R,G\})$ of the form:

$$\begin{aligned}\varphi \;=\; & 5\big|1,B,R\rangle + 6\big|1,B,G\rangle + 3\big|1,W,R\rangle + 4\big|1,W,G\rangle \\ & + 6\big|2,B,R\rangle + 3\big|2,B,G\rangle + 9\big|2,W,R\rangle + 5\big|2,W,G\rangle.\end{aligned}$$

1 Let $e\colon \{1,2\} \times \{B,W\} \rightsquigarrow \{R,G\}$ be the channel extracted (via disintegration) from the associated distribution $\mathit{Flrn}(\varphi)$. Describe e in detail and check that $e(i,-)\colon \{B,W\} \rightsquigarrow \{R,G\}$ is the channel $\mathit{Flrn} \circ c_i$ from Example 2.2.3, for $i \in \{1,2\}$.

2 Compute the adjustment channel $d\colon \{B,W\} \rightsquigarrow \{R,G\}$, as on the right in (9.4), in this example, and show that:

$$\begin{aligned} d(B) &= \tfrac{776}{1353}|R\rangle + \tfrac{577}{1353}|G\rangle \\ &\approx 0.574|R\rangle + 0.426|G\rangle \end{aligned} \qquad \begin{aligned} d(W) &= \tfrac{45}{82}|R\rangle + \tfrac{37}{82}|G\rangle \\ &\approx 0.549|R\rangle + 0.451|G\rangle. \end{aligned}$$

Conclude that in this adjusted case the black urn B is the best choice, with the highest probability of drawing a red ball R, for the convex combination of urns determined by the first marginal $\mathit{Flrn}(\varphi)[1,0,0] \in \mathcal{D}(\{1,2\})$, see Example 2.2.3 for details.

9.1.3 Here is another version of Simpson's paradox, copied from (Spiegelhalter, 2019, Table 4.2), about acceptance rates by gender at Cambridge University in 1996, with observation: "Although overall the acceptance rate was higher for men, the acceptance rate in each subject individually was higher for women."

	Women			**Men**		
	Applied	**Accepted**	%	**Applied**	**Accepted**	%
Computer Science	26	7	27%	228	58	25%
Economics	240	63	26%	512	112	22%
Engineering	164	52	32%	972	252	26%
Medicine	416	99	24%	578	140	24%
Veterinary Medicine	338	53	16%	180	22	12%
TOTAL	1,184	274	23%	2,470	584	24%

1 Use three sets $G = \{f,m\}$ for gender, $S = \{cs, ec, en, md, vm\}$ for subject and $A = \{a, a^{\perp}\}$ for admission (or not), and translate the above table into a multiset $\psi \in \mathcal{N}(G \times S \times A)$. Note that the number for non-accepted ($a^{\perp}$) is 'Applied' minus 'Accepted'.

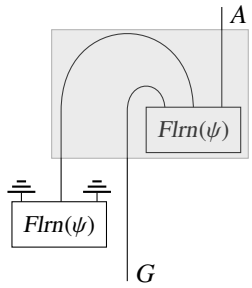
2 Extract the gender-subject channel $gs := \mathit{Flrn}(\psi)[0,1,0 \,\big|\, 1,0,0]\colon G \rightsquigarrow S$ and check that:

$$\begin{aligned} gs(f) &= \tfrac{13}{592}|cs\rangle + \tfrac{15}{74}|ec\rangle + \tfrac{41}{296}|en\rangle + \tfrac{13}{37}|md\rangle + \tfrac{169}{592}|vm\rangle \\ &\approx 0.022|cs\rangle + 0.203|ec\rangle + 0.139|en\rangle + 0.351|md\rangle + 0.285|vm\rangle \\ gs(m) &= \tfrac{6}{65}|cs\rangle + \tfrac{256}{1235}|ec\rangle + \tfrac{486}{1235}|en\rangle + \tfrac{289}{1235}|md\rangle + \tfrac{18}{247}|vm\rangle \\ &\approx 0.0923|cs\rangle + 0.207|ec\rangle + 0.394|en\rangle + 0.234|md\rangle + 0.0729|vm\rangle. \end{aligned}$$

3 Check that the gender-admission percentages in the bottom line of the table are obtained via the extracted channel $ga := Flrn(\psi)[0,0,1 \,\big|\, 1,0,0] \colon G \rightarrowtail A$ with:

$$\begin{aligned} ga(f) &= \tfrac{137}{592}|a\rangle + \tfrac{455}{592}|a^{\perp}\rangle \approx 0.231|a\rangle + 0.769|a^{\perp}\rangle \\ ga(m) &= \tfrac{292}{1235}|a\rangle + \tfrac{943}{1235}|a^{\perp}\rangle \approx 0.236|a\rangle + 0.764|a^{\perp}\rangle. \end{aligned}$$

4 Consider the following adjusted-gender-admission channel:

$aga \quad =$ 

(9.11)

Check now that the adjusted percentages for women are higher then for men, in line with the percentages for the separate subjects:

$$\begin{aligned} aga(f) &= \tfrac{523143643}{2025485280}|a\rangle + \tfrac{1502341637}{2025485280}|a^{\perp}\rangle \approx 0.258|a\rangle + 0.742|a^{\perp}\rangle \\ aga(m) &= \tfrac{1227130469}{5417310780}|a\rangle + \tfrac{4190180311}{5417310780}|a^{\perp}\rangle \approx 0.227|a\rangle + 0.773|a^{\perp}\rangle. \end{aligned}$$

In Spiegelhalter (2019) this adjustment (9.11) does not occur, but instead the paradox is explained away: "The explanation is that women were more likely to apply for the more popular and therefore more competitive subjects with the lowest acceptance rate, such as medicine and veterinary medicine, and tended not to apply to engineering which has a higher acceptance rate. In this case, therfore, we might conclude that there is no evidence of discrimination." These differences in applications are made explicit in the first item of this exercise. The marginal at the bottom of Diagram (9.11) does indeed add the proportions of the different subjects.

As an aside, the claim made in Spiegelhalter (2019) that "the acceptance rate in each subject individually was higher for women" is not entirely true since the percentages of 24% occurring twice in the above table row for Medicine are roundings. With one extra decimal precision one gets 23.8% for women and 24.2% for men. This is nitpicking.

9.2 Two causality challenges of Pearl

This section applies the causal surgery approach from the previous section to two examples of Pearl. The first one is classical, about whether smoking causes cancer, and can be found in many places in the literature. The second example exists only in graphical form.

9.2.1 Does smoking cause cancer, via tar?

Now that we have seen the causal surgery procedure in the previous section, we can apply it to a classical example in the causality literature, about a possible causal relation between smoking

and cancer. The example involves information about smoking (or not), presence of tar in lungs (or not), and having cancer or not, via the three sample spaces:

$$S = \{s, s^\perp\} \qquad T = \{t, t^\perp\} \qquad C = \{c, c^\perp\}.$$

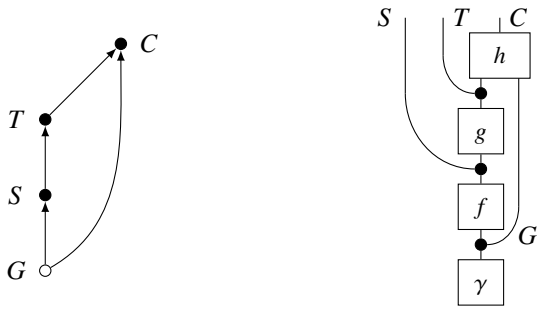
We are given a joint distribution $\sigma \in \mathcal{D}(S \times T \times C)$, namely[1]:

$$\begin{aligned}\sigma = {} & \tfrac{323}{800}|s,t,c\rangle + \tfrac{57}{800}|s,t,c^\perp\rangle + \tfrac{9}{400}|s,t^\perp,c\rangle + \tfrac{1}{400}|s,t^\perp,c^\perp\rangle \\ & + \tfrac{1}{800}|s^\perp,t,c\rangle + \tfrac{19}{800}|s^\perp,t,c^\perp\rangle + \tfrac{19}{400}|s^\perp,t^\perp,c\rangle + \tfrac{171}{400}|s^\perp,t^\perp,c^\perp\rangle.\end{aligned} \tag{9.12}$$

The (marginal) probability of getting cancer is just a bit less than one half:

$$\sigma[0,0,1] = \tfrac{19}{40}|c\rangle + \tfrac{21}{40}|c^\perp\rangle = 0.475|c\rangle + 0.525|c^\perp\rangle.$$

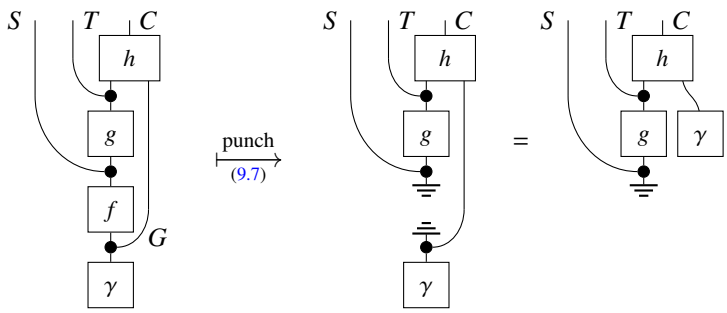
The assumed causal diagram involves a confounding genetic cause which makes people smoke and get cancer, via a sample space G that we do not further specify. The assumed causal connections are described informally on the left below, and more formally as string diagram on the right.



(9.13)

We see that not all wires are accessible. In particular, the genetic influence G is hidden. This model suggests that the influence of smoking happens via tar in lungs. But there is also 'confounding' genetic influence, via the distribution γ at the bottom. It influences both smoking and cancer. We are interested in obtaining a 'causal' channel $S \twoheadrightarrow C$, without confounding.

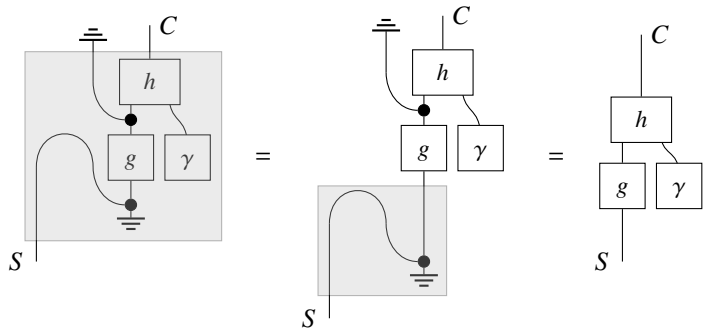
In order to determine the *causal* influence of smoking on cancer we would like to remove the confounding (backdoor) influence via the genetic factor. This can be done via the causal surgery procedure from Jacobs et al. (2021a), described in four steps in Subsection 9.1.2. In this case one punches a hole as below, removing the channel f.



Now we obtain the required intervention / do channel by discarding and disintegration, applied to

[1] We use the same numbers as in an elaborated example of Michael Nielsen in www.michaelnielsen.org/ddi/if-correlation-doesnt-imply-causation-then-what-does, consulted on March 1, 2020.

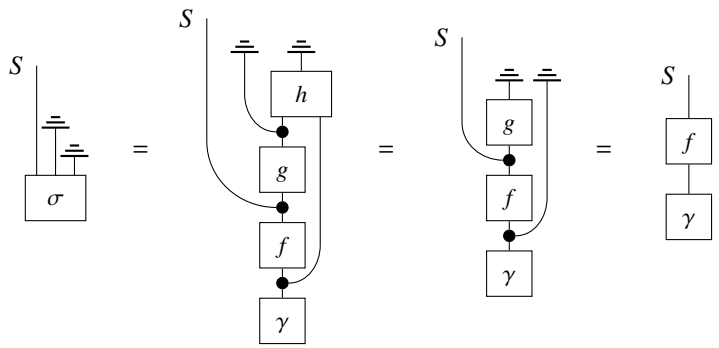
the diagram on the right (as in Jacobs et al. (2025)). Via the shaded box removal of Lemma 7.10.4 we get:



(9.14)

At this stage the question is: can we also perform the fourth 'semantic' step of causal diagram surgery and find interpretations for the boxes in the latter intervention diagram, on the right in (9.14)? We shall do so in three steps.

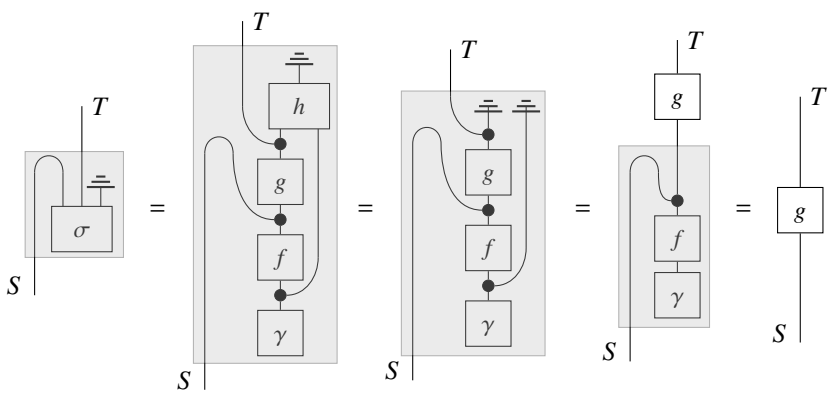
First, we take the first marginal $\sigma[1,0,0]$ for smoking. It is in fact a uniform distribution $\frac{1}{2}|c\rangle + \frac{1}{2}|c^{\perp}\rangle$. In terms of the assumed causal shape (9.13) of the joint distribution σ we get:



The second step of the surgery procudure is to extract an interpretation of the channel g. This can be done since both its domain wire S and codomain wire T are externally accessible, as in:

$$g \coloneqq \sigma[0,1,0 \,|\, 1,0,0] : S \twoheadrightarrow T.$$

Indeed, via disintegration of Bayesian networks, that is, via the rules of shaded boxes from Section 7.10, we get, using the assumed shape (9.13),



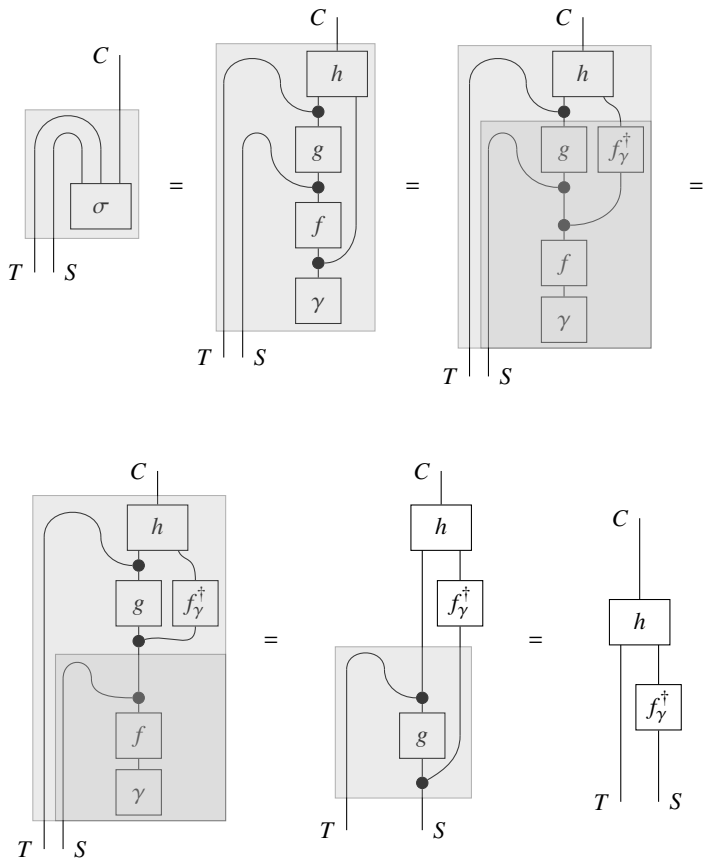


Figure 9.1 Disintegration of a Bayesian network using the rules from Section 7.10.

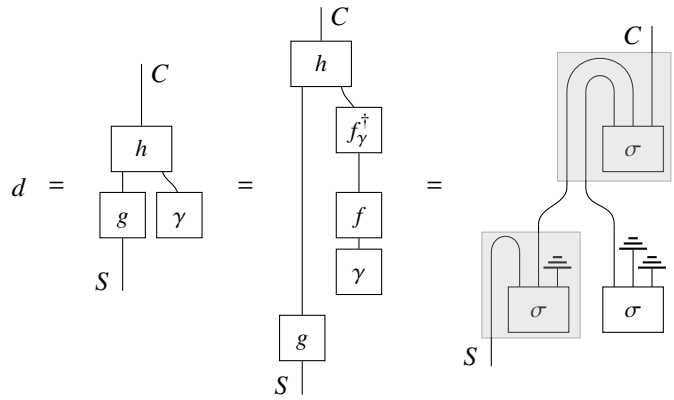
When we actually disintegrate and calculate we get:

$$g(s) \;=\; \tfrac{19}{20}|t\rangle + \tfrac{1}{20}|t^\perp\rangle \qquad\qquad g(s^\perp) \;=\; \tfrac{1}{20}|t\rangle + \tfrac{19}{20}|t^\perp\rangle.$$

For the third step we observe that we cannot extract the channel h from σ, since its G input is not externally accessible. But what we can reach turns out to be good enough for our purposes. This is elaborated in Figure 9.1, using the rewrite rules from Lemmas 7.10.3 and 7.10.4.

After these three steps we can interprete the 'do' channel $d\colon S \twoheadrightarrow C$ that we are interested in, after the punch, as described on the right in Diagram (9.14). It is re-described below, exclusively in terms of disintegration and marginalisations of the original joint distribution σ. The equation

in the middle uses the equality $\gamma = \left(f_\sigma^\dagger\right)_*(f_*(\gamma))$, from (7.4).



(9.15)

This intervention / do channel $d\colon S \twoheadrightarrow C$ can now be computed as:

$$\begin{aligned} d(s) &= \tfrac{181}{400}|c\rangle + \tfrac{219}{400}|c^\perp\rangle \approx 0.453|c\rangle + 0.547|c^\perp\rangle \\ d(s^\perp) &= \tfrac{199}{400}|c\rangle + \tfrac{201}{400}|c^\perp\rangle \approx 0.497|c\rangle + 0.503|c^\perp\rangle. \end{aligned}$$

In the beginning of this section we calculated the marginal cancer distribution as $\sigma[0,0,1] = 0.475|c\rangle + 0.525|c^\perp\rangle$. In addition, the joint distribution $\sigma \in \mathcal{D}(S \times T \times C)$ gives a do channel $d\colon S \twoheadrightarrow C$ telling us that smoking actually *reduces* the probability of attracting cancer, and not that smoking increases it.

9.2.2 The napkin problem

The next challenge exists as a sketch only, without concrete probabilities. The problem is attributed to Bareinboim in Pearl and Mackenzie (2019) and is presented as a kind of problem that Pearl solves 'on a napkin', using his do-calculus. The setting is formulated on the left in Figure 9.2, as causality diagram copied from (Pearl and Mackenzie, 2019, Fig. 7.5), where U_1 and U_2 are unobserved variables and X, Y, Z, W are observed. This leads to the formulation as string diagram, on the right in Figure 9.2.

The question is if we can find an intervention channel $X \twoheadrightarrow Y$ that captures the direct influence, without the backdoor influence via the confounding connections at the bottom. No interpretations are given for the various variables or distributions / channels. This is not needed, since diagram surgery is a graphical technique.

Our starting point is thus, as before, some joint distribution $\tau \in \mathcal{D}(W \times Z \times X \times Y)$ for which we assume a graphical structure, as on the right in Figure 9.2. A first useful observation[2] is that the situation is similar to the previous 'smoking' example, when we consider the product

[2] This observation comes from Robin Piedeleu in his online research note piedeleu.com/posts/diagrammatic-causal-inference-2-napkin-problem, consulted April 2, 2025.

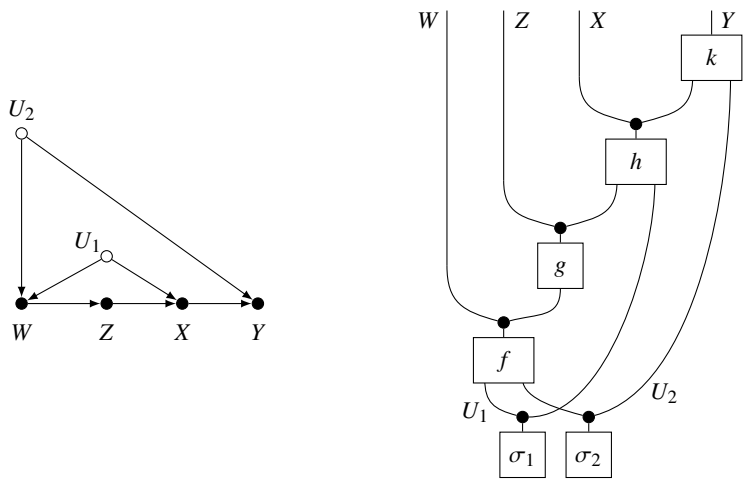
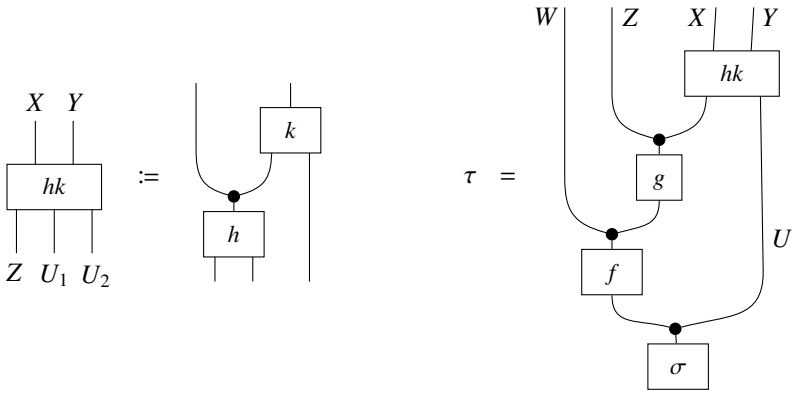
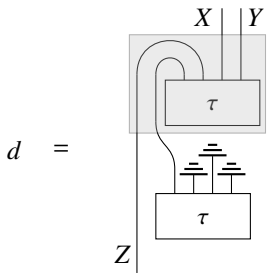


Figure 9.2 The napkin problem from Pearl and Mackenzie (2019) on the left, reformulated as string diagram on the right.

$U := U_1 \times U_2$ as a single line, write $\sigma := \sigma_1 \otimes \sigma_2$, and use hk as abbreviation in:



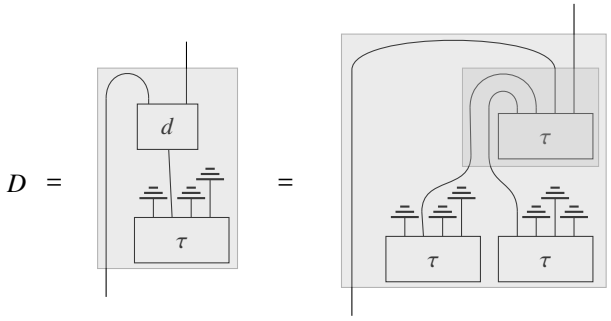
Using the techniques from the previous subsection we can form a causal channel $d\colon Z \twoheadrightarrow X \times Y$. This works as in (9.15), so that:



(9.16)

We are not interested in the input Z, so we can remove it by plugging the Z-marginal in that position. This results in a joint distribution on $X \times Y$, from which we obtain the 'napkin' do

channel $D\colon X \twoheadrightarrow Y$ via ordinary extraction:



(9.17)

The formulation on the right involves only operations on the original joint distribution τ. Notice that this do channel construction involves two nested disintegrations. Doing this in one single disintegration gives a slightly different, but incorrect outcome. Exercise 9.2.2 below relates this do channel D to the assumed string diagram on the right in Figure 9.2.

Exercises

9.2.1 In Jacobs et al. (2021a) a smoking-tar-cancer situation is analysed with the same causal structure (9.13) as used here, but with a different joint distribution, namely:

$$\begin{aligned}\sigma \;=\; & \tfrac{1}{5}\big|s,t,c\rangle + \tfrac{1}{50}\big|s,t,c^{\perp}\rangle + \tfrac{1}{20}\big|s,t^{\perp},c\rangle + \tfrac{1}{10}\big|s,t^{\perp},c^{\perp}\rangle \\ & + \tfrac{1}{50}\big|s^{\perp},t,c\rangle + \tfrac{1}{100}\big|s^{\perp},t,c^{\perp}\rangle + \tfrac{1}{10}\big|s^{\perp},t^{\perp},c\rangle + \tfrac{1}{2}\big|s^{\perp},t^{\perp},c^{\perp}\rangle.\end{aligned}$$

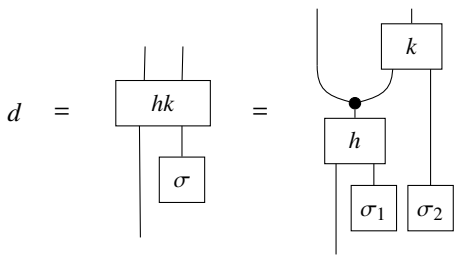
Calculate the do channel $d\colon S \twoheadrightarrow C$ as on the right in (9.15) and show that in this case:

$$\begin{aligned} d(s) &= \tfrac{4013}{7400}|c\rangle + \tfrac{3387}{7400}|c^{\perp}\rangle \;\approx\; 0.542|c\rangle + 0.458|c^{\perp}\rangle \\ d(s^{\perp}) &= \tfrac{8783}{34650}|c\rangle + \tfrac{25867}{34650}|c^{\perp}\rangle \;\approx\; 0.253|c\rangle + 0.747|c^{\perp}\rangle. \end{aligned}$$

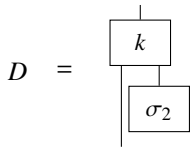
Does smoking cause cancer in this situation?

9.2.2 Assume that a joint distribution $\tau \in \mathcal{D}(W \times Z \times X \times Y)$ has a shape as described on the right in Figure 9.2.

1 Show that the do channel $d\colon Z \twoheadrightarrow X \times Y$ in (9.16) is of the form:



2 Similarly, show that $D\colon X \twoheadrightarrow Y$ in (9.17) takes the form:



This indeed gives the direct influence of X on Y, without confounders.

9.3 Two illustrations of counterfactual reasoning

This section gives an introduction to counterfactual reasoning by elaborating a 'party' example from the literature. A more systematic exposition will be given in the next section. Counterfactual reasoning is quite common and follows a 'had' pattern that is easy to recognise, like in: had we left earlier, we would have avoided the traffic jam. Such phrases contain an antecedent (about leaving early) and a conclusion (ending up in a traffic jam), where the antecedent is factually false — we did not leave early — but is still used to draw a conclusion. Such reasoning patterns form a challenge in logic. They change a small part of reality, but leave many other things the same, and then come to a conclusion. In a probabilistic setting, this small change — the 'had' part of the counterfactual — can be modelled as an intervention, of the kind that we have seen in the previous two sections on causality. We shall describe counterfactual reasoning by duplicating the situation at hand, into a 'twin network' with factual and counterfactual parts, as proposed in Balke and Pearl (1994) (see also Pearl (2009)), in such a way that the probabistic parts are shared, while the (deterministic) mechanisms in both worlds are duplicated (and thus the same). This involves decomposing a probabilistic channel into a deterministic channel together with an 'exogenous' distribution, in which all probability is concentrated. Such exogenous distributions capture the unobserved background factors that are left unexplained and are simply accepted as they are. This approach is illustrated below in terms of string diagrams where we shall, once again, exploit that we can simplify string diagrams after intervention, via the discard rules for $\overline{\top}$. In the party example, after the rewriting, the decomposition of the original probabilistic channels into a deterministic one and an exogenous distribution can be undone. We see this as a crucial fact, showing that the outcome of the counterfactual argument does not depend on the decomposition. This independence does not seem to have played a role in existing discussions.

9.3.1 Had Bob been at the party

The first example that we elaborate comes from Balke and Pearl (1994). It involves three people: Ann, Bob and Charly, which may or may not go to a party. This is captured by three two-element sets $A = \{a, a^\perp\}$, $B = \{b, b^\perp\}$ and $C = \{c, c^\perp\}$, where a indicates that Ann is going to the party, and $a^\perp$ that she is not going — and similarly for Bob and Charly. There is a fourth set $S = \{s, s^\perp\}$ whose elements indicate whether or not Bob and Charly scuffle. We assume a prior 60% chance that Ann will be at the party, as given by the distribution $\alpha = \frac{3}{5}|a\rangle + \frac{2}{5}|a^\perp\rangle$.

We have a string diagram as described on the left in Figure 9.3.1, with channels $f\colon A \twoheadrightarrow B$, $g\colon A \twoheadrightarrow C$ and $h\colon B \times C \twoheadrightarrow S$ given by:

$$\begin{aligned} f(a) &= \tfrac{9}{10}|b\rangle + \tfrac{1}{10}|b^\perp\rangle & h(b,c) &= \tfrac{19}{20}|s\rangle + \tfrac{1}{20}|s^\perp\rangle \\ f(a^\perp) &= \tfrac{3}{100}|b\rangle + \tfrac{97}{100}|b^\perp\rangle & h(b,c^\perp) &= 1|s^\perp\rangle \\ g(a) &= \tfrac{1}{10}|c\rangle + \tfrac{9}{10}|c^\perp\rangle & h(b^\perp,c) &= 1|s^\perp\rangle \\ g(a^\perp) &= \tfrac{19}{20}|c\rangle + \tfrac{1}{20}|c^\perp\rangle & h(b^\perp,c^\perp) &= \tfrac{1}{20}|s\rangle + \tfrac{19}{20}|s^\perp\rangle. \end{aligned}$$

The way to read these channels is:

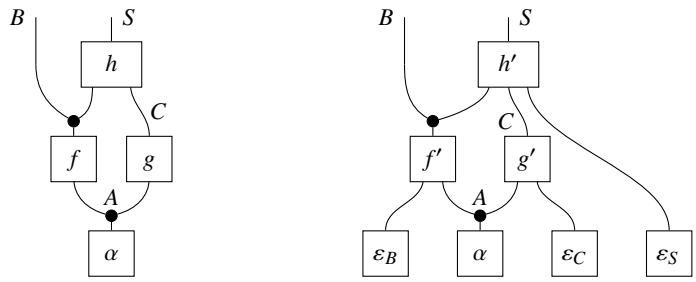


Figure 9.3 The party string diagram, with probabilistic channels f, g, h on the left, and on the right with these channels described as deterministic channels f', g', h' together with exogenous distributions $\varepsilon_B, \varepsilon_C, \varepsilon_S$

- Bob is very likely to go to the party if Ann is there, and is not likely to go when Ann is not there.
- For Charly it is the opposite: when Ann is there, it is unlikely that he goes to the party, and when Ann is not there, his presence is likely.
- When both Bob and Charly are both at the party it is likely that there will be scuffle; if they are both absent, there is still a small, 5% likelihood of scuffle.

An ordinary inference query in this situation is:

What is the likelihood of a scuffle if Bob is at the party?

We have seen many of these questions before, and so we leave the calculation of the outcome to the reader, see Exercise 9.3.2. We only mention here that the outcome is 11%, approximately.

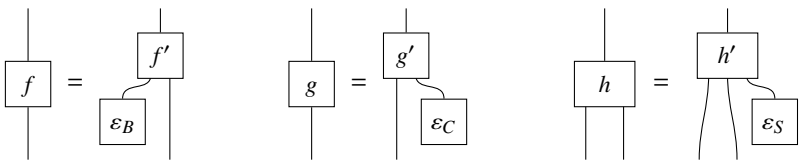
We now look at a counterfactual query:

If Bob is *not* at the party, what would be the likelihood of a scuffle had he been there?

The likelihood of a scuffle is now 79%, as will be computed below. This rise can be explained intuitively: if Bob is not a the party, then with high probability Ann wil be there neither, and so Charly can be expected to be present. But then, had Bob been there, a scuffle would be likely.

The formal analysis below involves a new technique, namely the isolation of the randomness of the channels f, g, h in so-called exogenous distributions (precisely as in Balke and Pearl (1994)). This allows us, as before to answer the counterfactual query, in a two-world situation, both factual and counterfactual, via a twinned string diagram, with shared randomness.

In the string diagram on the right in Figure 9.3.1 the channels f, g, h have been decomposed as deterministic channels (functions) f', g', h' together with exogenous distributions $\varepsilon_B, \varepsilon_C, \varepsilon_S$, as in:



(9.18)

Explicitly, this decomposition can be done via functions $f' \colon \mathbf{4} \times A \to B$, $g' \colon A \times \mathbf{4} \to C$ and

$h' \colon B \times C \times \mathbf{4} \to S$, where $\mathbf{4} = \{0, 1, 2, 3\}$ and:

$$\begin{array}{lll} f'(0,x) = b^\perp & f'(1,a) = b & f'(1,a^\perp) = b^\perp \\ f'(2,a) = b^\perp & f'(2,a^\perp) = b & f'(3,x) = b \\ g'(x,0) = c^\perp & g'(a,1) = c & g'(a^\perp,1) = c^\perp \\ g'(a,2) = c^\perp & g'(a^\perp,2) = c & g'(x,3) = c. \end{array}$$

And:

$$h'(x,y,1) = \begin{cases} s & \text{if } x = b, y = c \\ s^\perp & \text{otherwise} \end{cases} \qquad h'(x,y,2) = \begin{cases} s & \text{if } x = b, y = c \text{ or} \\ & \quad x = b^\perp, y = c^\perp \\ s^\perp & \text{otherwise.} \end{cases}$$

In addition, $h'(x,y,0) = s^\perp$ and $h'(x,y,3)$ is irrelevant, say uniform.

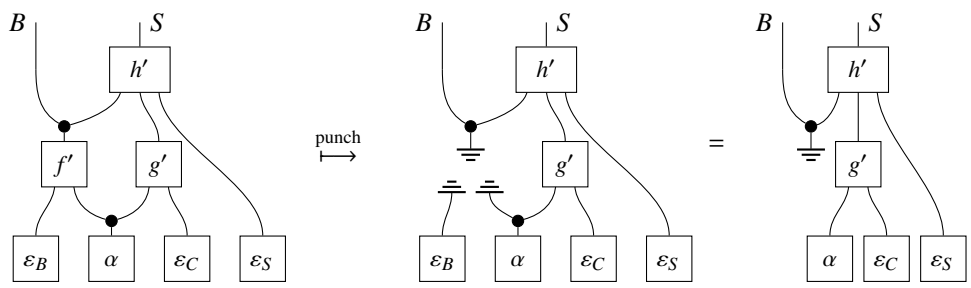
The decompositions (9.18) use three exogenous distributions $\varepsilon_B, \varepsilon_C, \varepsilon_S \in \mathcal{D}(4)$ for the set $\mathbf{4} = \{0, 1, 2, 3\}$. They are given by:

$$\varepsilon_B = \tfrac{7}{100}|0\rangle + \tfrac{9}{10}|1\rangle + \tfrac{3}{100}|2\rangle \qquad \varepsilon_C = \tfrac{1}{20}|0\rangle + \tfrac{17}{20}|2\rangle + \tfrac{1}{10}|3\rangle \qquad \varepsilon_S = \tfrac{1}{20}|0\rangle + \tfrac{9}{10}|1\rangle + \tfrac{1}{20}|2\rangle.$$

In Exercise 9.3.1 the goal is to verify that the channels f, g, h can indeed be decomposed as described in (9.18).

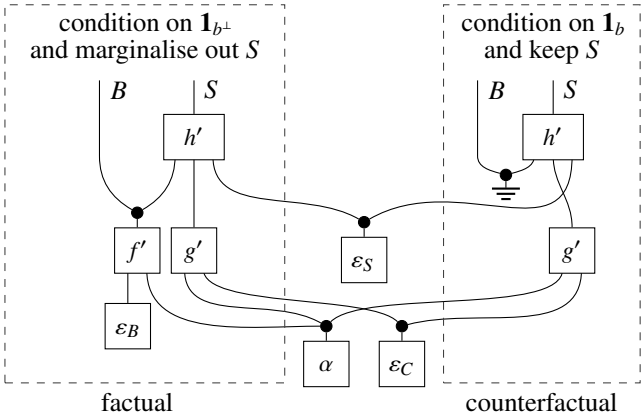
Our next step is to calculate the outcome of the above counterfactual query: given that Bob is not at the party, what would be the scuffle probability distribution, had he been there? The answer will be provided via a twin diagram approach, where a 'factual' and 'counterfactual' version of the diagram on the right in Figure 9.3.1 are combined, in such a way that the exogenous distributions are shared. In the factual part the information that Bob is not at the party is incorporated via conditioning with the point predicate $\mathbf{1}_{b^\perp}$. The counterfactual part the intervention happens to force the presence of Bob, via the point state $1|b\rangle$.

The counterfactual part of the twinned string diagram involves an intervention, consisting of a suitably placed punch, followed by a clean-up, as described in:



The factual part of the twinned diagram, described on the left below, involves conditioning with the factual assumption (Bob's absence) and marginalisation. The resulting twinned diagram then

looks as follows.



(9.19)

Notice how the exogenous distributions ε_C and ε_S are shared. The prior state α, describing Ann's presence, is also treated as a shared probabilistic resource. Notice also that the exogenous distribution ε_B is not shared, because it has disappeared from the counterfactual part during the clean-up after the punch.

In a next step we perform both conditionings and the discard (in the factual part), see Figure 9.4. The distribution $\tau \in \mathcal{D}(S)$ that is obtained there can be computed as:

$$\begin{aligned}\tau &= \Big(h \circ (id \otimes (g \circ f_\alpha^\dagger))\Big)_*(1|b\rangle \otimes 1|b^\perp\rangle) \\ &= \tfrac{35587}{44800}|s\rangle + \tfrac{9213}{44800}|s^\perp\rangle \;\approx\; 0.7944|s\rangle + 0.2056|s^\perp\rangle.\end{aligned}$$

Basically this same outcome is given in Balke and Pearl (1994), but as probability 0.79, with only two decimals.

An important observation is that in the final form in Figure 9.4 the decomposition of the channels f, g, h into deterministic channels with a distribution has disappeared. In particular, the outcome does not depend on the

9.3.2 Had the patient been treated

We include another illustration, taken from (Bareinboim et al., 2022, Ex. 27.2). It involves four distributions $flip(r) = r|1\rangle + (1-r)|0\rangle \in \mathcal{D}(\mathbf{2})$ that are used as exogenous states. Roughly following the original names, they are:

$$U_r = flip(\tfrac{1}{4}) \qquad U_z = flip(\tfrac{19}{20}) \qquad U_x = flip(\tfrac{9}{10}) \qquad U_y = flip(\tfrac{7}{10}).$$

There are three associated deterministic channels $f_z, f_x \colon \mathbf{2} \times \mathbf{2} \twoheadrightarrow \mathbf{2}$ and $f_z \colon \mathbf{2} \times \mathbf{2} \times \mathbf{2} \twoheadrightarrow \mathbf{2}$. Explicitly:

$$f_z(a,b) = \begin{cases} 1\big|1\rangle & \text{if } a = b = 1 \\ 1\big|0\rangle & \text{otherwise} \end{cases}$$

$$f_x(a,b) = \begin{cases} 1\big|1\rangle & \text{if } a = b = 1 \\ & \text{or } a = b = 0 \\ 1\big|0\rangle & \text{otherwise} \end{cases}$$

$$f_y(a,b,c) = \begin{cases} 1\big|1\rangle & \text{if } a = b = 1 \\ & \text{or } a = 0,\ b = c = 1 \\ & \text{or } a = b = c = 0 \\ 1\big|0\rangle & \text{otherwise.} \end{cases}$$

These distributions are composed as in the string diagram on the left below. Immediately, a

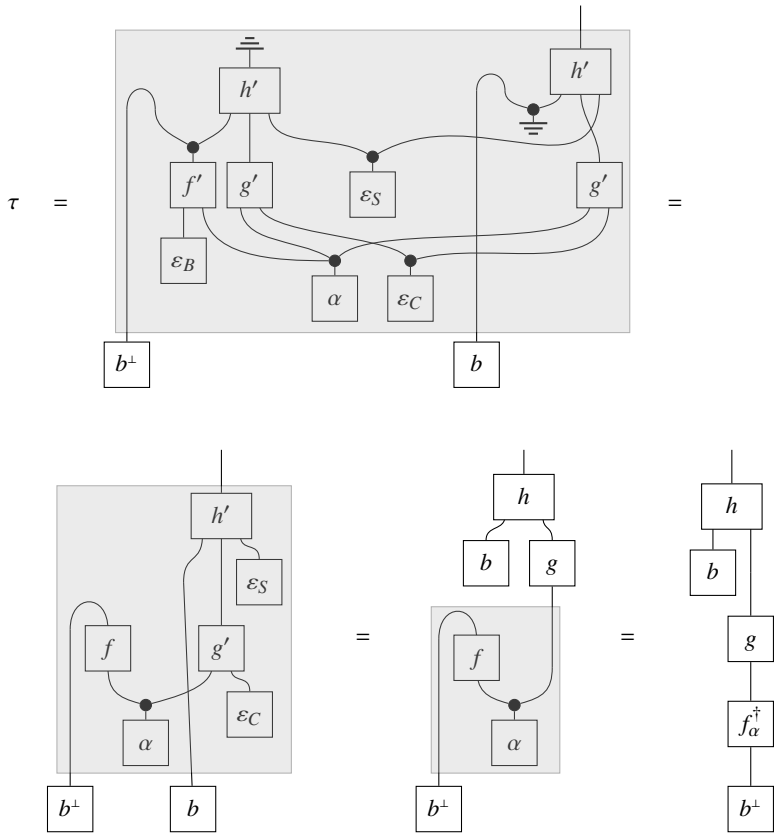
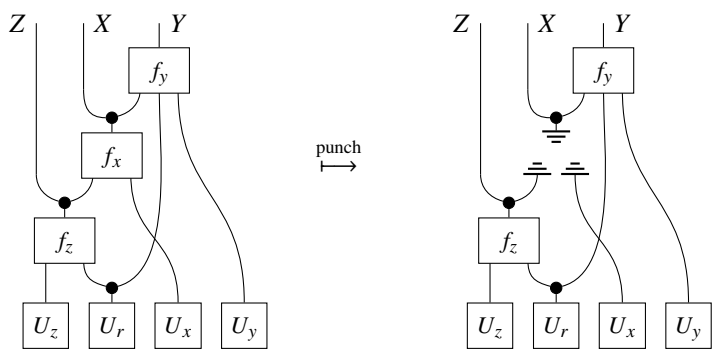


Figure 9.4 Calculation via string-diagrammatic rewriting of the counterfactual scuffle distribution τ that answers the question: given that Bob is not at the party, what would be the scuffle probability distribution, had he been there?

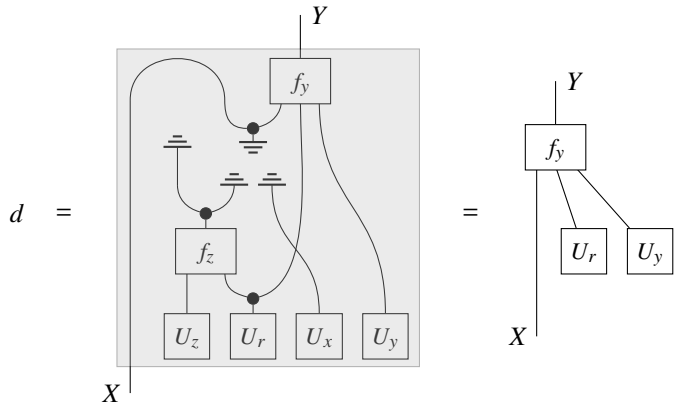
whole is punched to indicate where the intervention will happen. It is described in Bareinboim et al. (2022) as application of a treatment. Our description is as in Jacobs et al. (2025). The labels / types X, Y, Z are all equal to $\mathbf{2}$, but different names are used to distinghuish them.



(9.20)

The first challenge is to determine the direct influence from X (treatment) to Y (survival), without

confounding via U_r. We thus extract the do channel $c\colon X \twoheadrightarrow Y$ from the (punched) diagram on the right in (9.20). This gives:



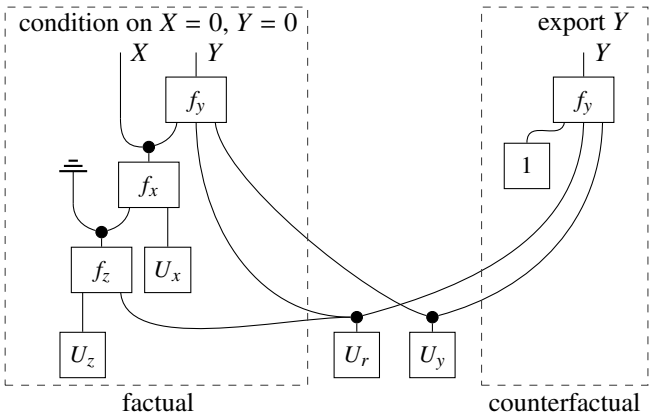
(9.21)

Then, as in Bareinboim et al. (2022),

$$d(1) = \tfrac{1}{4}\big|1\rangle + \tfrac{3}{4}\big|0\rangle \qquad d(0) = \tfrac{2}{5}\big|1\rangle + \tfrac{3}{5}\big|0\rangle.$$

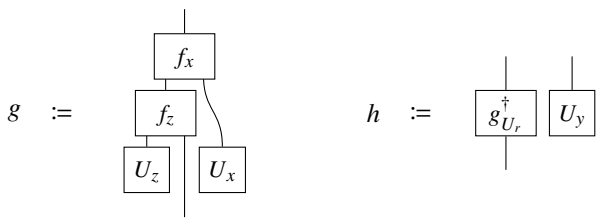
We see that the treatment does not really have effect.

We turn to counterfactuals. There is a patient who did not get treatment and died. Thus we can ask, had the patient been treated, would he or she have survived? We form a twinned diagram, coupling the factual and the counterfactual situation via the exogenous distributions, as in (9.19).



(9.22)

In order to increase readability we will use the following abbreviations.



The counterfactual reasoning that we are interested in can now be done via the channel c that is obtained by extraction in the upper left-corner in Figure 9.5. It is rewritten and simplified using the rules of Section 7.10. The last equation uses Exercise 7.10.5.

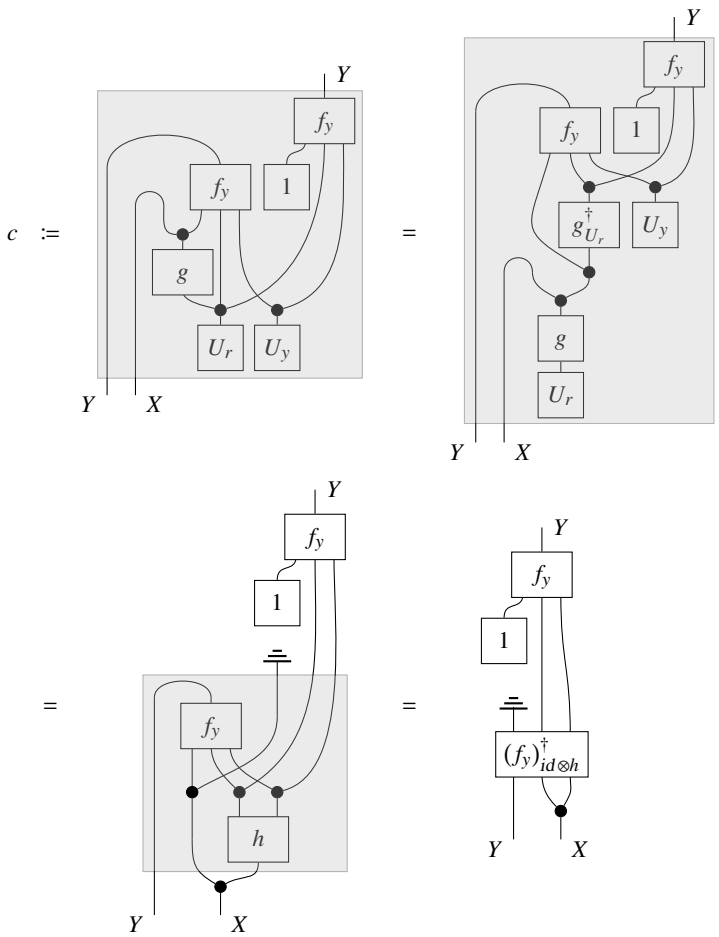


Figure 9.5 Extraction and simplification of the counterfactual channel c, obtained from the situation in Diagram (9.22).

Concretely, the counterfactual channel $c\colon X \times Y \twoheadrightarrow Y$, with $X = Y = \mathbf{2}$, is:

$$c(1,1) = 1\big|1\rangle \qquad c(1,0) = 1\big|0\rangle \qquad c(0,1) = \tfrac{49}{454}\big|1\rangle + \tfrac{405}{454}\big|0\rangle \qquad c(0,0) = \tfrac{1}{46}\big|1\rangle + \tfrac{45}{46}\big|0\rangle.$$

The latter outcome $\frac{1}{46} \approx 0.0217$ occurs also in Bareinboim et al. (2022). This confirms the earlier finding that the treatment has little positive effect. We conclude with an explicit counterfactual interpretation of the different outcomes of the channel c. The first two columns below describe the factual situation.

treatment	**survival**	**survival chance had treatment been given**
1	1	1
1	0	0
0	1	$\frac{49}{454}$
0	0	$\frac{1}{46}$

The third line is intriguing. It describes the situation where the patient survived without treatment, and where the survival probability had the treatment been given is only $\frac{49}{454} \approx 11\%$. Apparently, in this situation, the treatment has a negative effect: it lowers the chance of survival.

Exercises

9.3.1 Prove the equations in (9.18): the channels f, g, h can indeed be obtained from the deterministic channels f', g', h' with exogenous distributions, via:

$$f \;=\; f' \circ (\varepsilon_B \otimes id) \qquad g \;=\; g' \circ (id \otimes \varepsilon_C) \qquad h \;=\; h' \circ (id \otimes id \otimes \varepsilon_S).$$

9.3.2 Wite $\omega \in \mathcal{D}(B \times S)$ for the (interpretation of the) string diagram on the left in Figure 9.3.1.

1 Check that:

$$\omega \;=\; \Big((id \otimes h) \circ (\Delta \otimes id) \circ \langle f, g\rangle\Big)_*(\alpha).$$

2 The conditioning query that we asked ourselves at the beginning of this section involves the scuffle probability, given that Bob is at the party. It can be computed as the crossover inference on the left below. Show that it satisfies:

$$\omega|_{\mathbf{1}_b \otimes \mathbf{1}}[0, 1] \;=\; h_*\Big(\langle f, g\rangle_*(\alpha)\big|_{\mathbf{1}_b \otimes \mathbf{1}}\Big).$$

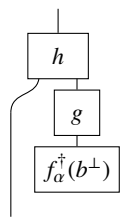
3 Show next that:

$$h_*\Big(\langle f, g\rangle_*(\alpha)\big|_{\mathbf{1}_b \otimes \mathbf{1}}\Big) \;=\; \tfrac{2071}{18400}|s\rangle + \tfrac{16329}{18400}|s^\perp\rangle \;\approx\; 0.1126|s\rangle + 0.8874|s^\perp\rangle.$$

This same query occurs in Balke and Pearl (1994), where the resulting scuffle probability is given as 0.11.

9.3.3 In this section we have elaborated the scuffle probability had Bob come to the party, given that Bob did not go, see Figure 9.4. In this exercise we provide a richer answer, via a 'counterfactual channel' $d\colon B \twoheadrightarrow S$, namely as:

$d \;:=$



1 Above we have shown that $d(b) = \frac{35587}{44800}|s\rangle + \frac{9213}{44800}|s^\perp\rangle$. Check that:

$$d(b^\perp) \;=\; \tfrac{367}{44800}|s\rangle + \tfrac{44433}{44800}|s^\perp\rangle.$$

2 Formulate the query to which this distribution forms the answer.

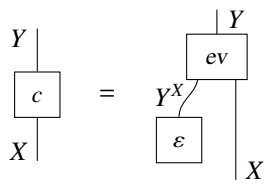
3 Compute the answer to the query: what is the scuffle probability had it been 50% certain that Bob came to the party, given that Bob did not go.

9.4 Counterfactual reasoning and randomness pushback

This section provides the background to what we have seen in the two examples in the previous section. This background consists of two parts. First, a systematic approach to exogenous distributions will be developed in terms of *randomness pushback* from Fritz (2020). Secondly, it is shown how counterfactual reasoning also involves the causal surgery procedure described in Subsection 9.1.2.

We start with a normal form theorem for channels, expressing that a channel can be decomposed into a deterministic channel together with a probabilistic distribution. In this way the randomness in the original channel is isolated in the distribution. This construction appears in Fritz (2020) where it is called randomness pushback. There is a canonical choice for the deterministic channel, namely an 'evaluation' function. There are many possible choices for the associated distribution. We shall call this distribution 'exogenous', as is common in the area of counterfactual reasoning. It represents randomness from the background that is factored into the situation at hand via the deterministic channel. As we have seen in twinned networks consisting of a factual and a counterfactual part, the deterministic channels are copied and exist (in the same form) on both sides, whereas the exogenous distributions are shared between the two parts, via a copier.

Proposition 9.4.1 *Let $c\colon X \twoheadrightarrow Y$ be a channel, whose domain X is a non-empty finite set. Then there is an 'exogenous' distribution $\varepsilon \in \mathcal{D}(Y^X)$ such that c can be written as:*



(9.23)

where:

- $\mathrm{ev}\colon Y^X \times X \to Y$ *is the evaluation function* $\mathrm{ev}(h, x) = h(x)$*; it is used as deterministic channel;*
- ε *is the distribution:*

$$\varepsilon := \sum_{h \in Y^X} \left(\prod_{x \in X} c(x)(h(x)) \right) |h\rangle. \tag{9.24}$$

One can restrict this summation to $h \in S^X$, for the finite set $S := \bigcup_x supp(c(x)) \subseteq Y$, since if $h(x) \notin S$, for some $x \in X$, then the whole product $\prod$ is zero.

The right-hand side in the above equation (9.23) describes the elements what are called 'structured causal models', see *e.g.* Pearl (2009).

Proof. We first check that ε is a distribution on Y^X. This yields a distribution since a function $h\colon X \to Y$ can be identified with an N-tuple $\langle y_1, \ldots, y_N \rangle$ for $N := |X| > 0$.

$$\sum_{h \in Y^X} \varepsilon(h) = \sum_{h \in Y^X} \prod_{x \in X} c(x)(h(x)) = \prod_{x \in X} \sum_{y \in supp(c(x))} c(x)(y) = \prod_{x \in X} 1 = 1.$$

Equation (9.23) holds since for each $x \in X$,

$$\begin{aligned}
\mathrm{ev}_*(\varepsilon \otimes 1|x\rangle) &= \sum_{y \in Y} \mathcal{D}(\mathrm{ev})\big(\varepsilon \otimes 1|x\rangle\big)(y)\,\big|y\rangle \\
&= \sum_{y \in Y} \sum_{h \in Y^X,\, \mathrm{ev}(h,x)=y} \varepsilon(h)\,\big|y\rangle \\
&= \sum_{y \in Y} \sum_{h \in Y^X,\, h(x)=y} \prod_{z \in X} c(z)(h(z))\,\big|y\rangle \\
&= \sum_{y \in Y} c(x)(y) \cdot \sum_{h \in Y^{X \setminus \{x\}}} \prod_{z \in X \setminus \{x\}} c(z)(h(z))\,\big|y\rangle \;=\; \sum_{y \in Y} c(x)(y)\,\big|y\rangle \;=\; c(x). \qquad \square
\end{aligned}$$

In concrete pushback situations, where the domain $X = \{x_1, \ldots, x_n\}$ of the channel $c\colon X \twoheadrightarrow Y$ has only a few elements, listed in a particular order, we shall often write the function space Y^X as an n-ary product $Y \times \cdots \times Y$. Evaluation at $x_i \in X$ then corresponds application of the i-th projection π_i. The exogenous distribution ε defined in the above proof can then be written as n-ary tensor product:

$$\varepsilon \;=\; c(x_1) \otimes \cdots \otimes c(x_n) \;\in\; \mathcal{D}(Y \times \cdots \times Y).$$

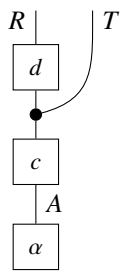
The channel's distribution at x_i can now be obtained as the i-th marginal:

$$c(x_i) \;=\; \mathrm{ev}_*(\varepsilon \otimes 1|x_i\rangle) \;=\; (\pi_i)_*(\varepsilon) \;=\; \varepsilon[0, \ldots, 0, 1, 0, \ldots, 0].$$

We thus see that an exogenous distribution for a channel c is in fact an n-ary coupling — see Section 3.1 — which marginalises to each of the distributions $c(x_i)$. The above version $\varepsilon = c(x_1) \otimes \cdots \otimes c(x_n)$ is the obvious such coupling, without entwinedness. But there may be many other couplings, as we shall see below.

In the remainder of this section we build on an illustration from the literature, namely from (Koller and Friedman, 2009, Ex. 21.23), where we introduce the exogenous distributions ourselves, following the approach described above. It turns out that in this illustration the counterfactual outcome depends on the choice of exogenous distribution. This differs from the examples in the previous section. One could argue that the dependence on the choice of exogenous distribution is not 'good' and indicates that the situation does not allow (proper, intrinsic) counterfactual reasoning.

Our illustration uses the set up from (Koller and Friedman, 2009, Ex. 21.23), but with different (simpler) numbers. The example involves a medical test situation where either a medicine is assigned to a patient or a placebo — via a sample space $A = \{m, p\}$ — where the patient actually takes this medicine / placebo or not — via a set $T = \{t, t^\perp\}$ — and where the patient recovers or not — via $R = \{r, r^\perp\}$. We have a string diagram as on the left below, with interpretion on the right.



$$\begin{aligned}
\alpha &= \tfrac{1}{2}|m\rangle + \tfrac{1}{2}|p\rangle \\
c(m) &= \tfrac{7}{10}|t\rangle + \tfrac{3}{10}|t^\perp\rangle \\
c(p) &= \tfrac{1}{10}|t\rangle + \tfrac{9}{10}|t^\perp\rangle \\
d(t) &= \tfrac{1}{2}|r\rangle + \tfrac{1}{2}|r^\perp\rangle \\
d(t^\perp) &= \tfrac{1}{4}|r\rangle + \tfrac{3}{4}|r^\perp\rangle.
\end{aligned} \tag{9.25}$$

We first concentrate on exogenous distributions for the channels c and d. Since their domains have two elements only, these exogenous distributions will be distributions on T^2 and R^2. Following the description of (9.24) yields:

$$\begin{aligned}\varepsilon_c &= c(a) \otimes c(a^\perp) = \tfrac{7}{100}\big|t,t\rangle + \tfrac{63}{100}\big|t,t^\perp\rangle + \tfrac{3}{100}\big|t^\perp,t\rangle + \tfrac{27}{100}\big|t^\perp,t^\perp\rangle \\ \varepsilon_d &= d(t) \otimes d(t^\perp) = \tfrac{1}{8}\big|r,r\rangle + \tfrac{3}{8}\big|r,r^\perp\rangle + \tfrac{1}{8}\big|r^\perp,r\rangle + \tfrac{3}{8}\big|r^\perp,r^\perp\rangle.\end{aligned}$$

The associated evaluation maps $ev_T \colon T^2 \times A \to T$ and $ev_R \colon R^2 \times T \to R$ are:

$$ev_T(t_1,t_2,x) = \begin{cases} t_1 & \text{if } x = m \\ t_2 & \text{if } x = p \end{cases} \qquad ev_R(r_1,r_2,y) = \begin{cases} r_1 & \text{if } y = t \\ r_2 & \text{if } y = t^\perp. \end{cases}$$

Then, for instance, one gets as first marginal:

$$\begin{aligned}\big(ev_T \circ (\varepsilon_T \otimes id)\big)(m) &= ev_{T*}(\varepsilon_T \otimes 1|m\rangle) = \sum_{t_1,t_2 \in T} \varepsilon_T(t_1,t_2)\,\big|ev_T(t_1,t_2,a)\rangle \\ &= \sum_{t_1,t_2 \in T} \varepsilon_T(t_1,t_2)\,\big|t_1\rangle \\ &= (\tfrac{7}{100} + \tfrac{63}{100})\big|t\rangle + (\tfrac{3}{100} + \tfrac{27}{100})\big|t^\perp\rangle = c(m).\end{aligned}$$

Similarly for argument $p \in A$. And in the same way one gets $ev_R \circ (\varepsilon_R \otimes id) = d$.

We explicitly mention two other exogenous distributions $\theta_T \in \mathcal{D}(T \times T)$ and $\theta_R \in \mathcal{D}(R \times R)$, with $ev_T \circ (\theta_T \otimes id) = c$ and $ev_R \circ (\theta_R \otimes id) = d$. This works for:

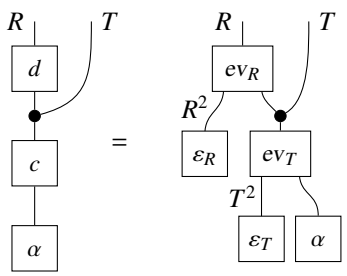
$$\theta_T = \tfrac{7}{10}\big|t,t^\perp\rangle + \tfrac{1}{10}\big|t^\perp,t\rangle + \tfrac{1}{5}\big|t^\perp,t^\perp\rangle \qquad \theta_R = \tfrac{1}{6}\big|r,r\rangle + \tfrac{1}{3}\big|r,r^\perp\rangle + \tfrac{1}{12}\big|r^\perp,r\rangle + \tfrac{5}{12}\big|r^\perp,r^\perp\rangle.$$

Via their first marginals we also recover the channels c and d.

We now come to the counterfactual query, after Koller and Friedman (2009):

Assume that we have a patient who, in the real world, was given the medicine (m), but did not take it ($t^\perp$) and did not recover ($r^\perp$). We are interested in the probability that he would have recovered had he taken the medicine.

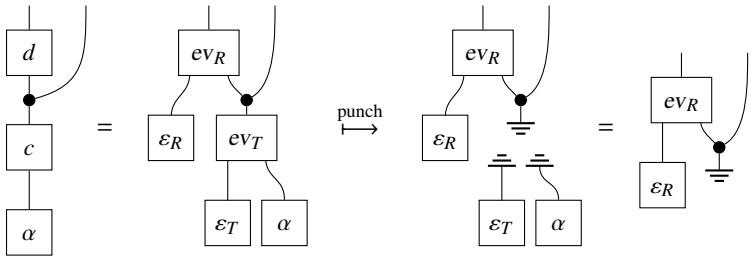
We follow the randomness pushback from Koller and Friedman (2009) which can be described in terms of a *joint* exogenous distribution $\varepsilon \in \mathcal{D}(R \times R \times T \times T)$, as in the following picture.



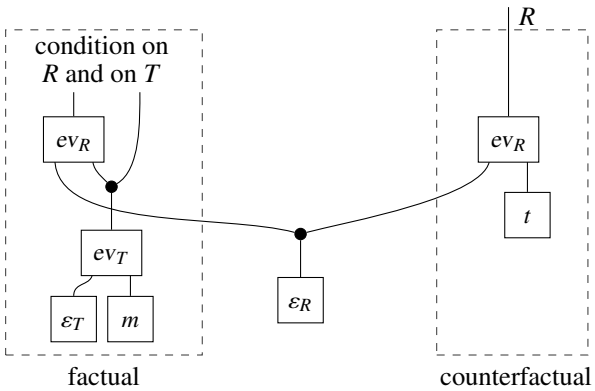
(9.26)

The counterfactual query involves the assumption "had he taken the medicine". This is modeled by an intervention, which is handled via causal diagram surgery, below the T label. This is

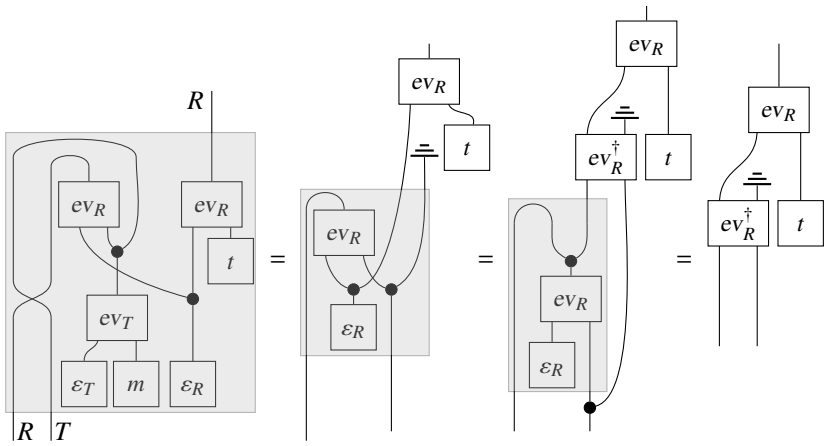
expressed by the punch, with a subsequent diagram simplification.



We are now ready to form a twin-diagram, with the factual world on the left, involving the assumptions from the real world, and the counterfactual world on the right, mutilated by diagram surgery (punching), where the medicine is taken — as represented by the point distribution box t. The two worlds are connected via the shared exogenous distribution ε.



We apply disintegration to this diagram and simplify to get a counterfactual channel $R \times T \rightsquigarrow$R, of the form:



The prior (channel) of the dagger $ev_R^\dagger$ has been left implicit, for readability. This prior is $\varepsilon_R \otimes id : T \rightsquigarrow R \times R \times T$. As a result, the dagger $ev_R^\dagger = (ev_R)^\dagger_{\varepsilon_R \otimes id} : R \times T \rightsquigarrow R \times R \times T$ depends on the choice of exogenous distribution. But we can also take θ_R as exogenous distribution and use it instead of ε_R in the dagger $(ev_R)^\dagger_{\theta_R \otimes id}$. These two choices yield different counterfactual channels

$R \times T \rightarrowtail R$, on the right-hand-side in the above string diagram equations.

With dagger prior $\varepsilon_R \otimes id$	**With dagger prior** $\theta_R \otimes id$
$(r,t) \longmapsto 1\vert r\rangle$	$(r,t) \longmapsto 1\vert r\rangle$
$(r,t^\perp) \longmapsto \frac{1}{2}\vert r\rangle + \frac{1}{2}\vert r^\perp\rangle$	$(r,t^\perp) \longmapsto \frac{2}{3}\vert r\rangle + \frac{1}{3}\vert r^\perp\rangle$
$(r^\perp,t) \longmapsto 1\vert r^\perp\rangle$	$(r^\perp,t) \longmapsto 1\vert r^\perp\rangle$
$(r^\perp,t^\perp) \longmapsto \frac{1}{2}\vert r\rangle + \frac{1}{2}\vert r^\perp\rangle$	$(r^\perp,t^\perp) \longmapsto \frac{4}{9}\vert r\rangle + \frac{5}{9}\vert r^\perp\rangle.$

In this section we have taken a more systematic look at the decomposition of a probabilistic channel into deterministic channel plus an exogenous distribution where all probability is located. The 'probability pushback' proposition 9.4.1 describes a canonical decomposition — among the many possible ones — using evaluation as deterministic channel. We have used these evaluation functions in a counterfactual example, adapted from Koller and Friedman (2009).

In the two examples in the previous section we have seen that the answer to the counterfactual queries did not depend on the specific channel decomposition that was used. We regard this independence as a well-definedness property of answers to counterfactual queries. The example in this section fails this test: choosing a different exogenous distribution — for which Equation (9.26) must of course still hold — yields a different counterfactual channel as outcome. This does not mean that there is no proper answer to the medical counterfactual query that we have considered here, but it does show, we think, that the graphical decomposition in (9.26) is not the right one for counterfactual reasoning.

Exercises

9.4.1 Consider the Medicine-Blood channel $b\colon M \rightarrowtail B$ from Example 6.2.6 for sets $M = \{0,1,2\}$ and $B = \{H,L\}$. In this exercise we elaborate randomness pushback for this channel.

1 Consider the set B^M containing the following $2^3 = 8$ functions $h_1,\ldots,h_8$.

$$\begin{cases} h_1(0) = H \\ h_1(1) = H \\ h_1(2) = H \end{cases} \quad \begin{cases} h_2(0) = H \\ h_2(1) = H \\ h_2(2) = L \end{cases} \quad \begin{cases} h_3(0) = H \\ h_3(1) = L \\ h_3(2) = H \end{cases} \quad \begin{cases} h_4(0) = H \\ h_4(1) = L \\ h_4(2) = L \end{cases}$$

$$\begin{cases} h_5(0) = L \\ h_5(1) = H \\ h_5(2) = H \end{cases} \quad \begin{cases} h_6(0) = L \\ h_6(1) = H \\ h_6(2) = L \end{cases} \quad \begin{cases} h_7(0) = L \\ h_7(1) = L \\ h_7(2) = H \end{cases} \quad \begin{cases} h_8(0) = L \\ h_8(1) = L \\ h_8(2) = L \end{cases}$$

Show that the exogenous distribution $\varepsilon \in \mathcal{D}(B^M)$ defined in the proof of Proposition 9.4.1 takes the form:

$$\tfrac{35}{108}\vert h_1\rangle + \tfrac{7}{36}\vert h_2\rangle + \tfrac{5}{54}\vert h_3\rangle + \tfrac{1}{18}\vert h_4\rangle + \tfrac{35}{216}\vert h_5\rangle + \tfrac{7}{72}\vert h_6\rangle + \tfrac{5}{108}\vert h_7\rangle + \tfrac{1}{36}\vert h_8\rangle.$$

2 The evaluation function $ev\colon B^M \times M \to B$ now gives an equation $b(x) = ev_*(\varepsilon \otimes 1\vert x\rangle)$. For instance:

$$\begin{aligned} ev_*(\varepsilon \otimes 1\vert 2\rangle)(H) &= \sum_{h \in B^M,\, h(2)=H} \varepsilon(h) = \varepsilon(h_1) + \varepsilon(h_3) + \varepsilon(h_5) + \varepsilon(h_7) \\ &= \tfrac{35}{108} + \tfrac{5}{54} + \tfrac{35}{216} + \tfrac{5}{108} = \tfrac{135}{216} = \tfrac{5}{8} = b(2)(H). \end{aligned}$$

Check yourself:

$$ev_*(\varepsilon \otimes 1|1\rangle)(L) \;=\; b(1)(L).$$

9.4.2 Write a channel $c\colon X \rightsquigarrow Y$ as $c = ev \circ (\varepsilon \otimes id)$, like in Proposition 9.4.1. The aim of this Exercise is to express the update rules of Pearl and Jeffrey from Section 7.9 in terms of the deterministic evaluation channel ev. Let $\sigma \in \mathcal{D}(X)$ be a prior distribution.

1 Show that for a predicate q on Y,

$$\sigma|_{c^*(q)} \;=\; \Big((\varepsilon \otimes \sigma)\big|_{ev^*(q)}\Big)[0,1].$$

2 Similarly, for a distribution $\tau \in \mathcal{D}(Y)$, prove that:

$$\Big(c_\sigma^\dagger\Big)_*(\tau) \;=\; \Big(ev_{\varepsilon\otimes\sigma}^\dagger\Big)_*(\tau)[0,1].$$

10

Probabilistic Automata

An automaton is a mathematical model of computation, consisting of positions (often called states) with possibilities to move from one position to one or more other positions. Such a move is often called a transition. Automata have been used since the 1930s to capture the dynamics of computation, and are used nowadays for instance in compiler design and in model checking. These computations can be deterministic or non-deterministic, with finitely many positions (corresponding to finite memory) or with infinitely many positions (like in Turing's tape models), see *e.g.* Arbib (1969); Eilenberg (1974) for an overview. Here we are interested in probabilistic automata, introduced in Rabin (1963). They typically involve a distribution in each position that gives the propabilities of moves to next positions. The (hidden) Markov models from Section 6.7 are examples of probabilistic automata. Important applications of such automata occur in system verification, via probabilistic model checking see Baier and Katoen (2008), and in reinforcement learning, via Markov decision processes (see below). In our approach to probabilistic automata we shall make use of monads and coalgebras. Our description will thus involve some basic categorical structures. Monads have already been described in Section 1.12, but coalgebras will be introduced below. This chapter's aim is to uncover and exploit the categorical structure that is underlying automata theory, but not to give an exhaustive description of the field.

There is a somewhat unfortunate clash of terminology that we have to deal with. In this book we have been using the term 'state' as synonym for probability distribution — as is common in for instance quantum probability. In automata theory the word 'state' is used for the source and target of transitions. Thus, in an automaton one can move from on state to another. We choose to use the word 'position' for what is called 'state' in automata theory and continue to use 'state' for distribution.

Recall from Section 6.7 that a Markov models can be described as a channel $c\colon X \rightarrowtail X$, that is, as a function $c\colon X \to \mathcal{D}(X)$. In the current setting we shall call X the set of positions. The channel gives for each position $x \in X$ a probability distribution $c(x) \in \mathcal{D}(X)$, say of the form $r_1|x_1\rangle + \cdots + r_n|x_n\rangle$ over successor positions x_i. It tells that there is a transition $x \to x_i$ with probability r_i. This chapter begins by generalising these Markov models to 'Markov models with outputs'. They are transition functions of the form $X \to \mathcal{D}(X + A)$, where A is a set of outputs, and where $+$ is disjoint union (coproduct). If such an output $a \in A$ is reached via a transition, no further transitions are possible. Interestingly, the mapping $X \mapsto \mathcal{D}(X+A)$, for a fixed set A, is not only a functor, but also a monad, see Section 1.12. This will be exploited to iterate transitions, via Kleisli composition of the monad. This may lead to a distribution over outputs in the limit. We illustrate this via the classical example of how a coin can be used to simulate a dice.

Sections 10.2 and 10.3 introduce two more variations of Markov models. A Markov models with rewards is of the form $X \to \mathcal{D}(X) \times \mathbb{R}$. We use Banach's fixed point theory to compute the

'return' as ultimate reward. A Markov descision process has the shape $X \to \mathcal{D}(X)^A \times \mathbb{R}$, and is a generalisation, involving a set of actions A. For such automata one is interested in computing an (optimal) policy, that tells in each position which action gives the highest (ultimate, successive) reward.

In Section 10.4 we switch gears and notice that a uniform description of automata can be given in terms of coalgebras. The latter are functions of the form $X \to F(X)$, where F is a functor **Sets** $\to$ **Sets**. This functor then captures the type of computation involved. The set X contains the position, and the map $X \to F(X)$ is the transition function. We identify an inductively defined class of probabilistic polynomial functors that we shall work with. They allow us to capture all previously described automata as coalgebras. Such coalgebras are more widely used to describe dynamical systems, including (non-)deterministic automata, see *e.g.* Rutten (2000); Jacobs (2016a); Adámek et al. (2025), but here we concentrate on their usage for probabilistic automata.

The last two sections 10.5 and 10.6 are about invariants and behavioural metrics, combining coalgebras with predicates and relations. Standardly, an invariant is a predicate that, once it holds for a position, continues to hold for all successor positions. Such invariants are important in reasoning, because they allow us to express for instance that a property "a variable v is non-zero" as an invariant, so that it is safe to divide by v in future positions. Similarly, a bisimulation is a relation that is closed under transitions. Bisimulations are used to express indistinguishability of positions. Standardly, invariants and bisimulations are sharp, giving only yes/no, or 1/0 outcomes. Our approach is 'quantitative' and uses fuzzy, $[0, 1]$-valued predicates and relations — in line with the rest of the book. In this quantitative approach one concentrates on special bisimulations, namely (pseudometric) distances. The idea is that two positions have distance zero when they are indistinguishable. An example is elaborated to give an impression. The literature on probabilistic automata is extensive, especially about their use in verification. This chapter provides only a first introduction.

Remark 10.0.1 A sequence $(x_n)_{n\in\mathbb{N}}$ of numbers $x_n \in \mathbb{R}$ is called geometric when $x_{n+1} = a \cdot x_n$ for a fixed number a. Then $x_n = x_0 \cdot a^n$. We are especially interested in the case where $a \in [0, 1)$, since then, by Theorem 1.7.4 (2):

$$\sum_{n\geq 0} a^n = \lim_{n\to\infty} a^0 + a^1 + \cdots + a^n = \frac{1}{1-a}. \tag{10.1}$$

10.1 Markov models with output

In this first section we elaborate several examples to set the scene. Without a general definition of a probabilistic automaton yet, we claim that a characteristic aspect is that there are 'positions' and probabilistic transitions between them, just like we have seen for (hidden) Markov models in Section 6.7. The exact shape of such transitions will occupy us later on in this chapter.

We recall that Markov models involve a transition function $X \to \mathcal{D}(X)$ from a space X to itself. In this section we add possible outputs in a set A and consider transition functions of the form $X \to \mathcal{D}(X + A)$, where $+$ is disjoint union (coproduct). Thus, from a position $x \in X$ there are possible transitions to a successor position $x' \in X$ and also to an output value $a \in A$. We call such a map $X \to \mathcal{D}(X + A)$ a *Markov model with output*, abbreviated as MMO.

In the literature the set A of outputs is often treated as a subset of X, namely as the subset of

final positions. The idea is that there are no outgoing transitions for these final positions, except to itself, with probability one. Here we like to deal with such final positions via a separate set A, unrelated to the set of positions. The type $X \to \mathcal{D}(X + A)$ of the transition function guarantees that there are no outgoing transitions for elements of A. The difference between ordinary Markov models and Markov models with output is not so big, see also Exercise 10.1.6.

The $+$ that we use for disjoint union is an instance of a general categorical operation called coproduct. It is dual to a product, but less well-known. We first introduce these coproducts, before using them in several examples of Makov models with output.

10.1.1 Coproducts and the exception monad

Products of sets, together with their projection functions $X \xleftarrow{\pi_1} X \times Y \xrightarrow{\pi_2} Y$ have been used extensively. Products $\times$ form a common construction that is well-known. Less well-known is a dual construction, called a *coproduct*, written as $+$. It can be formulated in an arbitrary category, but for our purposes coproducts $X + Y$ of sets X, Y suffice. They are disjoint unions and come with coprojection functions κ_i, of the form:

$$X \xrightarrow{\kappa_1} X + Y \xleftarrow{\kappa_2} Y \tag{10.2}$$

A concrete description of this coproduct set is the disjoint union:

$$X + Y := \{(x, 1) \mid x \in X\} \cup \{(y, 2) \mid y \in Y\}.$$

On the right-hand-side the elements from X and Y are 'tagged' with numbers 1 and 2 in order to ensure disjointness of the union $\cup$. Which tags are used is irrelevant, as long as they are different. There are two obvious *coprojection* functions $\kappa_1 \colon X \to X + Y$ and $\kappa_2 \colon Y \to X + Y$, as in (10.2), given by $\kappa_1(x) := (x, 1)$ and $\kappa_2(y) := (y, 2)$.

There is also an associated *cotuple* construction, which we have not seen so far. Suppose we have two functions $f \colon X \to Z$ and $g \colon Y \to Z$, then there is a unique function $X + Y \to Z$, written as cotuple $[f, g]$, with $[f, g] \circ \kappa_1 = f$ and $[f, g] \circ \kappa_2 = g$. This cotuple is (thus) defined by case distinction: for $w \in X + Y$,

$$[f, g](w) := \begin{cases} f(x) & \text{if } w = (x, 1) \\ g(y) & \text{if } w = (y, 2). \end{cases}$$

We also formulate this notion in general, in a category.

Definition 10.1.1 Let $\mathbb{C}$ be an arbitrary category, and let X, Y be objects in $\mathbb{C}$. The *coproduct* of X, Y, if it exists, is given by a new object, written as $X + Y$ with two arrows $\kappa_1 \colon X \to X + Y$ and $\kappa_2 \colon Y \to X + Y$ in $\mathbb{C}$ satisfying the following property: for each object Z with arrows $f \colon X \to Z$ and $g \colon Y \to Z$ there is a unique *cotuple* map $[f, g] \colon X + Y \to Z$ in $\mathbb{C}$ satisfying $[f, g] \circ \kappa_1 = f$ and $[f, g] \circ \kappa_2 = g$.

In a diagram, this looks as follows, where the cotuple arrow is dashed in order to express its uniqueness.

$$\begin{tikzcd} X \arrow[r, "\kappa_1"] \arrow[dr, "f"'] & X + Y \arrow[d, dashed, "{[f,g]}"] & Y \arrow[l, "\kappa_2"'] \arrow[dl, "g"] \\ & Z & \end{tikzcd}$$

For two maps $f\colon X \to A$ and $g\colon Y \to B$ we define:

$$f + g \coloneqq [\kappa_1 \circ f, \kappa_2 \circ g]\colon X + Y \longrightarrow A + B. \tag{10.3}$$

We say that $\mathbb{C}$ *has* coproducts — or is a category *with* coproducts — if such a coproduct exists in $\mathbb{C}$ for each pair of objects. The definition (10.3) then turns these coproducts into a functor $+\colon \mathbb{C} \times \mathbb{C} \to \mathbb{C}$.

Coproducts can be used to form a monad, see Subsection 1.12.4, which is often called the *exception monad*. It captures the idea that if a computation hits an exception, it halts, displaying the exception (object). Exception monad exist on categories with coproducts. We formulate the result in full generality, but will use it especially for the category **Sets** of sets and functions.

Lemma 10.1.2 *Let $\mathbb{C}$ be a category with coproducts $+$ and let $A \in \mathbb{C}$ be a fixed object — to be used for exceptions. Consider the functor $(-) + A\colon \mathbb{C} \to \mathbb{C}$, sending an object $X \in \mathbb{C}$ to $X + A$ and sending a morphism $f\colon X \to Y$ in $\mathbb{C}$ to $(f + id) = [\kappa_1 \circ f, \kappa_2]\colon X + A \to Y + A$.*

Then, this functor $(-) + A$ is a monad on $\mathbb{C}$, with unit and flatten maps:

$$X \xrightarrow{unit \,\coloneqq\, \kappa_1} X + A \qquad\qquad (X + A) + A \xrightarrow{flat \,\coloneqq\, [id, \kappa_2]} X + A$$

Proof. We briefly check the monad equations (1.60), using Exercise 10.1.7.

$$\begin{aligned}
flat \circ unit &= [id, \kappa_2] \circ \kappa_1 = id \\
flat \circ (unit + id) &\overset{(10.3)}{=} [id, \kappa_2] \circ [\kappa_1 \circ \kappa_1, \kappa_2] \\
&= \big[\, [id, \kappa_2] \circ \kappa_1 \circ \kappa_1,\ [id, \kappa_2] \circ \kappa_2 \,\big] = \big[\, \kappa_1, \kappa_2 \,\big] = id. \\
flat \circ (flat + id) &\overset{(10.3)}{=} [id, \kappa_2] \circ \big[\, \kappa_1 \circ [id, \kappa_2],\ \kappa_2 \,\big] \\
&= \big[\, [id, \kappa_2],\ \kappa_2 \,\big] = [id, \kappa_2] \circ [id, \kappa_2] = flat \circ flat.
\end{aligned}$$

❑

In a next step we combine the exception monad $(-) + A$ and the distribution monad $\mathcal{D}$, now on the category **Sets**. This yields another monad. In general, monads are not closed under composition. They do compose when there is a distributive law, see Exercise 3.9.4. The law involved in this case will be made explicit in Exercise 10.1.11. Here we concentrate on the (composite) monad.

Lemma 10.1.3 *Fix a set A and consider the functor $X \mapsto \mathcal{D}(X + A)$. This is a monad on* **Sets** *with $unit\colon X \to \mathcal{D}(X + A)$ and $flat\colon \mathcal{D}\big(\mathcal{D}((X + A) + A\big) \to \mathcal{D}(X + A)$ given as follows.*

$$\begin{aligned}
unit(x) &\coloneqq 1\big|\kappa_1(x)\big\rangle \\
flat(\Omega) &\coloneqq \sum_{u \in X+A} \left(\sum_{\omega \in \mathcal{D}(X+A)} \Omega(\kappa_1(\omega)) \cdot \omega(u) \right) \big|u\big\rangle + \sum_{a \in A} \Omega(\kappa_2(a)) \big|\kappa_2(a)\big\rangle.
\end{aligned}$$

The associated Kleisli composition $g \circ f\colon X \to \mathcal{D}(Z + A)$ of $f\colon X \to \mathcal{D}(Y + A)$ and $g\colon Y \to \mathcal{D}(Z + A)$ is given by:

$$(g \circ f)(x) \coloneqq \sum_{u \in Z+A} \left(\sum_{y \in Y} f(x)(\kappa_1(y)) \cdot g(y)(u) \right) \big|u\big\rangle + \sum_{a \in A} f(x)(\kappa_2(a)) \big|\kappa_2(a)\big\rangle. \tag{10.4}$$

Proof. One can verify the monad equation (1.60) directly, but one can also see this result as an instance of Exercise 10.1.11. ❑

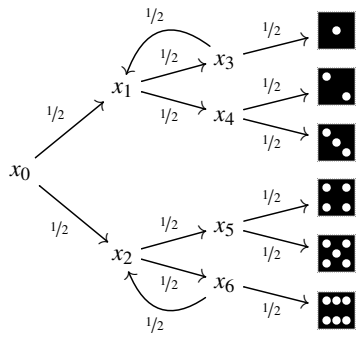


Figure 10.1 The transition system for simulating a dice via coin flips.

The coprojections κ_i in formulations like (10.4) do not add much, since they can be deduced from the types of the elements (like $y \in Y$ and $a \in A$), so we often ommit these κ_i for convenience.

In the remainder of this section we are interested in transition functions $f\colon X \to \mathcal{D}(X + A)$ with an initial position x_0. We will then look at subsequent distributions on $X + A$, of the form:

$$\begin{aligned} f^0(x_0) &= \mathit{unit}(x_0) \;=\; 1|x_0\rangle \\ f^1(x_0) &= f(x_0) \\ f^2(x_0) &= (f \circ f)(x_0) \\ f^3(x_0) &= (f \circ f \circ f)(x_0), \qquad \textit{etc.} \end{aligned}$$

The description of these iterations as Kleisli compositions for a monad gives us computation rules that we can evaluate, either by hand, or via a computer program.

10.1.2 Simulating a dice with coins

We elaborate a standard example from Knuth and Yao (1976). First, we describe it as a probabilistic transition system, see Figure 10.1. This example uses positions $X = \{x_0, x_1, \ldots, x_6\}$ and observations $A = \mathit{pips} = \{1, 2, 3, 4, 5, 6\}$ with transition function $f\colon X \to \mathcal{D}(X + A)$ given by:

$$\begin{aligned} f(x_0) &= \tfrac{1}{2}|x_1\rangle + \tfrac{1}{2}|x_2\rangle \\ f(x_1) &= \tfrac{1}{2}|x_3\rangle + \tfrac{1}{2}|x_4\rangle \\ f(x_2) &= \tfrac{1}{2}|x_5\rangle + \tfrac{1}{2}|x_6\rangle \end{aligned} \qquad \begin{aligned} f(x_3) &= \tfrac{1}{2}|x_1\rangle + \tfrac{1}{2}|1\rangle \\ f(x_4) &= \tfrac{1}{2}|2\rangle + \tfrac{1}{2}|3\rangle \\ f(x_5) &= \tfrac{1}{2}|4\rangle + \tfrac{1}{2}|5\rangle \\ f(x_6) &= \tfrac{1}{2}|x_2\rangle + \tfrac{1}{2}|6\rangle \end{aligned}$$

We will start in position x_0 and iterate f a couple of times, using composition $\circ$ from (10.4).

$$\begin{aligned} f(x_0) &= \tfrac{1}{2}|x_1\rangle + \tfrac{1}{2}|x_2\rangle \\ f^2(x_0) &= \tfrac{1}{4}|x_3\rangle + \tfrac{1}{4}|x_4\rangle + \tfrac{1}{4}|x_5\rangle + \tfrac{1}{4}|x_6\rangle \\ f^3(x_0) &= \tfrac{1}{8}|x_1\rangle + \tfrac{1}{8}|x_2\rangle + \tfrac{1}{8}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|3\rangle + \tfrac{1}{8}|4\rangle + \tfrac{1}{8}|5\rangle + \tfrac{1}{8}|6\rangle \\ f^4(x_0) &= \tfrac{1}{16}|x_3\rangle + \tfrac{1}{16}|x_4\rangle + \tfrac{1}{16}|x_5\rangle + \tfrac{1}{16}|x_6\rangle + \tfrac{1}{8}|1\rangle + \tfrac{1}{8}|2\rangle + \tfrac{1}{8}|3\rangle + \tfrac{1}{8}|4\rangle + \tfrac{1}{8}|5\rangle + \tfrac{1}{8}|6\rangle \\ f^5(x_0) &= \tfrac{1}{32}|x_1\rangle + \tfrac{1}{32}|x_2\rangle + \tfrac{5}{32}|1\rangle + \tfrac{5}{32}|2\rangle + \tfrac{5}{32}|3\rangle + \tfrac{5}{32}|4\rangle + \tfrac{5}{32}|5\rangle + \tfrac{5}{32}|6\rangle \\ f^6(x_0) &= \cdots \end{aligned}$$

One can prove, for each $a \in pips$, and $n \geq 1$,

$$f^{2n+1}(x_0)(a) \;=\; f^{2n+2}(x_0)(a) \;=\; \tfrac{1}{8} \;+\; \tfrac{1}{4} \cdot \tfrac{1}{8} \;+\; (\tfrac{1}{4})^2 \cdot \tfrac{1}{8} \;+\; \cdots \;+\; (\tfrac{1}{4})^{n-1} \cdot \tfrac{1}{8}.$$

When n goes to infinity this sum goes to:

$$\tfrac{1}{8} \cdot \sum_{k>0} (\tfrac{1}{4})^k \;\overset{(10.4)}{=}\; \tfrac{1}{8} \cdot \frac{1}{1 - 1/4} \;=\; \tfrac{1}{8} \cdot \tfrac{4}{3} \;=\; \tfrac{1}{6}.$$

Hence this map $f \colon X \to \mathcal{D}(X + A)$ can be used to mimic a dice. Notice that f itself involves only $\frac{1}{2}$ probabilities. In this way a coin can produce a dice. Thus:

$$\lim_{n\to\infty} f^n(x_0) \;=\; \tfrac{1}{6}|1\rangle + \tfrac{1}{6}|2\rangle + \tfrac{1}{6}|3\rangle + \tfrac{1}{6}|4\rangle + \tfrac{1}{6}|5\rangle + \tfrac{1}{6}|6\rangle \;=\; dice.$$

This is called the *limiting distribution* for position x_0. The limit is defined with respect to the total variation distance on $\mathcal{D}(X + A)$, see Section 4.8.1. Intuitively, this may be clear, but to make this precise we have to be little bit careful, since we are speaking about an equality between distributions $f^n(x_0) \in \mathcal{D}(X + A)$ and $dice \in \mathcal{D}(A)$ on different spaces. In order to make them match, we first need to map them to the same space $X + A$ via $\kappa_2 \colon A \to X + A$. Then, for $n \geq 3$,

$$\begin{aligned}
d\Big(f^n(x_0),\, \mathcal{D}(\kappa_2)(dice)\Big) \;&=\; \tfrac{1}{2} \cdot \sum_{i \in X+A} \Big| f^n(x_0)(u) - \mathcal{D}(\kappa_2)(dice)(u) \Big| \qquad \text{by Proposition 4.8.1} \\
&=\; \tfrac{1}{2} \cdot \sum_{x \in X} \Big| f^n(x_0)(x) \Big| + \tfrac{1}{2} \cdot \sum_{a \in A} \Big| f^n(x_0)(a) - dice(a) \Big| \\
&=\; \begin{cases} 2 \cdot \frac{1}{2} \cdot \frac{1}{2^n} \;+\; 6 \cdot \frac{1}{2} \cdot \Big((\frac{1}{4})^{n-3/2} \cdot \frac{1}{8} - \frac{1}{6} \Big) & \text{if } n \text{ is odd} \\ 4 \cdot \frac{1}{2} \cdot \frac{1}{2^n} \;+\; 6 \cdot \frac{1}{2} \cdot \Big((\frac{1}{4})^{n/2} \cdot \frac{1}{8} - \frac{1}{6} \Big) & \text{if } n \text{ is even.} \end{cases}
\end{aligned}$$

As shown above, this goes to zero as n goes to infinity.

One can ask: what is the average number of coin flips before seeing a particular dice pip, in the Markov model with output from Figure 10.1? This can be computed as mean of a distribution $\rho \in \mathcal{D}_\infty(\mathbb{N})$, where $\rho(n) \in [0, 1]$ is the probability of seeing a pip for the first time after n transitions. For instance, $\rho(0) = \rho(1) = \rho(2) = 0$, and:

$$\rho(3) \;=\; \tfrac{3}{4} \qquad \rho(4) \;=\; 0 \qquad \rho(5) \;=\; \tfrac{3}{16} \qquad \rho(6) \;=\; 0 \qquad \rho(7) \;=\; \tfrac{3}{64} \qquad \rho(8) \;=\; 0 \quad \textit{etc.}$$

Thus:

$$\begin{aligned}
Mean(\rho) \;=\; \sum_{n \geq 0} \rho(n) \cdot n \;&=\; \sum_{n \geq 1} \frac{3}{4^n} \cdot (2n + 1) \\
&=\; 6 \cdot \sum_{n \geq 1} \frac{n}{4^n} \;+\; \tfrac{3}{4} \cdot \sum_{n \geq 0} \frac{1}{4^n} \\
&=\; \frac{6 \cdot 1/4}{(1 - 1/4)^2} \;+\; \frac{3/4}{1 - 1/4} \qquad \text{by Theorem 1.7.4 (3) and (2)} \\
&=\; \frac{3/2}{9/16} \;+\; 1 \;=\; \tfrac{11}{3}.
\end{aligned}$$

We shall see a more systematic way to compute such averages later on in Example 10.5.6.

10.1.3 The game of craps

Craps is a well known game with one player throwing two dices repeatedly, until there is a 'win' or 'loose'. It works as follows. The player throws the two dices, say with outcomes called x and

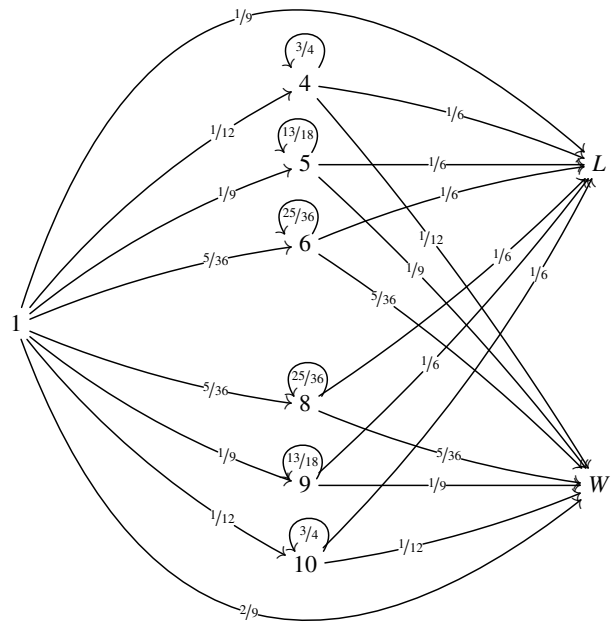


Figure 10.2 The transition diagram for the game of craps, using probabilities for the sum of two dices. The upper transition $1 \to L$ has the immediate-loose probability $tds(2) + tds(3) + tds(12) = \frac{1}{36} + \frac{1}{18} + \frac{1}{36} = \frac{1}{9}$. The lower transition $1 \to W$ is for immediate-win and has probability $tds(7) + tds(11) = \frac{1}{6} + \frac{1}{18} = \frac{2}{9}$.

y, and with sum $z := x + y$. Then z has the distribution called *tds* for 'two dices sum' that we have already seen in (4.9):

$$tds = \tfrac{1}{36}|2\rangle + \tfrac{1}{18}|3\rangle + \tfrac{1}{12}|4\rangle + \tfrac{1}{9}|5\rangle + \tfrac{5}{36}|6\rangle + \tfrac{1}{6}|7\rangle + \tfrac{5}{36}|8\rangle + \tfrac{1}{9}|9\rangle + \tfrac{1}{12}|10\rangle + \tfrac{1}{18}|11\rangle + \tfrac{1}{36}|12\rangle.$$

If z equals 7 or 11, player wins, and if z equals 2, 3 or 12 player loses. If there is no such immediate outcome, player goes into a loop, where the first sum z is re-used.

In this loop, a slight different rule applies: the player throws the two dices again, say with outcomes x, y. If the sum $x + y$ equals 7, the player wins; if $x + y$ equals the previous sum z, player looses. Otherwise, the player repeats this loop.

For the formalisation we use the set $X = \{1, 4, 5, 6, 8, 9, 10\}$ as positions, with 1 as initial point and the other positions representing the sum of the two dices, occurring in the distribution *tds* that do not immediately lead to win/loss. The initial throw of the two dices is represented by the left side of the transitions diagram in Figure 10.2, via the outgoing arrows from the initial position 1.

The 'loose' and 'win' observations are combined in a set $\{L, W\}$. The whole system is then described by a 'crap' transition function $c\colon X \to \mathcal{D}(X + \{L, W\})$, with:

$$\begin{aligned} c(1)(W) &= tds(7) + tds(11) = \tfrac{1}{6} + \tfrac{1}{18} = \tfrac{2}{9} \\ c(1)(L) &= tds(2) + tds(3) + tds(12) = \tfrac{1}{36} + \tfrac{1}{18} + \tfrac{1}{36} = \tfrac{1}{9}. \end{aligned}$$

For $i = 4, 5, 6, 8, 9, 10$ we have $c(1)(i) = tds(i)$, so that:

$$c(1) \;=\; \tfrac{1}{12}\big|4\rangle + \tfrac{1}{9}\big|5\rangle + \tfrac{5}{36}\big|6\rangle + \tfrac{5}{36}\big|8\rangle + \tfrac{1}{9}\big|9\rangle + \tfrac{1}{12}\big|10\rangle + \tfrac{2}{9}\big|W\rangle + \tfrac{1}{9}\big|L\rangle.$$

For the non-initial positions (in the middle of Figure 10.2), recall that throwing a sum of 7 means loosing, throwing the previous sum, recorded as the current position, is a win, everything else means a repeat, at this same position. This is incorporated in the following formula.

$$c(i) \;=\; tds(7)\big|L\rangle + (tds \models \mathbf{1}_i)\big|W\rangle + (\tfrac{5}{6} - tds \models \mathbf{1}_i)\big|i\rangle.$$

As a result:

$$\begin{aligned}
c(4) &= \tfrac{3}{4}|4\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{12}|W\rangle & c(8) &= \tfrac{25}{36}|8\rangle + \tfrac{1}{6}|L\rangle + \tfrac{5}{36}|W\rangle \\
c(5) &= \tfrac{13}{18}|5\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{9}|W\rangle & c(9) &= \tfrac{13}{18}|9\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{9}|W\rangle \\
c(6) &= \tfrac{25}{36}|6\rangle + \tfrac{1}{6}|L\rangle + \tfrac{5}{36}|W\rangle & c(10) &= \tfrac{3}{4}|10\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{12}|W\rangle.
\end{aligned}$$

What if we start (Kleisli) iterating the crap transition function $c\colon X \to \mathcal{D}(X + \{L, W\})$, starting in the initial position 1? It will give us in the limit a distribution on $\{L, W\}$ that captures the chances of loosing or winning the game of craps. Calculating these iterations works mechanically, bit is a bit of work. For instance, the second iteration step gives:

$$\begin{aligned}
c^2(1) &= \left(\tfrac{1}{12}\cdot\tfrac{3}{4}\right)|4\rangle + \left(\tfrac{1}{9}\cdot\tfrac{13}{18}\right)|5\rangle + \left(\tfrac{5}{36}\cdot\tfrac{25}{36}\right)|6\rangle + \left(\tfrac{5}{36}\cdot\tfrac{25}{36}\right)|8\rangle + \left(\tfrac{1}{9}\cdot\tfrac{13}{18}\right)|9\rangle + \left(\tfrac{1}{12}\cdot\tfrac{3}{4}\right)|10\rangle \\
&\quad + \left(\tfrac{1}{9} + \tfrac{1}{12}\cdot\tfrac{1}{6} + \tfrac{1}{9}\cdot\tfrac{1}{6} + \tfrac{5}{36}\cdot\tfrac{1}{6} + \tfrac{5}{36}\cdot\tfrac{1}{6} + \tfrac{1}{9}\cdot\tfrac{1}{6} + \tfrac{1}{12}\cdot\tfrac{1}{6}\right)|L\rangle \\
&\quad + \left(\tfrac{2}{9} + \tfrac{1}{12}\cdot\tfrac{1}{12} + \tfrac{1}{9}\cdot\tfrac{1}{9} + \tfrac{5}{36}\cdot\tfrac{5}{36} + \tfrac{5}{36}\cdot\tfrac{5}{36} + \tfrac{1}{9}\cdot\tfrac{1}{9} + \tfrac{1}{12}\cdot\tfrac{1}{12}\right)|W\rangle \\
&= \tfrac{1}{16}|4\rangle + \tfrac{13}{162}|5\rangle + \tfrac{125}{1296}|6\rangle + \tfrac{125}{1296}|8\rangle + \tfrac{13}{162}|9\rangle + \tfrac{1}{16}|10\rangle + \tfrac{2}{9}|L\rangle + \tfrac{97}{324}|W\rangle.
\end{aligned}$$

It is easy to see that the probabilities for non-initial positions $|i\rangle$ go to zero in multiple iterations. The interesting thing is to look at what happens to the probabilities of the outcomes L, W. For $n \geq 2$, the loose probability is:

$$\begin{aligned}
c^n(1)(L) &= \tfrac{1}{36} + \tfrac{1}{18} + \tfrac{1}{36} + 2\cdot\left(\tfrac{1}{12}\cdot(\tfrac{3}{4})^{n-2}\cdot\tfrac{1}{6} + \tfrac{1}{9}\cdot(\tfrac{13}{18})^{n-2}\cdot\tfrac{1}{6} + \tfrac{5}{36}\cdot(\tfrac{25}{36})^{n-2}\cdot\tfrac{1}{6}\right) \\
&= \tfrac{1}{9} + \tfrac{1}{108}\cdot\left(3\cdot(\tfrac{3}{4})^{n-2} + 4\cdot(\tfrac{13}{18})^{n-2} + 5\cdot(\tfrac{25}{36})^{n-2}\right).
\end{aligned}$$

As $n \to \infty$ we get by (10.4):

$$c^n(1)(L) \;\to\; \tfrac{1}{9} + \tfrac{1}{108}\cdot\left(\frac{3}{1 - {}^{3}\!/_{4}} + \frac{4}{1 - {}^{13}\!/_{18}} + \frac{5}{1 - {}^{25}\!/_{36}}\right) \;=\; \tfrac{251}{495} \;\approx\; 0.5070707.$$

Hence, the chance of loosing is eventually slightly higher than the chance of winning, see also Figure 10.3.

10.1.4 Stochastic processes

Assume that we have a Markov model with output $c\colon X \rightarrowtail X + A$, with an initial distribution $\sigma \in \mathcal{D}(X)$. Then we can start computing the pushforward $c_*(\sigma) \in \mathcal{D}(X + A)$, and subsequently $(c^2)_*(\sigma) \in \mathcal{D}(X + A)$, *etc.* The composition involved in iterating c^n is the one from (10.4). In this way we obtain a series of predicates $p_1, p_2, \ldots$ on A, namely:

$$p_n \;=\; \left(A \xrightarrow{\kappa_1} X + A \xrightarrow{(c^n)_*(\sigma)} [0, 1]\right).$$

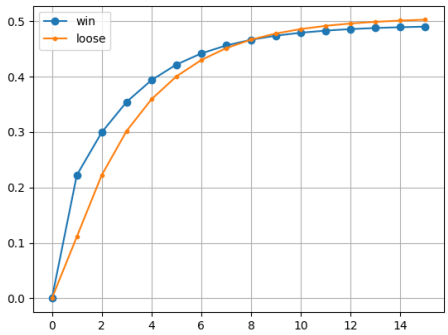


Figure 10.3 Probabilities of winning and loosing in the first 15 iterations of the craps transition function $X \to \mathcal{D}(X + \{L, W\})$.

Such series of predicates $p_n \in Pred(A)$ on the set of outputs A are used in (discrete time) stochastic processes. These predicates tell what can be observed at successive stages of the computation. Often, this is all what is (externally) available, and the model itself $X \rightsquigarrow X + A$ that gives rise to these predicates remains hidden.

One can also consider the set of outputs A as a subset of the coproduct $A + X$, via the inclusion $\kappa_2 \colon A \hookrightarrow X + A$. This yiels a sharp predicate $\mathbf{1}_A \colon X + A \to \{0, 1\}$, with $x \mapsto 0$ and $a \mapsto 1$. We can now compute successive updates:

$$\sigma_n \coloneqq (c^n)_*(\sigma)\big|_{\mathbf{1}_A} \in \mathcal{D}(A).$$

In presence of an observable $q \colon A \to \mathbb{R}$ we then obtain a sequence of random variables (σ_n, q).

We will make these constructions more concrete.

Example 10.1.4 Consider the last stage of a tennis game, at deuce (score of 40-40). Your chance of winning a point is $r \in (0, 1)$. We capture the situation via a Markov model with output, with three positions D = deuce, A = ahead (from your perspective), B = behind, and with 2 outputs W = win (for you) and L = loose. This model is:

$$\{D, A, B\} \xrightarrow{c} \mathcal{D}\Big(\{D, A, B\} + \{W, L\}\Big) \qquad \text{with} \qquad \begin{cases} c(D) \;=\; r|A\rangle + (1-r)|B\rangle \\ c(A) \;=\; (1-r)|D\rangle + r|W\rangle \\ c(B) \;=\; r|D\rangle + (1-r)|L\rangle. \end{cases}$$

Let's take probability $r = \frac{1}{3}$, and initial (deuce) distribution $\sigma = 1|D\rangle$. Then we can iterate a bit and condition on the output.

stage	$(c^n)_*(\sigma)$	$(c^n)_*(\sigma)\big\vert_{\mathbf{1}_{\{W,L\}}}$
$n = 1$	$\frac{1}{3}\vert A\rangle + \frac{2}{3}\vert B\rangle$	(undefined)
$n = 2$	$\frac{4}{9}\vert D\rangle + \frac{1}{9}\vert W\rangle + \frac{4}{9}\vert L\rangle$	$\frac{1}{5}\vert W\rangle + \frac{4}{5}\vert L\rangle$
$n = 3$	$\frac{4}{27}\vert A\rangle + \frac{8}{27}\vert B\rangle + \frac{1}{9}\vert W\rangle + \frac{4}{9}\vert L\rangle$	$\frac{1}{5}\vert W\rangle + \frac{4}{5}\vert L\rangle$
$n = 4$	$\frac{16}{81}\vert D\rangle + \frac{13}{81}\vert W\rangle + \frac{52}{81}\vert L\rangle$	$\frac{1}{5}\vert W\rangle + \frac{4}{5}\vert L\rangle$
$n = 5$	$\frac{16}{243}\vert A\rangle + \frac{32}{243}\vert B\rangle + \frac{13}{81}\vert W\rangle + \frac{52}{81}\vert L\rangle$	$\frac{1}{5}\vert W\rangle + \frac{4}{5}\vert L\rangle$

Interestingly, the win-loose distribution remains the same, eventhough it arises from a different distribution at each stage n, see Exercise 10.1.5 for more details.

Exercises

10.1.1 Show that in the dice-via-coins illustration in Subsection 10.1.2 iteration starting in positions x_1 and x_2 yields:

$$\lim_{n\to\infty} f^n(x_1) = \tfrac{1}{3}\vert 1\rangle + \tfrac{1}{3}\vert 2\rangle + \tfrac{1}{3}\vert 3\rangle \qquad \lim_{n\to\infty} f^n(x_2) = \tfrac{1}{3}\vert 4\rangle + \tfrac{1}{3}\vert 5\rangle + \tfrac{1}{3}\vert 6\rangle.$$

10.1.2 The following example goes back to Kozen (2009) and is a simpler version of the dice-via-coins illustration in Subsection 10.1.2. Consider a system with two positions $1, 2$ and with two outcomes H, T for head and tail, with a probabilistic transition system:

$$\begin{tikzcd}
1 \arrow[r, "1/2"] \arrow[d, bend right, "1/2"'] & H \\
2 \arrow[r, "1/2"'] \arrow[u, bend right, "1/2"'] & T
\end{tikzcd}$$

This example can be described with spaces $X = \{1, 2\}$ and $A = \{H, T\}$ via a map $c\colon X \to \mathcal{D}(X + A)$ with:

$$c(1) = \tfrac{1}{2}\vert 2\rangle + \tfrac{1}{2}\vert H\rangle \qquad \text{and} \qquad c(2) = \tfrac{1}{2}\vert 1\rangle + \tfrac{1}{2}\vert T\rangle.$$

Hence a fair coin can be used to create a biased coin.

1 Compute a number of (Kleisli) iterations, starting in position 1, namely $c(1), c^2(1), c^3(1), c^4(1), \ldots$

2 Show in general that:

$$c^{2n}(1) = \frac{1}{4^n}\vert 1\rangle + 2\cdot\left(\sum_{1\le k\le n}\frac{1}{4^k}\right)\vert H\rangle + \left(\sum_{1\le k\le n}\frac{1}{4^k}\right)\vert T\rangle$$
$$c^{2n+1}(1) = \frac{1}{2}\cdot\frac{1}{4^n}\vert 2\rangle + 2\cdot\left(\sum_{1\le k\le n+1}\frac{1}{4^k}\right)\vert H\rangle + \left(\sum_{1\le k\le n}\frac{1}{4^k}\right)\vert T\rangle$$

3 Conclude that:

$$\lim_{n\to\infty} c^n(1) = \tfrac{2}{3}\vert H\rangle + \tfrac{1}{3}\vert T\rangle.$$

4 Show also that starting in position 2 yields:

$$\lim_{n\to\infty} c^n(2) = \tfrac{1}{3}\vert H\rangle + \tfrac{2}{3}\vert T\rangle.$$

10.1.3 For $a \in A = \{1,2,3,4,5,6\}$ in the dice-via-coins example of Subsection 10.1.2 write $[a]\colon X \to [0,1]$ for the predicate $[a] := f^*(\mathbf{1}_a)$, so that:

$$[a](x) = f^*(\mathbf{1}_a)(x) = f(x) \models \mathbf{1}_{\kappa_2 a} = f(x)(a),$$

where $f\colon X \to \mathcal{D}(X+A)$ is the transition map in this example, acting on $X = \{x_0, \ldots, x_6\}$, see Figure 10.1.

1 Take $p = [1] \ovee [6]$ and check that:

$$p = \tfrac{1}{2} \cdot \mathbf{1}_{x_3} + \tfrac{1}{2} \cdot \mathbf{1}_{x_6}.$$

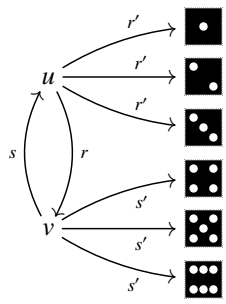
2 Show that the predicate $q = \frac{2}{7} \cdot \mathbf{1}_{x_0} + \frac{2}{7} \cdot \mathbf{1}_{x_1} + \frac{2}{7} \cdot \mathbf{1}_{x_2} + \frac{4}{7} \cdot \mathbf{1}_{x_3} + \frac{4}{7} \cdot \mathbf{1}_{x_6}$ satisfies the fixed point equation:

$$q = p \ovee \big(p^\perp \mathbin{\&} f^*([q, \mathbf{0}])\big).$$

where the predicate $[q, \mathbf{0}]\colon X + A \to [0,1]$ sends $x \in X$ to $q(x)$ and $a \in A$ to 0.

3 Now take $p = [2]+[3]+[4]+[5]$ and $q = \frac{2}{3}\cdot\mathbf{1}_{x_0} + \frac{2}{3}\cdot\mathbf{1}_{x_1} + \frac{2}{3}\cdot\mathbf{1}_{x_2} + \frac{1}{3}\cdot\mathbf{1}_{x_3} + 1\cdot\mathbf{1}_{x_4} + 1\cdot\mathbf{1}_{x_5} + \frac{1}{3}\cdot\mathbf{1}_{x_6}$. Show again that $q = p \ovee (p^\perp \mathbin{\&} f^*([q, \mathbf{0}]))$.

10.1.4 Consider the transition diagram below with two positions u, v and with two parameters $r, s \in (0,1)$.



where $\begin{cases} r' = \frac{1}{3}(1-r) \\ s' = \frac{1}{3}(1-s). \end{cases}$

1 Describe this transition diagram as a Markov model with output $c\colon X \rightsquigarrow X + A$.

2 Assume the relationship $r = \frac{1}{2-s}$. Show that iterating c, starting from position u, yields, in the limit, the (uniform) dice distribution.

3 Notice that with the relationship $r = \frac{1}{2-s}$ the number s can range all over $(0,1)$ but r only over $(\frac{1}{2}, 1)$. Explain how this works.

10.1.5 Consider the general description of the the tennis endgame model c in Example 10.1.4, with parameter $r \in (0,1)$ for your chance of winning a point.

1 Write for $n \in \mathbb{N}_{>0}$,

$$h(n) := \sum_{1 \le i \le n} 2^{i-1} \cdot r^{i-1} \cdot (1-r)^{i-1}.$$

Show that for $n \ge 1$,

$$\begin{aligned} c_*(\sigma) &= r\big|A\rangle + (1-r)\big|B\rangle \\ (c^{2n})_*(\sigma) &= 2^n \cdot r^n \cdot (1-r)^n\big|D\rangle + r^2 \cdot h(n)\big|W\rangle + (1-r)^2 \cdot h(n)\big|L\rangle. \\ (c^{2n+1})_*(\sigma) &= 2^n \cdot r^{n+1} \cdot (1-r)^n\big|A\rangle + 2^n \cdot r^n \cdot (1-r)^{n+1}\big|B\rangle \\ &\qquad + r^2 \cdot h(n)\big|W\rangle + (1-r)^2 \cdot h(n)\big|L\rangle. \end{aligned}$$

2 Prove now that that for $n \geq 2$

$$(c^n)_*(\sigma)\big|_{\mathbf{1}_{\{W,L\}}} = \frac{r^2}{r^2 + (1-r)^2}\big|W\rangle + \frac{(1-r)^2}{r^2 + (1-r)^2}\big|L\rangle.$$

Check that this matches the findings in Example 10.1.4 for $r = \frac{1}{3}$.

10.1.6 Fix a set A and turn a channel for the monad $\mathcal{D}((-) + A)$ into a channel for $\mathcal{D}$, as in:

$$\Big(X \xrightarrow{f} \mathcal{D}(Y + A)\Big) \longmapsto \Big(X + A \xrightarrow{G(f)} \mathcal{D}(Y + A)\Big)$$

where $G(f)(x) = f(x)$ and $G(f)(a) = 1|a\rangle$.

1 Show that G can be turned into a functor $\mathbf{Chan}(\mathcal{D}((-) + A)) \to \mathbf{Chan}(\mathcal{D})$. In this way a Markov chain with output can be turned into an ordinary Markov chain.

2 Show that, in the other direction, there is an 'inclusion' functor $F\colon \mathbf{Chan}(\mathcal{D}) \to \mathbf{Chan}(\mathcal{D}((-)+A))$, which is formally defined on morphisms as $F(f) = \mathcal{D}(\kappa_1) \circ f$.

3 Check that a $\mathcal{D}$-channel $X \rightsquigarrow G(Y)$ is the same as a $\mathcal{D}((-) + A)$-channel $F(X) \rightsquigarrow Y$. In categorical terms, this says that the functor F is 'left adjoint' to G.

10.1.7 Let $\mathbb{C}$ be a category with coproducts $(+, 0)$.

1 Use the uniqueness property of cotuples in Definition 10.1.1 to show that:

$$[\kappa_1, \kappa_2] = id \qquad h \circ [f, g] = [h \circ f, h \circ g].$$

2 Use the definition (10.3) for the coproduct $f+g$ of arrows to check that $+$ forms a functor $\mathbb{C} \times \mathbb{C} \to \mathbb{C}$.

10.1.8 We have been using products $\times$ throughout this book, but we never formally defined them in a categorical manner, like we did for coproducts in Definition 10.1.1. Here we go: the *product* of two objects X, Y in a category $\mathbb{C}$ is given by an object $X \times Y$ with two *projection* arrows $\pi_1\colon X \times Y \to X$ and $\pi_2\colon X \times Y \to Y$ such that for each pair $f\colon Z \to X$ and $g\colon Z \to Y$ there is a unique *tuple* map $\langle f, g\rangle\colon Z \to X \times Y$ with $\pi_1 \circ \langle f, g\rangle = f$ and $\pi_2 \circ \langle f, g\rangle = g$.

Diagrammatically this amounts to:

$$\begin{tikzcd} X & X \times Y \arrow[l, "\pi_1"'] \arrow[r, "\pi_2"] & Y \\ & Z \arrow[ul, "f"] \arrow[u, dashed, "\langle f,g\rangle"'] \arrow[ur, "g"'] & \end{tikzcd}$$

1 Check that the categories **Sets** and **Mon** (of monoids) have products, as just defined.

2 Check that a category $\mathbb{C}$ has coproducts if and only if $\mathbb{C}^{\mathrm{op}}$ has products (and vice-versa). Thus, from an abstract perspective, products and coproducts are the same structures, albeit in dual (opposite) universes.

10.1.9 Show that in **Sets** products $\times$ distribute over coproducts $+$ in the sense that the canonical map

$$(A \times X) + (A \times Y) \xrightarrow{[id \times \kappa_1, id \times \kappa_2]} A \times (X + Y)$$

is an isomorphism.

10.1.10 Let $T\colon \mathbf{Sets} \to \mathbf{Sets}$ be a monad. Show that the associated category of channels $\mathbf{Chan}(T)$ has coproducts.

10.1.11 Let $A \in \mathbb{C}$ be a fixed object in a category with coproducts, and let $T = (T, \mathit{unit}, \mathit{flat})$ be a monad on $\mathbb{C}$. Recall from Lemma 10.1.2 that $(-) + A$ is also a monad on $\mathbb{C}$.

1 Show that the maps $\lambda_X := [T(\kappa_1), \mathit{unit} \circ \kappa_2] : T(X) + A \to T(X + A)$ form a natural transformation.
2 Prove that λ is a distributive law of monads, see Exercise 3.9.4.
3 Conclude, via Exercise 3.9.4, that the composite $X \mapsto T(X + A)$ is a monad and show that the associated flatten map and Kleisli composition can be described as:

$$\begin{aligned} \mathit{flat}^{T(-+A)} &= \mathit{flat} \circ T([\mathit{id}, \mathit{unit} \circ \kappa_2]) \\ g \odot f &= \mathit{flat} \circ T([g, \mathit{unit} \circ \kappa_2]) \circ f. \end{aligned}$$

10.2 Markov models with rewards

This section will look into an extension of Markov models with a reward function, sometimes also called reinforcement signal. It assigns to each position a numerical reward, as real number. Such a reward can a be negative, and is then seen as a cost. The aim for such models is to obtain 'optimal' rewards, which can be obtained as fixed points. Therefor we need a classical result about complete metric spaces, namely Banach's Fixed Point Theorem. We include an associated reasoning principle.

Definition 10.2.1 Let (X, d_X), (Y, d_Y) be two metric spaces. A function $f : X \to Y$ is *Lipschitz* or *L-Lipschitz*, if there is a number $L \in \mathbb{R}_{>0}$ such that:

$$d_Y(f(x), f(x')) \leq L \cdot d_X(x, x'), \qquad \text{for all } x, x' \in X.$$

This number L is sometimes called the Lipschitz constant. There are two special cases of interest: the map f is called *short* when $L = 1$ and *contractive* when $L < 1$.

Theorem 10.2.2 *Let (X, d) be a complete metric space, with a contractive function $f : X \to X$.*

1 *The function f has a unique fixed point, written as $\mathrm{fix}(f) \in X$ with $f(\mathrm{fix}(f)) = \mathrm{fix}(f)$. This fixed point can be obtained from an arbitrary element $x \in X$ via iteration: $\mathrm{fix}(f) = \lim_{n\to\infty} f^n(x)$.*
2 *Let the metric space X now also be ordered, in such a way that for each $x \in X$ the subsets $\downarrow x = \{y \in X \mid y \leq x\}$ and $\uparrow x = \{y \in X \mid y \geq x\}$ are (metrically) closed. If the (contractive) function $f : X \to X$ is also monotone, then its fixed point $\mathrm{fix}(f) \in X$ is both:*
   - *the least pre-fixed point: $f(x) \leq x$ implies $\mathrm{fix}(f) \leq x$*
   - *the greatest post-fixed point: $x \leq f(x)$ implies $x \leq \mathrm{fix}(f)$.*

Proof. The fixed point result in the first item is completely standard, see *e.g.* (Granas and Dugundji, 2023, Thm. 1.1). For the second point we notice that if $f(x) \leq x$, then $f^n(x) \leq x$ for each n, by monotonicity. This means that $f^n(x) \in \downarrow x$, and so also that the limit $\mathit{fix}(f)$ is in $\downarrow x$. ❑

Let Y be an arbitrary set. An observable $f : Y \to \mathbb{R}$ is called *bounded* if there is a number $b \in \mathbb{R}_{>0}$ with $|f(y)| \leq b$ for all $y \in Y$. We write:

$$\mathit{BObs}(Y) := \{f : Y \to \mathbb{R} \mid f \text{ is bounded}\} \subseteq \mathit{Obs}(Y). \tag{10.5}$$

This set $BObs(Y)$ comes with the supremum distance, as in Lemma 4.8.14 for predicates:

$$d(f_1, f_2) = \bigvee_{y \in Y} \big| f_1(y) - f_2(y) \big|.$$

In this way $BObs(Y)$ is an ordered complete metric space. In fact, it is at the same time a vector space, forming a complete order unit space. Often it is written as $\ell^\infty(Y)$. Such spaces $BObs(Y)$ will be used below as setting for Banach's Fixed Point Theorem, in the form of so-called Bellman equations, from Bellman (1957). We notice that when the set Y is finite, say with n elements, then every observable $Y \to \mathbb{R}$ is bounded and so $BObs(Y) = Obs(Y) \cong \mathbb{R}^n$.

We are now ready to introduce the kind of probabilistic automata that we will consider in this section.

Definition 10.2.3 A *Markov Model with rewards* is a Markov model $t \colon X \to \mathcal{D}(X)$ together with a bounded *reward* observable $r \colon X \to \mathbb{R}$.

Let $L \in (0, 1)$ be a Lipschitz factor. The *return operator* $R \colon Obs(X) \to Obs(X)$ associated with a model (t, r) is defined in terms of observable transformation t^*, as:

$$R(f) \coloneqq r + L \cdot t^*(f). \tag{10.6}$$

The dependence of R on t, r, L is left implicit. Soon it will be shown that the operator R is contractive, with Lipschitz constant L. Its fixed point $fix(R) \colon X \to \mathbb{R}$ will be called the *return observable* of the model. It is the solution to the *Bellman equation* $R(f) = f$.

In a Markov model with rewards $\langle t, r \rangle \colon X \to \mathcal{D}(X) \times \mathbb{R}$ the *immediate* reward in position x is $r(x)$. The (expected) reward in the successor states of x is:

$$\sum_{x' \in X} t(x)(x') \cdot r(x') = t(x) \models r = t^*(r)(x).$$

The *future* rewards observables $X \to \mathbb{R}$ can be given as sum, say with finitely many steps:

$$\sum_{i=0}^{n} (t^i)^*(r) = r + t^*(r) + t^*(t^*(r)) + \cdots + (t^n)^*(r).$$

Even when r is bounded, this sum may diverge when n goes to infinity. That's why it is common to use a Lipschitz (or 'discount') factor $L \in (0, 1)$ which is used in each successive step, giving summands $L^n \cdot (t^n)^*(r)$. We shall soon see that this expression forms one way of describing the return $fix(R)$.

Proposition 10.2.4 *Let $\langle t, r \rangle \colon X \to \mathcal{D}(X) \times \mathbb{R}$ be a Markov model with rewards, with Lipschitz factor $L \in (0, 1)$.*

1 *The return operator $R \colon Obs(X) \to Obs(X)$ defined in (10.6) is contractive, when restricted to bounded functions. Its Lipschitz factor is L. Its unique fixed point can (thus) be described as:*

$$fix(R) = \lim_{n \to \infty} R^n(\mathbf{0}) = \lim_{n \to \infty} r_n \qquad \textit{where} \qquad \begin{cases} r_0 = \mathbf{0} \\ r_1 = r \\ r_{n+1} = r + L \cdot t^*(r_n). \end{cases} \tag{10.7}$$

2 *The return observable* $\mathrm{fix}(R)$ *can also be described as infinite 'discounted' sum of observable transformations:*

$$\mathrm{fix}(R) = \sum_{n\in\mathbb{N}} L^n \cdot (t^n)^*(r).$$

As a result, the function $\mathrm{fix}(R)\colon X \to \mathbb{R}$ *is bounded by* $\frac{b}{1-L}$*, where* $b \in \mathbb{N}$ *is a bound for* r*. This means that* $\mathrm{fix}(R) \in BObs(X)$ *and that* $R\colon Obs(X) \to Obs(X)$ *restricts to* $R\colon BObs(X) \to BObs(X)$*.*

Proof. 1 We first show that R is contractive when restricted to bounded observables. For $f_1, f_2 \in Obs(X)$,

$$\begin{aligned} d\big(R(f_1), F(f_2)\big) &= \bigvee_{x\in X} \big|R(f_1)(x) - R(f_2)(x)\big| = \bigvee_{x\in X} L \cdot \big|t^*(f_1)(x) - t^*(f_2)(x)\big| \\ &= L \cdot d(t^*(f_1), t^*(f_2)) \\ &\leq L \cdot d(f_1, f_2). \end{aligned}$$

The last inequality uses that observable transformation t^* is short, like predicate transformation in Lemma 4.8.15.

Next, as we know from Theorem 10.2.2, the fixed point of R can be obtained via iteration, starting at any point in the space. Hence we might as well start at the zero observable $\mathbf{0} \in BObs(X)$. This gives $\mathrm{fix}(R) = \lim_{n\to\infty} r_n$ with $r_0 = \mathbf{0}$ and:

$$r_1 = R(\mathbf{0}) = r + L \cdot t^*(\mathbf{0}) = r + L \cdot \mathbf{0} = r.$$

Next: $r_{n+1} = R(r_n) = r + L \cdot t^*(r_n)$.

2 Let $r\colon X \to \mathbb{R}$ be bounded by $b \in \mathbb{R}_{>0}$. We claim that the iterations $(t^n)^*(r)$ are then also bounded by b — so that $(t^n)^*(r) \in \ell^\infty(X)$. The case $n = 0$ holds by assumption; via the induction hypothesis we get:

$$\begin{aligned} \big|(t^{n+1})^*(r)(x)\big| = \left|\sum_{x'\in X} t(x)(x') \cdot (t^n)^*(r)(x')\right| &\leq \sum_{x'\in X} t(x)(x') \cdot \big|(t^n)^*(r)(x')\big| \\ &\leq \sum_{x'\in X} t(x)(x') \cdot b = b. \end{aligned}$$

Now we can find the bound for the observable $\mathrm{fix}(R)\colon X \to \mathbb{R}$, namely:

$$\begin{aligned} \big|\mathrm{fix}(R)(x)\big| = \left|\sum_{n\geq 0} L^n \cdot (t^n)^*(r)(x)\right| &\leq \sum_{n\geq 0} L^n \cdot \big|(t^n)^*(r)(x)\big| \\ &\leq \sum_{n\geq 0} L^n \cdot b \overset{(10.1)}{=} \frac{b}{1-L}. \end{aligned}$$

Next, since observable transformation t^* preserves sums and scalar multiplication we obtain that $\overline{r} := \sum_n L^n \cdot (t^n)^*(r)$ is the fixed point:

$$\begin{aligned} R(\overline{r}) = r + L \cdot t^*(\overline{r}) &= r + L \cdot t^*\left(\sum_{n\geq 0} L^n \cdot (t^n)^*(r)\right) \\ &= r + L \cdot \sum_{n\geq 0} L^n \cdot t^*((t^n)^*(r)) \\ &= L^0 \cdot (t^0)^*(r) + \sum_{n\geq 0} L^{n+1} \cdot (t^{n+1})^*(r) = \overline{r}. \end{aligned}$$

❑

The above descriptions embody the so-called infinite-horizon discounted approach: they use the infinite limit / sum of rewards, but discounted via L, so that the return is still bounded. In an alternative approach, not followed here, one can use only finitely many steps ahead, to determine the return in a particular position.

Example 10.2.5 The following elementary example is copied from Moore (2002). Consider the following transition diagram of a Markov model with rewards, involving three weather positions S = sun, W = wind, H = hail.

$$\begin{tikzcd} S \arrow[loop left, "1/2"] \arrow[r, bend left, "1/2"] & W \arrow[l, bend left, "1/2"] \arrow[r, bend left, "1/2"] & H \arrow[l, bend left, "1/2"] \arrow[loop right, "1/2"] \end{tikzcd} \quad \text{with reward} \quad \begin{cases} r(S) = 4 \\ r(W) = 0 \\ r(H) = -8. \end{cases}$$

The transition function $t\colon \{S, W, H\} \to \mathcal{D}(\{S, W, H\})$ is thus given by:

$$t(S) = \tfrac{1}{2}|S\rangle + \tfrac{1}{2}|W\rangle \qquad t(W) = \tfrac{1}{2}|S\rangle + \tfrac{1}{2}|H\rangle \qquad t(H) = \tfrac{1}{2}|W\rangle + \tfrac{1}{2}|H\rangle.$$

We consider two ways to find the return observable, as function $\{S, W, H\} \to \mathbb{R}$. We can algebraically solve the Bellman equation $r + L \cdot t^*(f) = f$, say for $L = \frac{1}{2}$. This gives three equations:

$$\begin{aligned} f(S) &= r(S) + \tfrac{1}{2} \cdot (t(S)(S) \cdot f(S) + t(S)(W) \cdot f(W)) \\ &= 4 + \tfrac{1}{4} \cdot f(S) + \tfrac{1}{4} \cdot f(W) \\ f(W) &= r(W) + \tfrac{1}{2} \cdot (t(W)(S) \cdot f(S) + t(W)(H) \cdot f(H)) \\ &= \tfrac{1}{4} \cdot f(S) + \tfrac{1}{4} \cdot f(H) \\ f(H) &= r(H) + \tfrac{1}{2} \cdot (t(H)(W) \cdot f(W) + t(H)(H) \cdot f(H)) \\ &= -8 + \tfrac{1}{4} \cdot f(W) + \tfrac{1}{4} \cdot f(H). \end{aligned}$$

It is not hard to see that this yields $f(S) = \frac{24}{5}$, $f(W) = -\frac{8}{5}$ and $f(H) = -\frac{56}{5}$. Thus, only the 'sun' position has positive return.

Alternatively, one can approximate the solution by computing the successive observables r_n from the return formulation in (10.7). We list the first few outcomes.

r_n	S	W	H
$n = 0$	4	0	-8
$n = 1$	5	-1	-10
$n = 2$	5	$\frac{5}{4}$	$\frac{43}{4}$
$n = 3$	$\frac{79}{16}$	$\frac{23}{16}$	-11
$n = 4$	$\frac{39}{8}$	$-\frac{97}{64}$	$-\frac{711}{64}$

This last row with $\frac{39}{8} = 4.875$ and $-\frac{97}{64} = -1.515625$ and $-\frac{711}{64} = -11.109375$ already comes close to the above exact solution $\frac{24}{5} = 4.8$ and $-\frac{8}{5} = -1.6$ and $-\frac{56}{5} = -11.2$.

We have seen that the return observable $\mathit{fix}(R)$ is the unique solution to the Bellman equation $R(f) = f$. Below is another characterisation, due to Feys et al. (2018). This approach will be further explored later on in this chapter. It uses the 'mean' operation from Definition 4.2.1 in the form of a function $\mathit{Mean}\colon \mathcal{D}(\mathbb{R}) \to \mathbb{R}$, as

$$\mathit{Mean}(\omega) = \omega \models \mathit{id} = \sum_{y \in \mathbb{R}} \omega(y) \cdot y.$$

This map $\mathit{Mean}\colon \mathcal{D}(\mathbb{R}) \to \mathbb{R}$ is an algebra of the distribution monad, see Exercise 4.2.5.

Proposition 10.2.6 *Let $L \in (0,1)$ be a fixed Lipschitz / discount factor. Define the* Bellman *map $\mathit{bm}_L\colon \mathcal{D}(\mathbb{R}) \times \mathbb{R} \to \mathbb{R}$ as:*

$$\mathit{bm}_L(\omega, z) \coloneqq z + L \cdot \mathit{Mean}(\omega) = z + L \cdot \sum_{y \in \mathbb{R}} \omega(y) \cdot y.$$

Let $\langle t, r\rangle\colon X \to \mathcal{D}(X) \times \mathbb{R}$ be a Markov model with rewards. The return observable $\mathit{fix}(R)$ is then the unique bounded function $f\colon X \to \mathbb{R}$ making the following diagram commute.

$$\begin{tikzcd}
\mathcal{D}(X) \times \mathbb{R} \arrow[r, "\mathcal{D}(f)\times id"] & \mathcal{D}(\mathbb{R}) \times \mathbb{R} \arrow[d, "\mathit{bm}_L"] \\
X \arrow[u, "\langle t,r\rangle"] \arrow[r, "f"] & \mathbb{R}
\end{tikzcd}$$

More abstractly one can formulate commutation of the above diagram as: the Bellman map $\mathit{bm}_L\colon \mathcal{D}(R) \times \mathbb{R} \to \mathbb{R}$ is a 'corecursive algebra' for the functor $\mathcal{D}(-) \times \mathbb{R}$, see Feys et al. (2018), or Definition 10.4.3 below, for the general concept.

Proof. We show that commutation of the diagram amounts to the Bellman equation $R(f) = f$, so that f is equal to its unique solution $\mathit{fix}(R) \in \ell^\infty(X)$. So let a bounded map $f\colon X \to \mathbb{R}$ make the diagram commute. Then, for $x \in X$,

$$\begin{aligned}
f(x) = \Big(\mathit{bm}_L \circ (\mathcal{D}(f) \times id) \circ \langle r, t\rangle\Big)(x) &= \mathit{bm}_L\Big(\mathcal{D}(f)(t(x)), r(x)\Big) \\
&= r(x) + L \cdot \left(\sum_{y \in \mathbb{R}} \mathcal{D}(f)(t(x))(y) \cdot y\right) \\
&= r(x) + L \cdot \left(\sum_{y \in \mathbb{R}} \sum_{x' \in f^{-1}(y)} t(x)(x') \cdot y\right) \\
&= r(x) + L \cdot \left(\sum_{x' \in X} t(x)(x') \cdot f(x')\right) \\
&= r(x) + L \cdot t^*(f)(x) = R(f)(x).
\end{aligned}$$

❑

Exercises

10.2.1 In Example 10.2.5 compute yourself:

1. the solution to the equations describing the observable f;
2. the values for r_n in the table.

10.2.2 Let X be an arbitrary set.

1. Let $f\colon X \to \mathbb{R}$ be a bounded observable, say with bound $b \in \mathbb{R}_{>0}$. Check that $\frac{1}{2} \cdot \mathbf{1} + \frac{1}{2b} \cdot f$ is a predicate on X.
2. Let $f\colon X \to \mathbb{R}_{\geq 0}$ be a bounded factor, say also with bound $b \in \mathbb{R}_{>0}$. Check that $\frac{1}{b} \cdot f$ is a predicate and that updating a distribution $\omega \in \mathcal{D}(X)$ with f gives the same outcome as updating ω with $\frac{1}{b} \cdot f$.

10.2.3 Let $\langle t, r\rangle \colon X \to \mathcal{D}(X) \times \mathbb{R}$ be a Markov model with rewards, with Lipschitz / discount factor L. Prove that for a distribution $\omega \in \mathcal{D}(X)$, the validity of the return observable $\mathit{fix}(R) \colon X \to \mathbb{R}$ in ω is the sum of the validities of the discounted reward observable r over all successor states (distributions):

$$\omega \models \mathit{fix}(R) \;=\; \sum_{n \in \mathbb{N}} (t^n)_*(\omega) \models L^n \cdot r.$$

10.2.4 Let $\langle t, r\rangle \colon X \to \mathcal{D}(X) \times \mathbb{R}$ be a Markov model with rewards, with Lipschitz / discount factor L. We assume that the set X is finite, say with N elements. Thus, we can identify the transition channel $t \colon X \rightsquigarrow X$ with an $N \times N$ stochastic matrix D_t, like in Remark 4.5.9.

1 Check that the (unique) solution $\mathit{fix}(R) \in \mathit{BObs}(N)$ can be described as:

$$\mathit{fix}(R) \;=\; \left(I - L \cdot D_t^T\right)^{-1} \cdot r,$$

where I is the $N \times N$ identity matrix, $(-)^T$ is matrix transpose, $(-)^{-1}$ is matrix inversion, and $\cdot$ is matrix-vector multiplication.

2 Check that the equations that are solved algebraically in Example 10.2.5 correspond to computing the matrix inverse from the previous point.

10.3 Markov decision processes

In the previous section we have looked at Markov models with rewards. Markov decision processes — commonly abbreviated as MDPs — are an extension of these models, with 'actions', see Puterman (1994). In each position there is a choice of actions, and each choice comes with a distribution over successor positions. The aim is to choose actions with highest long-run sum of rewards (called returns). MDPs have many applications in planning under uncertainty — of routes, schedules, paths *etc.* — and in reinforcement learning (see *e.g.* Kaelbling et al. (1996); Sutton and Barto (2005)), where one 'learns to act'. In cognitive science MDPs are used as models that make a connection between what one's beliefs and actions, see *e.g.* Parr et al. (2022); Smith et al. (2022).

Definition 10.3.1 Let's write A for a finite set of 'actions'. A *Markov decision process* (MDP) is a pair of functions $\langle t, r\rangle \colon X \to \mathcal{D}(X)^A \times \mathbb{R}$.

For a position $x \in X$ and action $a \in A$ the distribution $t(x)(a) \in \mathcal{D}(X)$ captures the successor positions following action a. We shall write $t_a \colon X \to \mathcal{D}(X)$ for the Markov model given by $t_a(x)(x') = t(x)(a)(x')$.

Given a discount factor $L \in (0, 1)$ we again associate a (contractive) operator $R \colon \mathit{BObs}(X) \to \mathit{BObs}(X)$ with this situation. It is now defined on bounded observable $f \colon X \to \mathbb{R}$ as:

$$R(f) \;\coloneqq\; r \,+\, L \cdot \max_{a \in A} \, (t_a)^*(f). \tag{10.8}$$

The (unique) solution $\mathit{fix}(R) \colon X \to \mathbb{R}$ is now called the *optimal return*. The equation $R(f) = f$ is called the *optimal Bellman equation*.

Thus, a Markov model with rewards, as used in the previous section, can be identified with an MDP with singleton set $A = 1$ of actions, corresponding to a single transition map t. In a Markov decision model one has a collection $(t_a)_{a \in A}$ of transition functions, indexed by actions $a \in A$.

Remark 10.3.2 The above definition of MDP has a rather simple formulation. In the literature more complicated variations exist — see *e.g.* Puterman (1994); Kaelbling et al. (1996); Sutton and Barto (2005) — of which we briefly discuss two versions.

1 In Definition 10.3.1 rewards are given via an observable $r\colon X \to \mathbb{R}$, like in the previous section. In the presence of actions one imagine that not just transitions but also rewards depend on actions. In that case one can work with a reward function of the $X \to \mathbb{R}^A$, where A is the set of actions.

2 We have used a set of actions A in such a way that in each position each $a \in A$ action is possible. In certain applications however the available actions may depend on the position. For instance, for a robot in a bounded maze, in the inner cells (positions) move actions north / east / south / west are possible, but in the cells on the upper boundary the robot may ony execute east / south / west actions. And in a corner cell only two actions are available.

One can formalise this by using an indexed collection $(A_x)_{x\in X}$ of actions, where A_x is the set of actions available in position x. There are different ways to describe the corresponding transitions functions, for instance as 'dependent' functions in:

$$\prod_{x\in X} \mathcal{D}(X)^{A_x}.$$

One can also form the disjoint union $A := \coprod_x A_x$ and use partial transition maps, as in:

$$X \longrightarrow (\mathcal{D}(X) + 1)^A$$

They are called 'reactive' systems in Glabbeek et al. (1995); Bartels et al. (2004); Sokolova (2011).

In the previous section we have seen how to obtain the return map of future rewards. The challenge for a MDP is to find a 'policy' that chooses the best action in each position that maximises the return. In general, a policy is a function $\pi\colon X \to \mathcal{D}(A)$, that is, a channel $X \rightsquigarrow A$, giving for each position a distribution of actions. Often such a policy is 'deterministic', forming a function $X \to A$, or a deterministic channel $X \rightsquigarrow A$. We collect some associated constructions.

Definition 10.3.3 Let $\pi\colon X \to \mathcal{D}(A)$ be a policy for a MDP $\langle t, r\rangle\colon X \to \mathcal{D}(X)^A \times \mathbb{R}$, with Lipschitz / discount factor $L \in (0, 1)$.

1 The associated Markov chain $t^\pi\colon X \to \mathcal{D}(X)$ is defined as composite channel:

$$t^\pi := \Big(X \xrightarrow{\langle id,\pi\rangle} X \times A \xrightarrow{\underline{t}} X\Big) \quad \text{where} \quad \underline{t}(x, a) := t(x)(a) = t_a(x). \tag{10.9}$$

2 We now write R^π for the return operator associated with the Markov model with rewards $\langle t^\pi, r\rangle\colon X \to \mathcal{D}(X) \times \mathbb{R}$, as in (10.6), so that $R^\pi(f) = r + L \cdot (t^\pi)^*(f)$. The associated return observable is then $fix(R^\pi)\colon X \to \mathbb{R}$.

3 The policy π is called optimal if its return is optimal, that is, if $fix(R^\pi) = fix(R)$, where R is as in (10.8). Equivalently, π is optimal if $fix(R^\pi)$ satisfies the optimal Bellman equation.

There is always a deterministic policy.

Proposition 10.3.4 *In an arbitrary MDP, with optimal return operator R as in* (10.8)*,*

1 *there is a deterministic 'greedy' policy function* $\chi\colon X \to A$ *which is optimal:* $\mathit{fix}(R^\chi) = \mathit{fix}(R)$*; this* χ *is chosen pointwise as:*

$$\chi(x) \in \operatorname*{argmax}_{a\in A} (t_a)^*(\mathit{fix}(R))(x). \tag{10.10}$$

2 *in general,* $\mathit{fix}(R) = \max_{\pi\colon X \rightsquigarrow A} \mathit{fix}(R^\pi)$*, pointwise.*

The latter equation explains why $\mathit{fix}(R)$ is called the 'optimal' return. Interestingly, the two points imply that the maximum over *probabilistic* policies $X \to \mathcal{D}(A)$ is the same as the maximum over *deterministic* policies $X \to A$.

Proof. Consider MDP $\langle t, r\rangle\colon X \to \mathcal{D}(X)^A \times \mathbb{R}$ with $L \in (0, 1)$.

1 By definition we have for each $x \in X$,

$$\mathit{fix}(R)(x) = R(\mathit{fix}(R))(x) \overset{(10.8)}{=} r(x) + L \cdot \max_{a\in A} (t_a)^*(\mathit{fix}(R))(x).$$

Let $\chi(x) \in A$ be an element for which the maximum is reached, as in (10.10). Then by construction:

$$\begin{aligned}
R^\chi(\mathit{fix}(R))(x) &\overset{(10.6)}{=} r(x) + L \cdot (t^\chi)^*(\mathit{fix}(R))(x)\\
&= r(x) + L \cdot \sum_{x'\in X} t_{\chi(x)}(x)(x') \cdot \mathit{fix}(R)(x')\\
&= r(x) + L \cdot \max_{a\in A} \sum_{x'\in X} t_a(x)(x') \cdot \mathit{fix}(R)(x')\\
&= r(x) + L \cdot \max_{a\in A} \sum_{x'\in X} (t_a)^*(\mathit{fix}(R))(x) \overset{(10.8)}{=} R(\mathit{fix}(R))(x) = \mathit{fix}(R)(x).
\end{aligned}$$

This proves the equality of functions $R^\chi(\mathit{fix}(R)) = \mathit{fix}(R)$. Thus, $\mathit{fix}(R) = \mathit{fix}(R^\chi)$ by uniqueness of fixed points.

2 The direction ($\leq$) is easy by the previous point: $\mathit{fix}(R)$ is equal to one of the $\mathit{fix}(R^\pi)$, namely when π is the greedy policy χ. For ($\geq$) we observe that $R^\pi(f) \leq R(f)$ for each policy π, since:

$$\begin{aligned}
R^\pi(f)(x) = r(x) + L \cdot (t^\pi)^*(f)(x) &= r(x) + L \cdot \sum_{x'\in X} t_{\pi(x)}(x)(x') \cdot f(x')\\
&\leq r(x) + L \cdot \max_{a\in A} \sum_{x'\in X} t_a(x)(x') \cdot f(x') = R(f)(x).
\end{aligned}$$

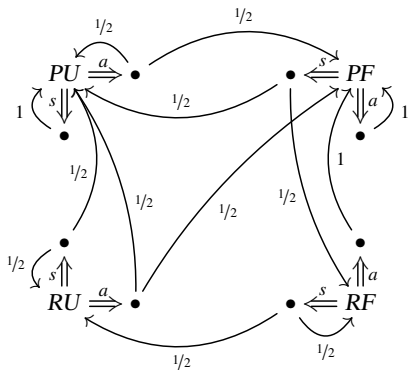
But then:

$$\mathit{fix}(R^\pi) = R^\pi(\mathit{fix}(R^\pi)) \leq R(\mathit{fix}(R^\pi)) \qquad \text{and thus} \qquad \mathit{fix}(R^\pi) \leq \mathit{fix}(R).$$

The latter step works via Theorem 10.2.2 (2), using that R is monotone. The join of these $\mathit{fix}(R^\pi)$ is then also below $\mathit{fix}(R)$. ❑

Example 10.3.5 We use a well known example from Moore (2002) that is often used for illustrating MDPs. It describes four possible positions of a start-up entrepreneur, abbreviated as PU = Poor & Unknown, PF = Poor & Famous, RU = Rich & Unknown, RF = Rich & Famous. Figure 10.4 contains the transition diagram between these positions. In each position there are two possible actions a = advertise and s = save, indicated by the double arrow $\Rightarrow$, followed by a distribution over (successor) positions. Thus, the example involves sets of positions and actions:

$$X = \{PU, PF, RU, RF\} \qquad \text{and} \qquad A = \{a, s\}.$$



$$r(PU) = 0$$
$$r(PF) = 0$$
$$r(RU) = 10$$
$$r(RF) = 10.$$

Figure 10.4 Transition diagram of a Markov decision process with four positions: PU = Poor & Unknown, PF = Poor & Famous, RU = Rich & Unknown, RF = Rich & Famous. In each position there are two actions a = advertise and s = save. The column on the right-hand side describes the reward in each position.

The transition map $t\colon X \to \mathcal{D}(X)^A$ is given by:

$$\begin{aligned} t(PU)(a) &= \tfrac{1}{2}\big|PU\rangle + \tfrac{1}{2}\big|PF\rangle & t(PU)(s) &= 1\big|PU\rangle \\ t(PF)(a) &= 1\big|PF\rangle & t(PF)(s) &= \tfrac{1}{2}\big|PU\rangle + \tfrac{1}{2}\big|RF\rangle \\ t(RU)(a) &= \tfrac{1}{2}\big|PU\rangle + \tfrac{1}{2}\big|PF\rangle & t(RU)(s) &= \tfrac{1}{2}\big|PU\rangle + \tfrac{1}{2}\big|RU\rangle \\ t(RF)(a) &= 1\big|PF\rangle & t(RF)(s) &= \tfrac{1}{2}\big|RU\rangle + \tfrac{1}{2}\big|RF\rangle. \end{aligned}$$

There is the optimal return operator R as in (10.8), with optimal return $fix(R) = \lim_{n\to\infty} R^n(\mathbf{0})$. The first few steps are:

$R^n(\mathbf{0})$	PU	PF	RU	RF
$n = 1$	0	0	10	10
$n = 2$	0	4.5	14.5	19
$n = 3$	2.02	8.55	16.52	25.1
$n = 4$	4.76	12.2	18.3	28.7
$n = 5$	7.63	15.1	20.4	31.2
$n = 6$	10.2	17.5	22.6	33.2
$\vdots$				
$n = 100$	31.6	38.6	44.0	54.2

The optimal greedy policy χ obtained via (10.10) is:

$$\chi(PU) = a \qquad \chi(PF) = s \qquad \chi(RU) = s \qquad \chi(RF) = s.$$

Thus, the greedy policy chooses to advertise (action a) in position PU and to save (action s) in all other positions. One can check that iterating $(R^\chi)^n(\mathbf{0})$ gives the same limit as in the above table.

In this example we can clearly distinguish an *exploration* phase, namely computation of $fix(R)$,

from an *exploitation* phase, namely the choice of (optimal) greedy policy. For more discussion on the trade-off between exploration and exploitation in interactive goal-seeking problem solving under uncertainty, see Sutton and Barto (2005).

We conclude this section with the analogue of Proposition 10.2.6 for MDPs, following Feys et al. (2018).

Proposition 10.3.6 *Let $L \in (0,1)$ be a fixed Lipschitz / discount factor. For a finite set of actions A, define the* optimal Bellman map $obm_L \colon \mathcal{D}(\mathbb{R})^A \times \mathbb{R} \to \mathbb{R}$ *as:*

$$obm_L(c,z) \coloneqq z + L \cdot \max_{a \in A} \mathit{Mean}(c(a)) = z + L \cdot \max_{a \in A} \sum_{y \in X} c(a)(y) \cdot y.$$

Let $\langle t, r \rangle \colon X \to \mathcal{D}(X)^A \times \mathbb{R}$ be an MDP. The optimal return $\mathit{fix}(R) \colon X \to \mathbb{R}$ is the unique bounded observable $f \colon X \to \mathbb{R}$ making the following diagram commute.

$$\begin{tikzcd}
\mathcal{D}(X)^A \times \mathbb{R} \arrow[r, "\mathcal{D}(f)^A \times id"] & \mathcal{D}(\mathbb{R})^A \times \mathbb{R} \arrow[d, "obm_L"] \\
X \arrow[u, "\langle t,r \rangle"] \arrow[r, "f"] & \mathbb{R}
\end{tikzcd}$$

Proof. Let a bounded map $f \colon X \to \mathbb{R}$ make the diagram commute. Then, for $x \in X$,

$$\begin{aligned}
f(x) &= \Big(obm_L \circ (\mathcal{D}(f)^A \times id) \circ \langle r, t \rangle\Big)(x) \\
&= obm_L\Big(\mathcal{D}(f)^A(t(x)), r(x)\Big) \\
&= r(x) + L \cdot \max_{a \in A} \sum_{y \in \mathbb{R}} \mathcal{D}(f)\big(t(x)(a)\big)(y) \cdot y \\
&= r(x) + L \cdot \max_{a \in A} \sum_{y \in \mathbb{R}} \sum_{x' \in f^{-1}(y)} t_a(x)(x') \cdot y \\
&= r(x) + L \cdot \max_{a \in A} \sum_{x' \in X} t_a(x)(x') \cdot f(x') \;=\; r(x) + L \cdot \max_{a \in A} (t_a)^*(f)(x) \overset{(10.8)}{=} R(f)(x).
\end{aligned}$$

Thus $R(f) = f$ and so $f = \mathit{fix}(R)$ by uniquenss of fixed points. ❑

Exercises

10.3.1 1 Check that the Markov chain $t^\pi \colon X \to \mathcal{D}(X)$ associated in (10.9) with policy $\pi \colon X \to \mathcal{D}(A)$ can be described explicitly as:

$$t^\pi(x)(x') = \sum_{a \in A} \pi(x)(a) \cdot t_a(x)(x') = \sum_{a \in A} \pi(x)(a) \cdot t(x)(a)(x').$$

2 Check that this means that the distribution $t^\pi(x)$ of successors of position x can be described as a convex sum of distributions of successors.

3 Show that when the policy π is deterministic one gets $t^\pi(x) = t_{\pi(x)}(x)$, where π chooses a particular action $\pi(x) \in A$ in each position x.

10.3.2 Consider the Poor / Rich and Unknown / Famous Example 10.3.5.

1 Compute the expected rewards after one step $t(x)(y) \models r$ for each $x \in \{PU, PF, RU, RF\}$ and $y \in \{a, s\}$.

2 Suppose you are in position RF and you can take two consecutive actions (both either a or s). Which sequence of two actions gives the highest reward?

10.3.3 Show that the operator $R\colon BObs(X) \to BObs(X)$ defined in (10.8) is contractive. (This is a bit more tricky than the proof of Proposition 10.2.4 (1); if needed, a proof can be found in (Puterman, 1994, Prop. 6.2.4).)

10.4 Probabilistic automata as coalgebras

So far we have seen probabilistic automata with transition (and emission/reward) maps of the following forms.

name	map
Markov model	$X \longrightarrow \mathcal{D}(X)$
Hidden Markov model	$X \longrightarrow \mathcal{D}(X) \times \mathcal{D}(Y)$
Markov model with output	$X \longrightarrow \mathcal{D}(X + A)$
Markov model with rewards	$X \longrightarrow \mathcal{D}(X) \times \mathbb{R}$
Markov decision process	$X \longrightarrow \mathcal{D}(X)^A \times \mathbb{R}$

The common pattern for the maps in this table is that they have the form $X \to F(X)$, for a suitable functor $F\colon \mathbf{Sets} \to \mathbf{Sets}$. This functor captures the kind of transitions involved, and thus also the kind of computation.

Such a map $X \to F(X)$ is called a *coalgebra* of the functor F. It turns out that a significant part of the theory of automata, transition systems and dynamical systems can be handled at an abstract, generic level in terms of such coalgebras, see *e.g.* Rutten (2000); Adámek (2005); Jacobs (2016a) for an overview. This coalgebraic abstraction provides clarification and makes the most essential aspects explicit. We sketch this coalgebraic approach, but we restrict ourselves to 'probabilistic' coalgebras, where the functor F involves the distribution functor $\mathcal{D}$. A special role is played by *final* coalgebras. They are final objects in the category of coalgebas.

Definition 10.4.1 Let $\mathbb{C}$ be category with an object A.

1 This A is called *initial* in $\mathbb{C}$ if for each object X there is precisely one arrow $A \to X$ in $\mathbb{C}$.
2 Dually, A is *final* if for each object X there is precisely one $X \to A$.

Usually one writes an initial object as 0 and a final object as 1. Such objects need not exist, but if they do, they are determined up-to-isomorphism.

In **Sets** any singleton set is final, and so in particular the set that we use as $1 = \{0\}$. Obviously, for each set X there is only one function $X \to 1$, namely the function that maps each element of X to the sole element 0 of 1. The empty set 0 (or) is initial in **Sets**, since for each set X there is only one function $0 \to X$, namely the 'empty' function.

We shall first describe what a coalgebra is in general, and then identify a relevant class of functors.

Definition 10.4.2 Let $F\colon \mathbb{C} \to \mathbb{C}$ be an arbitrary functor, from a category $\mathbb{C}$ to itself. A *coalgebra* for F is a pair (X, c) where X is an object of $\mathbb{C}$ and c is a map in $\mathbb{C}$ of the form $c\colon X \to F(X)$ in $\mathbb{C}$. We often call X the carrier or space (of positions) of the coalgebra, and c the transition map.

A *homomorphism* (or map) of coalgebras from $X \xrightarrow{c} F(X)$ to $Y \xrightarrow{d} F(Y)$ is a morphism $f\colon X \to Y$ in $\mathbb{C}$ which makes the following diagram commute:

$$\begin{tikzcd} F(X) \arrow[r, "F(f)"] & F(Y) \\ X \arrow[u, "c"] \arrow[r, "f"] & Y \arrow[u, "d"'] \end{tikzcd}$$

This yields a category of F-coalgebra, written as $\mathbf{CoAlg}(F)$. It comes with an obvious functor $\mathbf{CoAlg}(F) \to \mathbb{C}$ which sends a coalgebra to its underlying carrier.

Coalgebras will be used as abstractions of automata, or more generally as abstract systems Rutten (2000); Jacobs (2016a). A coalgebra is the dual of an algebra. For a functor F as above, an *algebra* is a map of the form $a\colon F(X) \to X$. Algebras can be used to captures the operations of certain algebraic structures. For instance, the unit and multiplication operations of a monoid M can be described jointly as an algebra $1 + (M \times M) \to M$. Such a description as algebra of a functor does not capture the (monoid) equations, only the types of the operation. Algebras of the functor F, together with algebra homomorphisms that commute with them, form a category which we write as $\mathbf{Alg}(F)$.

Of special interest are *final* coalgebras and *initial* algebras. They are final objects in categories $\mathbf{CoAlg}(F)$ and initial objects in $\mathbf{Alg}(F)$. Explicitly, for coalgebras: a coalgebra $\zeta\colon Z \to F(Z)$ is final if for any coalgebra $c\colon X \to F(X)$ there is a unique homomorphism of coalgebras $f\colon X \to Z$ in:

$$\begin{tikzcd} F(X) \arrow[r, dashed, "F(f)"] & F(Z) \\ X \arrow[u, "c"] \arrow[r, dashed, "f"] & Z \arrow[u, "\zeta"'] \end{tikzcd}$$

The dashed arrows are used to indicate uniqueness. A basic result is Lambek's Lemma, see *e.g.* Lambek (1968) (or Jacobs (2016a)). It says that if $\zeta\colon Z \to F(Z)$ is a final coalgebra, then the map ζ is an isomorphism, forming a fixed point $\zeta\colon Z \xrightarrow{\cong} F(Z)$ of the functor F. Dually, each initial algebra $\alpha\colon F(A) \to A$ is an isomorphism $F(A) \cong A$.

In addition, the following notion of corecursive algebra is relevant in relation to probabilistic automata / coalgebras, see also Eppendahl (1999); Capretta et al. (2009). We have already seen such situations in Propositions 10.2.6 and 10.3.6, but we shall encounter several more instances below.

Definition 10.4.3 Let $a\colon F(Y) \to Y$ be an algebra for a functor $F\colon \mathbb{C} \to \mathbb{C}$. This algebra is called *corecursive* if for each coalgebra $c\colon X \to F(X)$ there is a unique morphism $f\colon X \to Y$ making the following diagram commute.

$$\begin{tikzcd} F(X) \arrow[r, "F(f)"] & F(Y) \arrow[d, "a"] \\ X \arrow[u, "c"] \arrow[r, "f"] & Y \end{tikzcd}$$

In the above diagram the algebra a re-assembles what the coalgebra c takes apart. It is easy to see that if $\zeta\colon Z \to F(Z)$ is a final coalgebra, then its inverse $\zeta^{-1}\colon F(Z) \to Z$ is a corecursive algebra. But not all corecursive algebras arise in this way, simply because not all of them are isomorphisms.

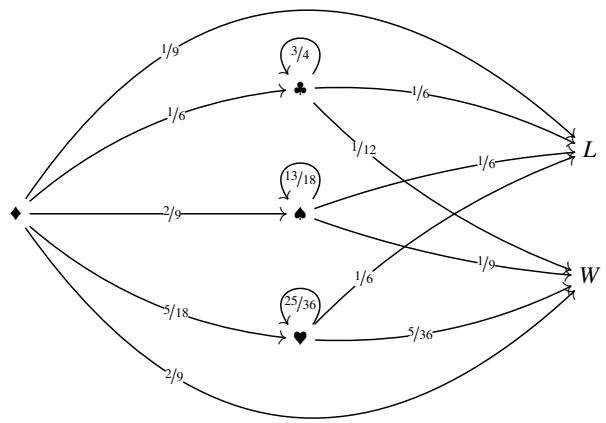


Figure 10.5 A simplified version of the game of craps from Figure 10.2, copied from (Baier and Katoen, 2008, Fig. 10.2).

We first illustrate how morphism of coalgebras can connect different automata — of the same kind, that is, of the same functor.

Example 10.4.4 Recall the game of craps from Subsection 10.1.3 and its transition graph described in Figure 10.2. An alternative, smaller transition graph is given in Figure 10.5, based on (Baier and Katoen, 2008, Fig. 10.2).

Let's write the two relevant sample spaces and outputs as:

$$X = \{1, 4, 5, 6, 8, 9, 10\} \quad \text{and} \quad Y = \{\diamondsuit, \clubsuit, \spadesuit, \heartsuit\} \quad \text{and} \quad A = \{L, W\}.$$

The craps version in Figure 10.2 is described as a coalgebra $c \colon X \to \mathcal{D}(X + A)$, whereas the one in Figure 10.5 will be denoted by $e \colon Y \to \mathcal{D}(Y + A)$, where:

$$\begin{aligned}
e(\diamondsuit) &= \tfrac{1}{6}|\clubsuit\rangle + \tfrac{2}{9}|\spadesuit\rangle + \tfrac{5}{18}|\heartsuit\rangle + \tfrac{1}{9}|L\rangle + \tfrac{2}{9}|W\rangle \\
e(\clubsuit) &= \tfrac{3}{4}|\clubsuit\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{12}|W\rangle \\
e(\spadesuit) &= \tfrac{13}{18}|\spadesuit\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{9}|W\rangle \\
e(\heartsuit) &= \tfrac{25}{36}|\heartsuit\rangle + \tfrac{1}{6}|L\rangle + \tfrac{5}{36}|W\rangle.
\end{aligned}$$

Both c and e are coalgebras for the functor $F = \mathcal{D}(\,-\,+A)$ on **Sets**. We claim that there is a homomorphism of coalgebras $f \colon X \to Y$, namely:

$$f(1) = \diamondsuit \qquad \begin{aligned} f(4) &= f(10) = \clubsuit \\ f(5) &= f(9) = \spadesuit \\ f(6) &= f(8) = \heartsuit. \end{aligned} \tag{10.11}$$

One can then prove the equation that defines a morphism of coalgebras, see Deftinition 10.4.2:

$$\mathcal{D}(f + id_A) \circ c = e \circ f. \tag{10.12}$$

For instance, for position $1 \in X$ one has:

$$\begin{aligned}
&\big(\mathcal{D}(f + id_A) \circ c\big)(1) \\
&= \mathcal{D}(f + id_A)\big(\tfrac{1}{12}\big|4\rangle + \tfrac{1}{9}\big|5\rangle + \tfrac{5}{36}\big|6\rangle + \tfrac{5}{36}\big|8\rangle + \tfrac{1}{9}\big|9\rangle + \tfrac{1}{12}\big|10\rangle + \tfrac{2}{9}\big|id_A(W)\rangle + \tfrac{1}{9}\big|id_A(L)\rangle\big) \\
&= \tfrac{1}{12}\big|f(4)\rangle + \tfrac{1}{9}\big|f(5)\rangle + \tfrac{5}{36}\big|f(6)\rangle + \tfrac{5}{36}\big|f(8)\rangle + \tfrac{1}{9}\big|f(9)\rangle + \tfrac{1}{12}\big|f(10)\rangle + \tfrac{2}{9}\big|W\rangle + \tfrac{1}{9}\big|L\rangle \\
&= \tfrac{1}{12}\big|\clubsuit\rangle + \tfrac{1}{9}\big|\spadesuit\rangle + \tfrac{5}{36}\big|\heartsuit\rangle + \tfrac{5}{36}\big|\heartsuit\rangle + \tfrac{1}{9}\big|\spadesuit\rangle + \tfrac{1}{12}\big|\clubsuit\rangle + \tfrac{2}{9}\big|W\rangle + \tfrac{1}{9}\big|L\rangle \\
&= \tfrac{1}{6}\big|\clubsuit\rangle + \tfrac{2}{9}\big|\spadesuit\rangle + \tfrac{5}{18}\big|\heartsuit\rangle + \tfrac{2}{9}\big|W\rangle + \tfrac{1}{9}\big|L\rangle \;=\; e(\diamondsuit) \;=\; e(f(1)).
\end{aligned}$$

The same equations holds for positions $4, 5, 6, 8, 9, 10$ instead of 1.

In the current setting we shall be using a restricted class of functors from **Sets** to **Sets**, built up inductively.

Definition 10.4.5 1 The class of *simple polynomial functors* **Sets** $\to$ **Sets** contains the identity functor and the constant functor $X \mapsto A$, for each set A, and is closed under binary products and coproducts, in the following way:

- if F_1, F_2 are simple polynomial, then so are $F_1 \times F_2$ and $F_1 + F_2$ given by:

$$(F_1 \times F_2)(X) \coloneqq F_1(X) \times F_2(X) \quad \text{and} \quad (F_1 + F_2)(X) \coloneqq F_1(X) + F_2(X).$$

2 The class of *probabilistic polynomial functors* is defined as in the previous point, with the addition that the distribution functor $\mathcal{D}$ is included. If it suits, we also allow $\mathcal{D}_\infty$ instead of $\mathcal{D}$.

3 A coalgebra of a probabilistic polynomial functors will be called a *probabilistic automaton*.

We can now see that the three Markov models in the table in the beginning of this section are all coalgebras of probabilistic polynomial functors, namely of:

$$X \mapsto \mathcal{D}(X) \qquad X \mapsto \mathcal{D}(X) \times \mathcal{D}(Y) \qquad X \mapsto \mathcal{D}(X + A).$$

For a finite set A, say with n elements, we can identify a set of functions X^A from A to X with the n-fold cartesian product $X^n = X \times \cdots \times X$. With this understanding, a Markov decision process $X \to \mathcal{D}(X)^A \times \mathbb{R}$ is also a coalgebra of a probabilistic polynomial functor — and thus a probabilistic automaton — when the set A of actions is finite — see Definition 10.3.1.

We finish with deterministic probabilistic automata as coalgebras.

Example 10.4.6 A deterministic probabilistic automaton can be described as a coalgebra $c\colon X \to \mathcal{D}(X \times A)$, for some set of labels A. From each position $x \in X$ one can make transitions to (x', a) with a certain probability $r = c(x)(x', a)$. This can be written as a labelled transition, with probability, as in:

$$x \xrightarrow{r,a} x' \qquad \text{or in:} \qquad x \xrightarrow{r\,/\,a} x'.$$

Notice that there may be multiple successor positions, say x' and x'' with the same label a, namely with both (x', a) and (x'', a) are in the support of the distribution $c(x) \in \mathcal{D}(A \times X)$.

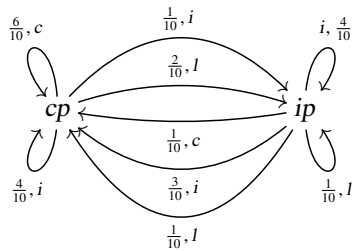
As concrete example we re-describe the soft drink example from Subsection 8.9.2. There is forms a hidden Markov model of the form:

$$X \xrightarrow{\langle t,e\rangle} \mathcal{D}(X) \times \mathcal{D}(A) \qquad \text{for} \qquad \begin{cases} X = \{cp, ip\} \\ A = \{c, i, l\}. \end{cases}$$

Here we massage it into a different coalgebra, making it a deterministic probabilistic automaton:

$$\begin{array}{c} \mathcal{D}(X \times A) \\ {\scriptstyle c := \otimes \circ \langle t, e \rangle} \uparrow \\ X \end{array} \qquad \text{where} \qquad \begin{cases} c(cp) \;=\; \frac{6}{10}|cp, c\rangle + \frac{1}{10}|cp, l\rangle \\ \qquad\qquad + \frac{1}{10}|ip, i\rangle + \frac{2}{10}|ip, l\rangle \\ c(ip) \;=\; \frac{1}{10}|cp, c\rangle + \frac{3}{10}|cp, i\rangle + \frac{1}{10}|cp, l\rangle \\ \qquad\qquad + \frac{4}{10}|ip, i\rangle + \frac{1}{10}|ip, l\rangle. \end{cases}$$

We can write this coalgebra as a transition diagram:



We can now define iterations $X \to \mathcal{D}(X \times A^n)$ for $n \geq 1$. The case $n = 1$ uses simply the coalgebra c itself, and subsequent steps can be obtained as:

$$X \longrightarrow \mathcal{D}(X \times A^n) \xrightarrow{\mathcal{D}(c \otimes id)} \mathcal{D}\Big(\mathcal{D}(X \times A \times A^n)\Big) \xrightarrow{flat} \mathcal{D}(X \times A^{n+1}) \tag{10.13}$$

By discarding the X in the output we get distributions of traces (of length n). For instance, for $n = 2$, starting in position ip gives:

$$\begin{aligned} &\tfrac{171}{1000}\big|c, c\rangle + \tfrac{9}{250}\big|c, i\rangle + \tfrac{157}{1000}\big|c, l\rangle + \tfrac{27}{500}\big|i, c\rangle + \tfrac{9}{1000}\big|i, i\rangle \\ &\quad + \tfrac{19}{1000}\big|i, l\rangle + \tfrac{51}{200}\big|l, c\rangle + \tfrac{9}{200}\big|l, i\rangle + \tfrac{61}{500}\big|l, l\rangle \end{aligned}$$

This will be described more systematically below.

In order to continue with probabilities of sequences of labels, let's return to a general deterministic probabilistic automaton of the form $c\colon X \to \mathcal{D}(X \times A)$. The probability of seeing a label $a \in A$ after one step starting in position x can be described as:

$$c(x) \models \mathbf{1} \otimes \mathbf{1}_a \;=\; c(x)[0,1] \models \mathbf{1}_a \;=\; c(x)[0,1](a) \;=\; \sum_{x' \in X} c(x)(x', a).$$

Next, the probability of seeing a after the first and b after the second step can be obtained also in various ways:

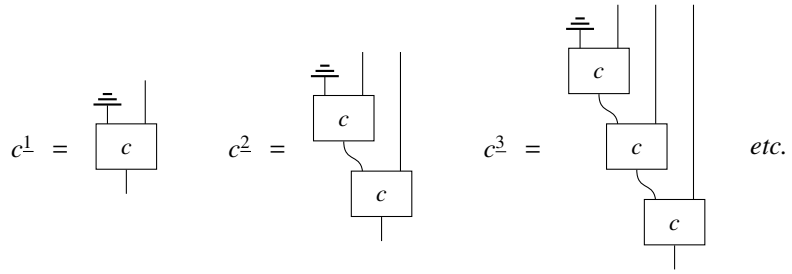
$$\begin{aligned} ((c \otimes id) \circ c)(x) \models \mathbf{1} \otimes \mathbf{1}_b \otimes \mathbf{1}_a \;&=\; (c \otimes id)_*(c(x)) \models \mathbf{1} \otimes \mathbf{1}_b \otimes \mathbf{1}_a \\ &=\; c(x) \models (c \otimes id)^*(\mathbf{1} \otimes \mathbf{1}_b \otimes \mathbf{1}_a) \;=\; c(x) \models c^*(\mathbf{1} \otimes \mathbf{1}_b) \otimes \mathbf{1}_a. \end{aligned}$$

Thus, the probability in position x of seeing a word $[a_1, \ldots, a_n]$ of length $n > 0$ can be obtained via computing n steps forward and taking the validity of the predicate $\mathbf{1} \otimes \mathbf{1}_{a_n} \otimes \cdots \otimes \mathbf{1}_{a_1}$. Alternatively, this probability can be obtained by pulling back this predicate appropriately and computing its validity in $c(x)$.

We shall concentrate on the 'forward' approach and now define n-fold 'iterations' of an automaton $c\colon X \rightsquigarrow X \times A$, giving a channel $c^{\underline{n}}\colon X \rightsquigarrow A^n$. It is obtained basically by discarding the X in the output in (10.13). We now define it explicitly, by induction, for $n \geq 1$.

$$c^{\underline{1}} := \pi_2 \circ c \qquad \text{and} \qquad c^{\underline{n+1}} := (c^{\underline{n}} \otimes id_A) \circ c.$$

Thus, as string diagrams, we can write:



The distribution $c^{\underline{n}}(x) \in \mathcal{D}(A^n)$ gives the probability of generating n-letter words, starting at position x. The question arises whether we can combine these distributions to a single distribution on the set $A^+ := \bigcup_{n \geq 1} A^n$ of non-empty words. Adding up all probabilities does not work, so we shall need a scaling factor.

In general, a subset of words is called a *language*. A distribution on words is called a *probabilistic language*. Such distributions typically have infinite support, and so we will use the distribution functor $\mathcal{D}_\infty$ from Definition 2.1.6, instead of $\mathcal{D}$ — which allows only finite support. Here we shall be interested in probabilistic languages of non-empty words only, that is, in distributions $\omega \in \mathcal{D}_\infty(A^+)$.

We switch to the functor $F(X) = \mathcal{D}_\infty(X \times A)$, now with $\mathcal{D}_\infty$ instead of $\mathcal{D}$, and use coalgebras $c \colon X \to F(X)$ as automata. We also use the following F-algebra with the set $\mathcal{D}_\infty(A^+)$ as carrier, which we call pl for 'probabilistic language'.

$$\begin{array}{c} \mathcal{D}_\infty(\mathcal{D}_\infty(A^+) \times A) \\ \big\downarrow{\scriptstyle pl} \\ \mathcal{D}_\infty(A^+) \end{array} \qquad \begin{cases} pl(\Omega)([a]) \;=\; \frac{1}{2} \cdot \displaystyle\sum_{\omega \in \mathcal{D}_\infty(A^+)} \Omega(\omega, a) \\ pl(\Omega)(a \cdot \sigma) \;=\; \frac{1}{2} \cdot \displaystyle\sum_{\omega \in \mathcal{D}_\infty(A^+)} \Omega(\omega, a) \cdot \omega(\sigma). \end{cases} \tag{10.14}$$

We need to rescale with one half in order to obtain a distribution, like in Exercise 2.1.17. Indeed, probabilities in $pl(\Omega)$ add up to one:

$$\begin{aligned} \sum_{\sigma \in A^+} pl(\Omega)(\sigma) &= \sum_{a \in A} pl(\Omega)([a]) + \sum_{a \in A,\, \sigma \in A^+} pl(\Omega)(a \cdot \sigma) \\ &= \tfrac{1}{2} \cdot \sum_{a \in A,\, \omega \in \mathcal{D}_\infty(A^+)} \Omega(\omega, a) \;+\; \tfrac{1}{2} \cdot \sum_{a \in A,\, \sigma \in A^+,\, \omega \in \mathcal{D}_\infty(A^+)} \Omega(\omega, a) \cdot \omega(\sigma) \\ &= \tfrac{1}{2} \;+\; \tfrac{1}{2} \cdot \sum_{a \in A,\, \omega \in \mathcal{D}_\infty(A^+)} \Omega(\omega, a) \cdot \sum_{\sigma \in A^+} \omega(\sigma) \\ &= 1. \end{aligned}$$

Proposition 10.4.7 *The algebra* $pl \colon \mathcal{D}_\infty(\mathcal{D}_\infty(A^+) \times A) \to \mathcal{D}_\infty(A^+)$ *in* (10.14) *is* corecursive. *This means that for each coalgebra* $c \colon X \to \mathcal{D}_\infty(X \times A)$ *there is a unique map* $f \colon X \to \mathcal{D}_\infty(A^+)$ *in:*

$$\begin{array}{ccc} \mathcal{D}_\infty(X \times A) & \xrightarrow{\mathcal{D}_\infty(f \times id_A)} & \mathcal{D}_\infty(\mathcal{D}_\infty(A^+) \times A) \\ {\scriptstyle c}\big\uparrow & & \big\downarrow{\scriptstyle pl} \\ X & \xrightarrow{\quad f \quad} & \mathcal{D}_\infty(A^+) \end{array}$$

Proof. We define $f\colon X \to \mathcal{D}_\infty(A^+)$ as:

$$f(x)([a_1,\dots,a_n]) \coloneqq \frac{c^{\underline{n}}(x)(a_1,\dots,a_n)}{2^n}.$$

This yields a well-defined distribution, since:

$$\begin{aligned}\sum_{\sigma\in A^+} f(x)(\sigma) &= \sum_{n\geq 1}\sum_{\sigma\in A^n} f(x)(\sigma) = \sum_{n\geq 1}\frac{\sum_{\sigma\in A^n} c^{\underline{n}}(x)(\sigma)}{2^n} = \sum_{n\geq 1}\frac{1}{2^n}\\ &= \frac{1}{2}\cdot\sum_{n\geq 0}\frac{1}{2^n} \overset{(10.1)}{=} \frac{1}{2}\cdot\frac{1}{1-{}^1\!/_2} = 1.\end{aligned}$$

Next we show that f makes the above diagram commute. First, for singleton words we have:

$$\begin{aligned}\Big(pl\circ\mathcal{D}(f\times id_A)\circ c\Big)(x)([a]) &= pl\Big(\mathcal{D}(f\times id)(c(x))\Big)([a])\\ &= \tfrac{1}{2}\cdot\sum_{\omega\in\mathcal{D}_\infty(A^+)}\mathcal{D}(f\times id)(c(x))(\omega,a)\\ &= \tfrac{1}{2}\cdot\sum_{\omega\in\mathcal{D}_\infty(A^+)}\sum_{x'\in f^{-1}(\omega)} c(x)(x',a)\\ &= \tfrac{1}{2}\cdot\sum_{x'\in X} c(x)(x',a) = \tfrac{1}{2}\cdot c^{\underline{1}}(x)(a) = f(x)([a]).\end{aligned}$$

Similarly, for words of length $n+1>1$,

$$\begin{aligned}\Big(pl\circ\mathcal{D}(f\times id)\circ c\Big)(x)(a\cdot\sigma) &= pl\Big(\mathcal{D}(f\times id)(c(x))\Big)(a\cdot\sigma)\\ &= \tfrac{1}{2}\cdot\sum_{\omega\in\mathcal{D}_\infty(A^+)}\mathcal{D}(f\times id)(c(x))(\omega,a)\cdot\omega(\sigma)\\ &= \tfrac{1}{2}\cdot\sum_{\omega\in\mathcal{D}_\infty(A^+)}\sum_{x'\in f^{-1}(\omega)} c(x)(x',a)\cdot\omega(\sigma)\\ &= \tfrac{1}{2}\cdot\sum_{x'\in X} c(x)(x',a)\cdot f(x')(\sigma)\\ &\overset{(\mathrm{IH})}{=} \tfrac{1}{2}\cdot\sum_{x'\in X} c(x)(x',a)\cdot\frac{c^{\underline{n}}(x')(\sigma)}{2^n}\\ &= \frac{1}{2^{n+1}}\cdot((c^{\underline{n}}\otimes id)\circ c)(x)(\sigma,a)\\ &= \frac{1}{2^{n+1}}\cdot c^{\underline{n+1}}(x)(\sigma,a) = f(x)(a\cdot\sigma).\end{aligned}$$

Uniqueness is left as an exercise to the reader. ❑

We conclude by observing that in the literature a coalgebra of the form $X\to\mathcal{D}(X)^A$ is called *reactive* because the elements of A can be seen as actions coming from the environment. A coalgebra of the form $X\to\mathcal{D}(X\times A)$ is called *generative* since it can be used to generate a probabilistic language — as we have seen in Proposition 10.4.7. One can combine reactive and generative, and even outputs, in a coalgebra of the form $X\to\mathcal{D}(X\times B+C)^A$. This illustraties how the functor of a coalgebra determines the type of computation involved.

As an aside, these differences between reactive and generative do not make sense in non-deterministic computation, since there one has an isomorphism $\mathcal{P}(X)^A\cong\mathcal{P}(X\times A)$, see (1.21).

Exercises

10.4.1 Prove the coalgebra homomorphism equation (10.12) for the positions $4, 5, 6, 8, 9, 10 \in X$ too.

10.4.2 Consider the small version of craps in Example 10.4.4.

1 Prove that there are limit equations for the 'inner positions':

$$\begin{aligned} \lim_{n\to\infty} e^n(\clubsuit) &= \tfrac{2}{3}|L\rangle + \tfrac{1}{3}|W\rangle \\ \lim_{n\to\infty} e^n(\spadesuit) &= \tfrac{3}{5}|L\rangle + \tfrac{2}{5}|W\rangle \\ \lim_{n\to\infty} e^n(\heartsuit) &= \tfrac{6}{11}|L\rangle + \tfrac{5}{11}|W\rangle. \end{aligned}$$

2 Deduce, via a convex sum of distributions, that for the initial position one has, as in Example 10.1.3,

$$\lim_{n\to\infty} d^n(\diamondsuit) = \tfrac{251}{495}|L\rangle + \tfrac{244}{495}|W\rangle.$$

10.4.3 Consider a Markov model with 'input and output' $c\colon X \to \mathcal{D}(X+A)^B$, where B is a finite set. We illustrate how it can be extended to $c^\star\colon X \to \mathcal{D}(X+A)^{B^\star}$, using sequences of inputs.

1 Recall the product-exponent correspondences (1.7) and use them to obtain:

$$\begin{array}{c} X \longrightarrow \mathcal{D}(X+A)^B \\ \hline\hline X \times B \longrightarrow \mathcal{D}(X+A) \\ \hline\hline B \longrightarrow \mathcal{D}(X+A)^X \end{array}$$

Describe the latter map concretely.

2 Notice that the set $\mathcal{D}(X+A)^X$ on the right-hand side in the last line is a monoid, via composition of MMOs as described in (10.4). Use that $\mathcal{L}(B) = B^\star$ is the free monoid on B, see Proposition 1.4.3, to obtain a function $B^\star \to \mathcal{D}(X+A)^X$.

3 Translate the latter function back to a function $c^\star\colon X \to \mathcal{D}(X+A)^{B^\star}$, again using the correspondences (1.7), and describe this Markov model concretely.

10.4.4 Let $F\colon \mathbb{C} \to \mathbb{C}$ be a functor on a category $\mathbb{C}$ with coproducts. Show that the category $\mathbf{CoAlg}(F)$ then also has coproducts.

10.4.5 Recall that we write $\mathbf{0}$ for the empty set $\emptyset$ and $\mathbf{1} = \{0\}$ for a singleton set.

1 Show that $\mathbf{0}$ is both initial and final the category $\mathbf{Chan}(\mathcal{P})$ of the powerset monad $\mathcal{P}$.

2 Check that $\mathbf{0}$ is initial in $\mathbf{Chan}(\mathcal{D})$ and that $\mathbf{1}$ is final in $\mathbf{Chan}(\mathcal{D})$.

10.4.6 For a fixed set A take the functor $F(X) := A \times X$. Consider the set $A^{\mathbb{N}}$ of infinite sequences of elements of A as a head-and-tail F-coalgebra:

$$A^{\mathbb{N}} \xrightarrow{\langle hd, tl\rangle} A \times A^{\mathbb{N}} \qquad \text{with} \qquad \begin{cases} hd(\sigma) = \sigma(0) \\ tl(\sigma)(n) = \sigma(n+1). \end{cases}$$

Show that this is the final F-coalgebra, that is, the final object in the category $\mathbf{CoAlg}(F)$.

10.4.7 (From Feys et al. (2018)) Let $\langle t, r\rangle \colon X \to \mathcal{D}(X) \times \mathbb{R}$ be a Markov model with rewards. Adapt it to a pair:

$$\mathcal{D}(X) \xrightarrow{\langle t', r'\rangle} \mathcal{D}(X) \times \mathbb{R} \qquad \text{with} \qquad \begin{cases} t'(\omega) \coloneqq t_*(\omega) \\ r'(\omega) \coloneqq \omega \models r. \end{cases}$$

This pair $\langle t', r'\rangle$ forms a coalgebra for the functor $(-) \times \mathbb{R}$. From the previous exercise we know that its final coalgebra is $\mathbb{R}^{\mathbb{N}}$. Check that the resulting unique coalgebra homomorphism $h \colon \mathcal{D}(X) \to \mathbb{R}^{\mathbb{N}}$ is given by successive rewards:

$$h(\omega)(n) = (t^n)_*(\omega) \models r = \omega \models (t^n)^*(r).$$

10.4.8 Prove that the function $f \colon X \to \mathcal{D}_\infty(A^+)$ as defined in the beginning of the proof of Proposition 10.4.7 satisfies the required uniqueness property.

10.4.9 We sketch a non-deterministic analogue of Proposition 10.4.7. Fix a set A, considered as alphabet for the set $\mathcal{P}(A^+)$ of languages of non-empty words over A. We use the functor $F \colon \mathbf{Sets} \to \mathbf{Sets}$ given by $F(X) = \mathcal{P}(X \times A)$, with its 'non-deterministic language' algebra:

$$\begin{array}{c} \mathcal{P}(\mathcal{P}(A^+) \times A) \\ \Big\downarrow{\scriptstyle ndl} \\ \mathcal{P}(A^+) \end{array} \qquad \text{where} \qquad \begin{cases} [a] \in ndl(U) \iff \exists L.\, (L, a) \in U \\ a \cdot \sigma \in ndl(U) \iff \exists L.\, (L, a) \in U \\ \qquad\qquad\qquad\qquad \text{and } \sigma \in L. \end{cases}$$

Prove that this algebra is corecursive, for the functor F.

10.4.10 (See (Jacobs et al., 2021b, Prop. 4.3)) Consider a functor $F \colon \mathbf{Sets} \to \mathbf{Sets}$ with two sets A, B and an algebra $\alpha \colon F(A) \to A$. We define a new functor $F_{A,B} \colon \mathbf{Sets} \to \mathbf{Sets}$ as $F_{A,B}(X) \coloneqq F(X)^B \times A$. We equip it with an algebra $\beta \colon F_{A,B}\big(A^{B^\star}\big) \to A^{B^\star}$ given on $g \in F(A^{B^\star})^B$ and $a \in A$ as:

$$\beta(g, a)([]) \coloneqq a \qquad \text{and} \qquad \beta(g, a)(b \cdot \sigma) \coloneqq \alpha\Big(F(ev_\sigma)(g(b))\Big),$$

where the function $ev_\sigma \colon A^{B^\star} \to A$ is given by $ev_\sigma(f) = f(\sigma)$. Show that β is a corecursive algebra.

This setting can be instantiated, for instance, with $F = \mathcal{D}$ and $A = [0, 1]$ with mean map $\mathcal{D}([0, 1]) \to [0, 1]$.

10.5 Invariants and nexttime operators

Now that we have seen several examples of probabilistic automata, and also a general formalism to describe them as coalgebras of probabilistic polynomial functors, we turn to predicates, in order to reason about such automata. This section concentrates on invariants and the next one on bisimulations. Invariants are predicates which are closed under transitions. This means that if they hold in a certain position, they continue to hold in all successor positions. Since our predicates are 'fuzzy' and not necessarily 'sharp', the definition of invariant involves an order. In a similar manner, bisimulations are relations that are closed under transitions.

We shall proceed in a principled manner, deriving logical structure from the structure of the polynomial functor that captures the type of probabilistic automaton. We start by organising predicates in a suitable category, and to 'lift' these functors to this category of predicates.

Definition 10.5.1 We shall write **FPred** for the category of (fuzzy) predicates, with pair (X, p) as objects, where $p\colon X \to [0,1]$ is a predicate on X. A morphism $f\colon (X,p) \to (Y,q)$ in **FPred** is a function $f\colon X \to Y$ with $p \leq q \circ f$, pointwise, using the standard order on the unit interval $[0,1] \subseteq \mathbb{R}$.

There is an obvious functor **FPred** $\to$ **Sets** that sends a predicate (X,p) to its underlying set X.

Sometimes we loosely say: "let $p\colon X \to [0,1]$ be an object of **FPred**" where the object really is of the form (X,p). This functor **FPred** $\to$ **Sets** is an example of a *fibration*, a fundamental structure in categorical logic, see Jacobs (1999) for more information.

Now that we have a category of predicates, we can see that various operations on predicates, that we have seen and used before, can be represented as functors.

Lemma 10.5.2 *1 There are truth and falsum functors:*

$$\begin{array}{ccc} \mathbf{Sets} & \xrightarrow{\top} & \mathbf{FPred} \\ X & \longmapsto & \big(X \xrightarrow{\mathbf{1}} [0,1]\big) \end{array} \qquad \begin{array}{ccc} \mathbf{Sets} & \xrightarrow{\bot} & \mathbf{FPred} \\ X & \longmapsto & \big(X \xrightarrow{\mathbf{0}} [0,1]\big) \end{array}$$

The objects $\top(\mathbf{1})$ and $\bot(\mathbf{0})$ are final and initial in the category **FPred**.

2 The category **FPred** *has coproducts, via cotupling of predicates:*

$$\big(X \xrightarrow{p} [0,1]\big) + \big(Y \xrightarrow{q} [0,1]\big) := \big(X + Y \xrightarrow{[p,q]} [0,1]\big).$$

This yields a commuting diagram:

$$\begin{array}{ccc} \mathbf{FPred} \times \mathbf{FPred} & \xrightarrow{+} & \mathbf{FPred} \\ \downarrow & & \downarrow \\ \mathbf{Sets} \times \mathbf{Sets} & \xrightarrow{+} & \mathbf{Sets} \end{array}$$

3 There is also a parallel conjunction (tensor) on the category **FPred**, *via:*

$$\big(X \xrightarrow{p} [0,1]\big) \otimes \big(Y \xrightarrow{q} [0,1]\big) := \big(X \times Y \xrightarrow{p \otimes q} [0,1]\big).$$

The unit for $\otimes$ is the final object $\top(\mathbf{1})$ in **FPred**, *since in* **FPred**,

$$\top(\mathbf{1}) \otimes (X,p) = (\mathbf{1} \times X, \mathbf{1} \otimes p) \cong (X,p) \cong (X \times \mathbf{1}, p \otimes \mathbf{1}) = (X,p) \otimes \top(\mathbf{1}).$$

We now have a commuting diagram:

$$\begin{array}{ccc} \mathbf{FPred} \times \mathbf{FPred} & \xrightarrow{\otimes} & \mathbf{FPred} \\ \downarrow & & \downarrow \\ \mathbf{Sets} \times \mathbf{Sets} & \xrightarrow{\times} & \mathbf{Sets} \end{array}$$

Proof. 1 On morphisms one defines $\top(f) = f$ and $\bot(f) = f$. This is well-defined, since $\mathbf{1} \circ f = \mathbf{1}$ and $\mathbf{0} \circ f = \mathbf{0}$. For an arbitrary predicate (X,p) there is the unique function $f\colon X \to \mathbf{1} = \{0\}$ in **Sets**. It is a map of predicates $(X,p) \to (\mathbf{1},\mathbf{1})$ since $p(x) \leq 1 = (\mathbf{1} \circ f)(x)$. Clearly, it is the only possible map of predicates $(X,p) \to (\mathbf{1},\mathbf{1})$. Similarly, $\bot(\mathbf{0})$ is an initial object in **FPred**.

2 We follow the formulation of Definition 10.1.1: the coprojections $X \xrightarrow{\kappa_1} X + Y \xleftarrow{\kappa_2} Y$ in **Sets** are also morphisms of predicates in **FPred**:

$$(X,p) \xrightarrow{\kappa_1} (X + Y, [p,q]) \xleftarrow{\kappa_2} (Y,q)$$

since $p \leq p = [p, q] \circ \kappa_1$ and $q \leq [p, q] \circ \kappa_2$.

Let (Z, r) now be an arbitrary predicate in **FPred**, with two morphisms $f\colon (X, p) \to (Z, r)$ and $g\colon (Y, q) \to (Z, r)$. This means that $p(x) \leq r(f(x))$ for $x \in X$ and $q(y) \leq r(g(y))$ for $y \in Y$. The cotuple map $[f, g]\colon X + Y \to Z$ in **Sets** is then also a map of predicates $(X + Y, [p, q]) \to (Z, r)$ since $[p, q] \leq r \circ [f, g]$. This follows by a simple case distinction for elements of $X + Y$:

$$\begin{aligned} [p, q](\kappa_1 x) &= p(x) \leq r(f(x)) = (r \circ [f, g])(\kappa_1 x) \\ [p, q](\kappa_2 y) &= q(y) \leq r(g(y)) = (r \circ [f, g])(\kappa_2 y). \end{aligned}$$

3 We concentrate on functoriality. Suppose we have two maps $f_i\colon (X_i, p_i) \to (Y_i, q_i)$ in **FPred**, for $i = 1, 2$, then $f_1 \times f_2\colon X_1 \times X_2 \to Y_1 \times Y_2$ in **Sets** is a morphism $(X_1 \times X_2, p_1 \otimes p_2) \to (Y_1 \times Y_2, q_1 \otimes q_2)$ in **FPred**. Indeed, by assumption we have $p_i(x_i) \leq q_i(f_i(x_i))$ for $x_i \in X_i$ and thus:

$$\begin{aligned} (p_1 \otimes p_2)(x_1, x_2) = p_1(x_1) \cdot p_2(x_2) &\leq q_1(f_1(x_1)) \cdot q_2(f_2(x_2)) \\ &= (q_1 \otimes q_2)(f_1(x_1), f_2(x_2)) \\ &= ((q_1 \otimes q_2) \circ (f_1 \times f_2))(x_1, x_2). \end{aligned}$$

The swap isomorphism $\langle \pi_2, \pi_1 \rangle\colon X_1 \times X_2 \xrightarrow{\cong} X_2 \times X_1$ in **Sets** is also an isomorphism $(X_1, p_1) \otimes (X_2, p_2) \xrightarrow{\cong} (X_2, p_2) \otimes (X_1, p_1)$ in **FPred**. Finally, it is not hard to see that the truth predicate on $\mathbf{1}$ forms the unit object $\top(\mathbf{1})$ for this tensor of predicates in **FPred**. ❑

In a next step we are going to lift the distribution functor $\mathcal{D}\colon \mathbf{Sets} \to \mathbf{Sets}$ to a functor $FP(\mathcal{D})$ on the category **FPred** of fuzzy predicates. This works in such a way that the next diagram commutes.

$$\begin{tikzcd} \mathbf{FPred} \arrow[r, "FP(\mathcal{D})"] \arrow[d] & \mathbf{FPred} \arrow[d] \\ \mathbf{Sets} \arrow[r, "\mathcal{D}"] & \mathbf{Sets} \end{tikzcd} \tag{10.15}$$

Commutation means that for $(X, p) \in \mathbf{FPred}$, one must have $FP(\mathcal{D})(X, p)$ of the form $(\mathcal{D}(X), ?)$, where ? is a predicate $\mathcal{D}(X) \to [0, 1]$. The obvious choice is to use validity $\models$ of predicate p. Indeed, we define:

$$\begin{aligned} FP(\mathcal{D})\big(X \xrightarrow{p} [0, 1]\big) &:= \big(\mathcal{D}(X) \xrightarrow{(-)\models p} [0, 1]\big) \\ FP(\mathcal{D})\big(X \xrightarrow{f} Y\big) &:= \mathcal{D}(f). \end{aligned} \tag{10.16}$$

We need to check that the latter is well-defined: if $p(x) \leq q(f(x))$ for all $x \in X$, then $\omega \models p \leq \mathcal{D}(f)(\omega) \models q$. This follows directly from Lemma 4.1.5, since $\mathcal{D}(f)(\omega) \models q = \omega \models q \circ f \geq \omega \models p$. This lifting $FP(\mathcal{D})\colon \mathbf{FPred} \to \mathbf{FPred}$ is not only a functor, but also a monad, see Exercise 10.5.7.

Definition 10.5.3 Let $F\colon \mathbf{Sets} \to \mathbf{Sets}$ be a probabilistic polynomial functor. We define $FP(F)\colon \mathbf{FPred} \to \mathbf{FPred}$ to be the *predicate lifting* of F that is defined following the structure of F.

- The identity functor id on **Sets** is lifted to the identity functor id on **FPred**, that is: $FP(id) := id$.
- For a set A, the constant functor $X \mapsto A$ on **Sets** is lifted to the constant functor $(X, p) \mapsto \top(A) = (A, \mathbf{1})$, using the truth predicate $\mathbf{1}$ on A.
- The distribution functor $\mathcal{D}$ is lifted to $FP(\mathcal{D})$, as defined in (10.16).

- Coproducts are lifted to coproducts: $FP(F_1 + F_2) := FP(F_1) + FP(F_2)$, see Lemma 10.5.2 (2).
- Products are lifted to parallel conjunctions, via: $FP(F_1 \times F_2) := FP(F_1) \otimes FP(F_2)$, see Lemma 10.5.2 (3).

By construction, the following diagram commutes.

$$\begin{array}{ccc} \mathbf{FPred} & \xrightarrow{FP(F)} & \mathbf{FPred} \\ \downarrow & & \downarrow \\ \mathbf{Sets} & \xrightarrow{F} & \mathbf{Sets} \end{array} \tag{10.17}$$

For a predicate $p\colon X \to [0,1]$ we sometimes write $FP(F)(X,p)$ as a predicate $FP(F)(p)\colon F(X) \to [0,1]$ on $F(X)$.

The category **FPred** of predicates that we use, and the lifting of funtors to this category, can be described as a special case of a more general construction in Bonchi et al. (2018) in terms of quantales (see also Sprunger et al. (2018)). We use it here in a more concrete setting for defining invariants.

Definition 10.5.4 Let $c\colon X \to F(X)$ be a coalgebra of a probabilistic polynomial functor F.

1 We define a *weak nexttime* ⓦ and a *strong nexttime* ⓢ operator on predicates on the space X of the coalgebra, as:

$$ⓦ\, p \;:=\; \Big(X \xrightarrow{c} F(X) \xrightarrow{FP(F)(p)} [0,1]\Big) \qquad \text{and} \qquad ⓢ\, p \;:=\; \Big(ⓦ\, p^{\perp}\Big)^{\perp}.$$

The dependence of these operators on the functor F and on the coalgebra c is left implicit.

2 A predicate p on X is called a *weak invariant* if $p \leq ⓦ\, p$ and a *strong invariant* if $p \leq ⓢ\, p$.

Notice that a weak invariant p for the coalgebra c, as defined above, can be described as a morphism $c\colon p \to FP(F)(p)$ in the category **FPred**, of the lifted functor $FP(F)$.

We like to spare on parentheses and write $ⓦ\, p(x)$ for $(ⓦ(p))(x)$, and similarly for ⓢ.

Example 10.5.5 We illustrate predicate lifting and invariants for some elementary examples.

1 We first take $F(X) = \mathcal{D}(X + A)$, as used in Section 10.1 for Markov models with outputs in the set A. By following the structure of the functor F we see that the associated lifting operation is given by validity of the cotuple predicate: for a predicate $p\colon X \to [0,1]$,

$$\begin{array}{ccc} \mathcal{D}(X+A) & \xrightarrow{FP(F)(p)} & [0,1] \\ \omega & \longmapsto & \omega \models [p, \mathbf{1}]. \end{array} \tag{10.18}$$

The double-negation version is:

$$FP(F)(p^{\perp})^{\perp}(\omega) \;=\; 1 - (\omega \models [p^{\perp}, \mathbf{1}]) \;=\; \omega \models [p^{\perp}, \mathbf{1}]^{\perp} \;=\; \omega \models [p, \mathbf{0}].$$

The last equation holds since cotupling preserves orthosupplement — in fact, all effect module structure — see Exercise 10.5.1.

Let $c\colon X \to \mathcal{D}(X + A) = F(X)$ be an F-coalgebra, that is, a Markov model with outputs in

A. A predicate $p\colon X \to [0,1]$ is a *weak* invariant if $p \leq \textcircled{w}\, p$. This means that for each position $x \in X$,

$$\begin{aligned} p(x) \;\leq\; \textcircled{w}\, p(x) \;=\; FP(F)(p)(c(x)) &\overset{(10.18)}{=} c(x) \models [p, \mathbf{1}] \\ &= c^*([p,\mathbf{1}])(x) \\ &= \sum_{y \in X} c(x)(\kappa_1 y) \cdot p(y) + \sum_{a \in A} c(x)(\kappa_2 a) \\ &= \sum_{y \in X} c(x)(\kappa_1 y) \cdot p(y) + \big(c(x) \models \mathbf{1}_A\big). \end{aligned}$$

In contrast, p is a *strong* invariant if:

$$p(x) \;=\; \textcircled{s}\, p(x) \;=\; c(x) \models [p, \mathbf{0}] \;=\; c^*([p,\mathbf{1}])(x) \;=\; \sum_{y \in X} c(x)(\kappa_1 y) \cdot p(y).$$

The idea of an invariant p is that if it holds in position x, then it also holds in each successor position y of x. This is expressed above in a fuzzy, probabilistic setting, where the value of $p(x)$ is below the combination of all values of p in successor positions y, plus the probability of an output in A. In the strong case one only takes the actual successor positions into account, and not the possible outputs.

2 For the second example we take as functor $F(X) := \mathcal{D}(X) \times \mathcal{D}(X) = \mathcal{D}(X)^2$, so that a coalgebra $X \to F(X)$ involves in each step an input from $2 = \{0,1\}$. The associated lifting turns $p\colon X \to [0,1]$ to $FP(F)(p)\colon \mathcal{D}(X) \times \mathcal{D}(X) \to [0,1]$, where:

$$\begin{aligned} FP(F)(p)(\omega_1, \omega_2) &= \big(FP(\mathcal{D})(p) \otimes FP(\mathcal{D})(p)\big)(\omega_1, \omega_2) \\ &= FP(\mathcal{D})(p)(\omega_1) \cdot FP(\mathcal{D})(p)(\omega_2) \\ &= (\omega_1 \models p) \cdot (\omega_2 \models p) \;=\; \omega_1 \otimes \omega_1 \models p \otimes p. \end{aligned}$$

This predicate p is a weak invariant for a coalgebra $c = \langle c_0, c_1 \rangle\colon X \to \mathcal{D}(X) \times \mathcal{D}(X)$ if for each position x,

$$\begin{aligned} p(x) \;\leq\; \textcircled{w}\, p(x) \;=\; FP(F)(p)(c(x)) &= c(x) \models p \otimes p \\ &= \langle c_0, c_1 \rangle^*(p \otimes p)(x) \\ &= \big((c_0)^*(p) \;\&\; (c_1)^*(p)\big)(x) \\ &= (c_0)^*(p)(x) \cdot (c_1)^*(p)(x) \;=\; (c_0(x) \models p) \cdot (c_1(x) \models p). \end{aligned}$$

This means that p continues to hold along *both* paths/legs c_0 and c_1 of the coalgebra c.

In contrast, p is a strong invariant if:

$$\begin{aligned} p(x) \;\leq\; \textcircled{s}\, p(x) \;=\; c(x) \models (p^\perp \otimes p^\perp)^\perp &\overset{(*)}{=} c(x) \models (p \otimes p) \,\textcircled{\vee}\, (p \otimes p^\perp) \,\textcircled{\vee}\, (p^\perp \otimes p) \\ &= c^*\big((p \otimes p) \,\textcircled{\vee}\, (p \otimes p^\perp) \,\textcircled{\vee}\, (p^\perp \otimes p)\big)(x). \end{aligned}$$

The marked equation $\overset{(*)}{=}$ holds by Excercise 4.4.9. Informally, this invariant inequality $\leq$ says that if p holds in x, then it must continue to hold along *some* path / leg of the coalgebra c. This 'some' is expressed via a sum of tensors $(p \otimes p) \,\textcircled{\vee}\, (p \otimes p^\perp) \,\textcircled{\vee}\, (p^\perp \otimes p)$. In the fuzzy setting in which we are working we need to be careful to count probabilities of options only once. Notice by the way that this sum of tensors equals $(p \otimes \mathbf{1}) \,\textcircled{\vee}\, (p^\perp \otimes p) = (\mathbf{1} \otimes p) \,\textcircled{\vee}\, (p \otimes p^\perp)$.

3 We look at a minor variation of the functor in the previous point and now take $F(X) := \mathcal{D}(X)^2 \times \mathbb{R}$. This is (essentially) the functor for the Markov decision process in Example 10.3.5. We like to make explicit what happens with the additional 'contant' $\mathbb{R}$ occurring in the functor F.

For a predicate p on X we now get $FP(F)(p)\colon \mathcal{D}(X)^2 \times \mathbb{R} \to [0,1]$ given by:

$$FP(F)(p)(\omega_1, \omega_2, r) \;=\; \Big(((-) \models p) \otimes ((-) \models p) \otimes \mathbf{1}\Big)(\omega_1, \omega_2, r) \;=\; (\omega_1 \models p) \cdot (\omega_2 \models p).$$

Hence, for a coalgebra $c = \langle c_0, c_1, c_2 \rangle\colon X \to \mathcal{D}(X) \times \mathcal{D}(X) \times \mathbb{R}$ we get:

$$\begin{aligned}
\textcircled{w}\, p(x) \;&=\; c(x) \models p \otimes p \otimes \mathbf{1} \;=\; c^*(p \otimes p \otimes \mathbf{1})(x) \\
\textcircled{s}\, p(x) \;&=\; c(x) \models (p^\perp \otimes p^\perp \otimes \mathbf{1})^\perp \;=\; c^*\Big((p \otimes p \otimes \mathbf{1}) \,\textcircled{v}\, (p \otimes p^\perp \otimes \mathbf{1}) \,\textcircled{v}\, (p^\perp \otimes p \otimes \mathbf{1})\Big)(x).
\end{aligned}$$

Thus the weak and strong nexttime operators ignore the constant.

We illustrate how to use nexttime operators and invariants, as described above for concrete logical caclulations.

Example 10.5.6 Recall the Markov model with output that we used to obtain a dice via coin flips in Subsection 10.1.2. It involves a coalgebra $f\colon X \to \mathcal{D}(X + A)$ for a set of positions $X = \{x_0, x_1, \dots x_6\}$ and a set of observations $A = \mathit{pips} = \{1, 2, \dots, 6\}$, with transitions as described in Figure 10.1. At the end of Subsection 10.1.2 we calculated the mean number of coin flips needed to get an outcome in *pips*, as mean of a distribution $\rho \in \mathcal{D}_\infty(\mathbb{N})$. At this stage we can also calculate these means in a logical manner.

We are interested in positions which nexttime produce a pip, and thus not a successor state. This is happens via the predicate:

$$\textcircled{w}\,\mathbf{0} \;=\; f^*([\mathbf{0}, \mathbf{1}]) \;=\; \tfrac{1}{2} \cdot \mathbf{1}_{x_3} + 1 \cdot \mathbf{1}_{x_4} + 1 \cdot \mathbf{1}_{x_5} + \tfrac{1}{2} \cdot \mathbf{1}_{x_6}.$$

Thus, starting from initial position x_0 the probability of getting a pip in $n + 1$ steps is $f^n(x_0) \models \textcircled{w}\,\mathbf{0}$. Hence $\rho \in \mathcal{D}_\infty(\mathbb{N})$ is:

$$\rho \;=\; \sum_{n \in \mathbb{N}} \Big(f^n(x_0) \models \textcircled{w}\,\mathbf{0}\Big) \big|n + 1\big\rangle \;=\; \sum_{n \in \mathbb{N}} \frac{3}{4^n} \big|2n + 1\big\rangle.$$

This gives $\mathit{Mean}(\rho) = \sum_n \frac{3}{4^n} \cdot (2n + 1) = \frac{11}{3}$, as computed at the end of Subsection 10.1.2.

Next we look at a more complicated illustration.

Example 10.5.7 The following description is quoted from (McIver and Morgan, 2004, §2.5.1).

> Imagine a gambler who is placing 50/50 bets: that is he expects to win each bet with probability 1/2 and receives twice his bet back if he does. He starts with a bet of 1 unit; if he wins, he has increased his capital by 1. If he loses, however, he doubles his bet to 2 and tries again; winning this time also increases his original capital by 1, since he wins 2 but lost 1 before. If he again loses, he continues with 4 units, and the same applies (might win 4, but lost 1 + 2 before). Eventually he must win, he reasons, in which case he will receive 1 unit more than the total he has already lost; and he will have increased his capital by exactly 1.
>
> ...
>
> In fact the gambler's reasoning is only "partially correct," and the flaw is that he has not assured (successful) termination ...

This situation is formalsed in McIver and Morgan (2004) as a probablistic program. Here we capture it as a 'bet' coalgebra b of the form:

$$\mathbb{Z}^2 \xrightarrow{\;b\;} \mathcal{D}\Big(\mathbb{Z}^2 + \mathbb{Z} + \mathbf{1}\Big).$$

It is a Markov model with output $X \to \mathcal{D}(X + A)$ with $X = \mathbb{Z}^2$ as set of positions and $\mathbb{Z} + \mathbf{1}$ as set

of outputs. Here we use $\mathbf{1}$ as singleton set $\mathbf{1} = \{*\}$. We avoid $\mathbf{1} = \{0\}$ to prevent confusion with $0 \in \mathbb{Z}$. This element $*$ is used for unsuccesful termination (or 'fail').

In a position $(x, y) \in \mathbb{Z}^2$ we understand x as the current bet (size) and y as the current capital that can still be used in a next bet. Thus:

$$b(x,y) \coloneqq \begin{cases} 1\big|*\rangle & \text{if } x > y \ \text{ (insufficent capital)} \\ \frac{1}{2}\big|2x, y-x\rangle + \frac{1}{2}\big|x+y\rangle & \text{otherwise.} \end{cases}$$

(As before we omit the burocracy of coprojections $\kappa_3(*)$, $\kappa_2(x+y)$ and $\kappa_1(2x, y-x)$ since the types should make sufficiently clear which coproduct component is meant.)

This bet coalgebra b is a coalgebra of a monad, so we can iterate it as $b^n = b \odot \cdots \odot b$, see (10.4) for details. We shall perform a few steps, starting from bet 1 and credit 10.

$$\begin{aligned} b(1,10) &= \tfrac{1}{2}\big|2,9\rangle + \tfrac{1}{2}\big|11\rangle \\ b^2(1,10) &= \tfrac{1}{2}\cdot b(2,9) + \tfrac{1}{2}\big|11\rangle = \tfrac{1}{4}\big|4,7\rangle + \tfrac{1}{2}\big|11\rangle + \tfrac{1}{4}\big|11\rangle = \tfrac{1}{4}\big|4,7\rangle + \tfrac{3}{4}\big|11\rangle \\ b^2(1,10) &= \tfrac{1}{4}\cdot b(4,7) + \tfrac{3}{4}\big|11\rangle = \tfrac{1}{8}\big|8,3\rangle + \tfrac{3}{4}\big|11\rangle + \tfrac{1}{8}\big|11\rangle = \tfrac{1}{8}\big|8,3\rangle + \tfrac{7}{8}\big|11\rangle \\ b^3(1,10) &= \tfrac{1}{8}\cdot b(8,3) + \tfrac{7}{8}\big|11\rangle = \tfrac{1}{8}\big|*\rangle + \tfrac{7}{8}\big|11\rangle. \end{aligned}$$

We see that three times there is an outcome (gain) of 11, but in the end there is a fail $*$ when the capital runs out. This means that the gambler can not be completely certain of a profit of 1.

For each number $N \in \mathbb{Z}$ we define a sharp predicate $p_N \colon \mathbb{Z}^2 \to \{0,1\}$ by $p_N(x,y) = 1$ iff $x + y = N$. We claim that for each N, the predicate p_N is a *weak* invariant for the above bet coalgebra b. We shall use the formulation of Example 10.5.5 (1). First of all,

$$(\textcircled{w}\, p_N)(x,y) = b(x,y) \models [p_N, \mathbf{1}] = \begin{cases} 1 & \text{if } x > y \\ \frac{1}{2}\cdot p_N(2x, y-x) + \frac{1}{2} & \text{otherwise.} \end{cases}$$

We claim $p_N \leq \textcircled{w}\, p_N$. For the proof, let's assume $p_N(x,y) = 1$, so that $x + y = N$. When $x > y$ also $(\textcircled{w}\, p_N)(x,y) = 1$, so this case is trivial. Let's assume $x \leq y$. We notice that $p_N(2x, y-x) = 1$ as well, since $2x + (y-x) = x + y = N$. Thus, also in this case $(\textcircled{w}\, p_N)(x,y) = \frac{1}{2}\cdot p_N(2x, y-x) + \frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1$.

Further aspects about this coalgebra are elaborated in Exercise 10.5.4. There it is shown that rich people — who can start with a higher credit value y — can play longer and thus have a higher chance of ending up with outcome $y + 1$, and thus with a profit of 1.

We consider one more example that further illustrates the use of nexttime operators.

Example 10.5.8 Recall the dice-via-coins coalgebra $f \colon X \to \mathcal{D}(X+A)$ from Subsection 10.1.2, with positions $X = \{x_0, x_1, \ldots, x_6\}$ and observations $A = \{1,2,3,4,5,6\}$. Like in Exercise 10.1.3 we turn each $a \in A$ into a predicate $[a] \colon X \to [0,1]$ via $[a](x) \coloneqq f^*(\mathbf{1}_a)(x) = f(x) \models [\mathbf{0}, \mathbf{1}_a] = f(x)(a)$. It describes the probability of seeing output a after the next step, where f is the coalgebra from Subsection 10.1.2. This gives:

$$[1] = \tfrac{1}{2}\cdot \mathbf{1}_{x_3} \qquad [2] = [3] = \tfrac{1}{2}\cdot \mathbf{1}_{x_4} \qquad [4] = [5] = \tfrac{1}{2}\cdot \mathbf{1}_{x_5} \qquad [6] = \tfrac{1}{2}\cdot \mathbf{1}_{x_6}.$$

With the strong nexttime operator $\textcircled{s}\, p = f^*([p, \mathbf{0}])$ we can start expressing more. For instance:

$$\textcircled{s}[1] = \textcircled{s}[2] = \textcircled{s}[3] = \tfrac{1}{4}\cdot \mathbf{1}_{x_1} \qquad \textcircled{s}[4] = \textcircled{s}[5] = \textcircled{s}[6] = \tfrac{1}{4}\cdot \mathbf{1}_{x_2}.$$

These outcomes make sense when one looks at the transition graph in Figure 10.1. Adding more strong nextttime operators gives:

$$\begin{aligned}
\text{Ⓢ}^2[1] &= \text{Ⓢ}^2[2] = \text{Ⓢ}^2[3] = \tfrac{1}{8}\cdot\mathbf{1}_{x_0} + \tfrac{1}{8}\cdot\mathbf{1}_{x_3}\\
\text{Ⓢ}^2[4] &= \text{Ⓢ}^2[5] = \text{Ⓢ}^2[6] = \tfrac{1}{8}\cdot\mathbf{1}_{x_0} + \tfrac{1}{8}\cdot\mathbf{1}_{x_6}\\
\text{Ⓢ}^3[1] &= \text{Ⓢ}^3[2] = \text{Ⓢ}^3[3] = \tfrac{1}{16}\cdot\mathbf{1}_{x_1}\\
\text{Ⓢ}^3[4] &= \text{Ⓢ}^3[5] = \text{Ⓢ}^3[6] = \tfrac{1}{16}\cdot\mathbf{1}_{x_2}
\end{aligned}$$

It is not hard to see that this patterns repeats: for each $n \geq 1$,

$$\begin{aligned}
\text{Ⓢ}^{2n}[1] &= \text{Ⓢ}^{2n}[2] = \text{Ⓢ}^{2n}[3] = \tfrac{1}{2}\cdot\tfrac{1}{4^n}\cdot\mathbf{1}_{x_0} + \tfrac{1}{2}\cdot\tfrac{1}{4^n}\cdot\mathbf{1}_{x_3}\\
\text{Ⓢ}^{2n+2}[4] &= \text{Ⓢ}^{2n+2}[5] = \text{Ⓢ}^{2n+2}[6] = \tfrac{1}{2}\cdot\tfrac{1}{4^n}\cdot\mathbf{1}_{x_0} + \tfrac{1}{2}\cdot\tfrac{1}{4^n}\cdot\mathbf{1}_{x_6}\\
\text{Ⓢ}^{2n+1}[1] &= \text{Ⓢ}^{2n+1}[2] = \text{Ⓢ}^{2n+1}[3] = \tfrac{1}{4^{n+1}}\cdot\mathbf{1}_{x_1}\\
\text{Ⓢ}^{2n+1}[4] &= \text{Ⓢ}^{2n+1}[5] = \text{Ⓢ}^{2n+1}[6] = \tfrac{1}{4^{n+1}}\cdot\mathbf{1}_{x_2}
\end{aligned}$$

Now we can compute, like in Subsection 10.1.2, but now in logical form, that the 'eventual' value of the predicate $[i]$ for $1 \leq i \leq 6$ is:

$$\left(\bigvee_{n\geq 0} \text{Ⓢ}^n[i]\right)(x_0) = \left(\bigvee_{n\geq 0} \text{Ⓢ}^{2(n+1)}[i]\right)(x_0) = \sum_{n\geq 0} \tfrac{1}{2}\cdot\tfrac{1}{4^{n+1}} = \tfrac{1}{8}\cdot\sum_{n\geq 0}\left(\tfrac{1}{4}\right)^n \overset{(10.1)}{=} \tfrac{1}{8}\cdot\frac{1}{1-{}^1\!/_4} = \tfrac{1}{6}.$$

Exercises

10.5.1 Let $p\colon X \to [0,1]$ and $q\colon Y \to [0,1]$ be two predicates, giving rise to the cotuple predicate $[p,q]\colon X+Y \to [0,1]$. Show that:

$$[p,q]^\perp = [p^\perp, q^\perp].$$

Prove more generally that the cotuple map:

$$Pred(X) \times Pred(Y) \xrightarrow{\;[-,-]\;} Pred(X+Y)$$

is an isomorphism of effect modules — where on the left we use the componentwise effect module structure on a product.

This isomorphism exists in the much more general setting of 'effectuses', see Jacobs (2015).

10.5.2 Consider predicate lifting $FP(F)$ as introduced in Definition 10.5.3 for a *simple* polynomial functor F, not containing $\mathcal{D}$. Show that $FP(F)$ maps sharp predicates to sharp predicates.

10.5.3 Prove for predicates p, q,

$$(p^\perp \otimes q^\perp \otimes \mathbf{1})^\perp = (p \otimes q \otimes \mathbf{1}) \ovee (p \otimes q^\perp \otimes \mathbf{1}) \ovee (p^\perp \otimes q \otimes \mathbf{1}).$$

An instance of this equation, with $q = p$, is used at the end of Example 10.5.5. See also Exercise 4.4.9.

10.5.4 Consider the bet coalgebra $b\colon \mathbb{Z}^2 \to \mathcal{D}(\mathbb{Z}^2 + \mathbb{Z} + 1)$ from Example 10.5.7.

1 First prove by induction on $n \in \mathbb{N}$ that:

$$\sum_{i=0}^{i=n} 2^i \;=\; 2^{n+1} - 1.$$

2 Next show that for $x, y \in \mathbb{Z}$ with $n \in \mathbb{N}$ such that $x \cdot 2^{n+1} \leq y$ one has for each $1 \leq i \leq n$,

$$b^i(x, y) \;=\; (1 - \tfrac{1}{2^i})|x + y\rangle + \tfrac{1}{2^i}|x \cdot 2^i, y - x \cdot (2^i - 1)\rangle.$$

3 Now let $N \in \mathbb{N}$ be the least number with $x \cdot 2^{N+1} > y$, that is, with $N > \log(y/x) - 1$. Check that:

$$c^N(x, y) \;=\; (1 - \tfrac{1}{2^N})|x + y\rangle + \tfrac{1}{2^N}|*\rangle.$$

Thus, rich people have a higher chance of getting a profit out of the gamble of Example 10.5.7.

10.5.5 Consider an arbitrary subset/event $E \subseteq X$ of a coalgebra $c\colon X \to \mathcal{D}(X)$. Show that the following points are equivalent.

1 The sharp preciate $\mathbf{1}_E$ is an invariant for c;
2 $supp(c(x)) \subseteq E$ for each $x \in E$;
3 The coalgebra c restricts to a coalgebra on E, as in:

$$\begin{tikzcd} \mathcal{D}(E) \arrow[r, hook] & \mathcal{D}(X) \\ E \arrow[u, dotted, "c'"] \arrow[r, hook] & X \arrow[u, "c"'] \end{tikzcd}$$

Hint: Recall Exercise 4.4.5.

10.5.6 Consider Exercise 10.1.4 with $r = \frac{3}{4}$ and $s = \frac{2}{3}$. As before we write $[a] = c(-)(\kappa_2 a)$ for the coalgebra $c\colon X \to \mathcal{D}(X+A)$ in this situation. Prove that for $i = 1, 2, 3$ and for $j = 4, 5, 6$ one has:

$$\left(\bigvee_{n \geq 0} \text{Ⓢ}^n[i]\right)(u) \;=\; \tfrac{1}{6} \qquad\qquad \left(\bigvee_{n \geq 0} \text{Ⓢ}^n[j]\right)(v) \;=\; \tfrac{2}{9}.$$

10.5.7 Consider the lifted functor $FP(\mathcal{D})\colon \mathbf{FPred} \to \mathbf{FPred}$ from (10.16).

1 Prove that $FP(\mathcal{D})$ is a monad, essentially by showing that the unit and flatten from $\mathcal{D}$ are suitable morphisms in **FPred**.
2 Check that a channel $c\colon (X, p) \rightsquigarrow (Y, q)$ for the monad $FP(\mathcal{D})$ is a channel $c\colon X \rightsquigarrow Y$ for the monad $\mathcal{D}$ satisfying $p \leq c \mathrel{=\!\!\ll} q$.

10.5.8 Check that Exercise 10.1.3 involves predicates p, q satisfying:

$$q \;=\; p^\perp \supset \text{Ⓢ}q,$$

where $\supset$ is Reichenbach implication, introduced in Exercise 4.4.10.

10.5.9 Let F be a probabilistic polynomial functor.

1 Prove that predicate lifting preserves truth: $FP(X, \mathbf{1}) = (F(X), \mathbf{1})$. Conclude that truth $\mathbf{1}$ is a weak invariant for any coalgebra, and that falsum $\mathbf{0}$ is a strong invariant.
2 Show that predicate lifting is monotone: if predicates $p, q\colon X \to [0, 1]$ on X satisfy $p \leq q$, then $FP(F)(p) \leq FP(F)(q)$, as predicates $F(X) \to [0, 1]$.

3 Prove that predicate lifting commutes with composition, in the following sense: for a function $f\colon X \to Y$ and a predicate $q\colon Y \to [0,1]$ one has:

$$FP(F)(q \circ f) \;=\; FP(F)(q) \circ F(f).$$

Conclude that if f is a homomorphism of coalgebras, and q is a weak invariant, then so is $q \circ f$.

10.6 Behavioural distances

Bisimulation relations, as subsets of a product, form a fundamental notion in process theory to describe equivalence, or behavioural indistinguishability, see Rutten (2000); Staton (2011); Jacobs (2015). There are relational (sharp) formulations of what a bisimulation is for the probabilistic case, see Larsen and Skou (1991); Desharnais et al. (2002); Danos et al. (2006); Sangiorgi (2011); Panangaden (2015). However, in a quantative probabilistic setting it is more common to use 'behavioural' distances instead of bisimulations, see *e.g.* Desharnais et al. (2004); Breugel et al. (2005, 2006); Bacci et al. (2015); Alfaro et al. (2009); Baldan et al. (2018); Bonchi et al. (2018). The idea is that the distance function tells how far apart positions are, where the bisimilar (indistinguishable) positions have distance zero. Here we illustrate this metric approach. In the previous section we have introduced a category **FPred** of fuzzy predicates, defined the predicate lifting of a probabilistic polynomial functor, and used it to define invariants as coalgebras of this lifted functor. In the present section we similarly define a category **FRel** of fuzzy relations, with relation lifting. Behavioural metrics then appear as coalgebras (actually fixed points) of such lifted functors.

The distances that are used in this section are 1-bounded pseudometrics. This means that the distance function, on a set X, is of the form $d\colon X \times X \to [0,1]$, taking values in the unit interval $[0,1]$. It is thus a binary fuzzy predicate, *i.e.* a fuzzy relation. The discrete distance and the total variation distance (between distributions, see Subsection 4.8.1) are 1-bounded. Further, such a distance function is called a pseudometric if unequal elements may have distance zero, that is, if the implication $d(x,x') = 0 \Rightarrow x = x'$ need not hold. These pseudometrics work best in this setting, since indistinguishable elements (with distance zero) need not be equal. A more technical reason is that proper metrics need not be preserved under lifting.

In this setting we use the Kantorovich distance between distributions. We only sketch the main points and refer to Baldan et al. (2018); Jacobs (2025c) for more details.

Definition 10.6.1 The category **FRel** has pairs (X,r) as objects, where $r\colon X \times X \to [0,1]$ is a binary predicate on the product $X \times X$, called a (fuzzy) relation. In **FRel** a morphism

$$\big(X \times X \xrightarrow{r} [0,1]\big) \xrightarrow{\;f\;} \big(Y \times Y \xrightarrow{s} [0,1]\big)$$

consists of a function $f\colon X \to Y$ with $r \leq s \circ (f \times f)$. This means $r(x_1,x_2) \leq s\big(f(x_1), f(x_2)\big)$ for all $x_1, x_2 \in X$.

This category **FRel** comes with an obvious forgetful functor:

$$\begin{array}{ccc} \mathbf{FRel} & (X,r) & f \\ \downarrow & \downarrow & \downarrow \\ \mathbf{Sets} & X & f \end{array}$$

Since this functor is so obvious, we do not even give them a name.

We proceed like in the previous section: we first identify relevant categorical / logical structure for relations, use it for lifting a probabilistic polynomial functor, and then look at coalgebras of the lifted functor. In the predicate-case these coalgebras are invariants; here in the relation-case they are bisimulations.

Lemma 10.6.2 *1 There are truth and false relation functors* $\top, \bot \colon \mathbf{Sets} \to \mathbf{FRel}$*, given by*

$$\top(X) = \Big(X \times X \xrightarrow{\mathbf{1}} [0,1]\Big) \qquad \bot(X) = \Big(X \times X \xrightarrow{\mathbf{0}} [0,1]\Big).$$

2 There are also coproducts of relations via:

$$\Big(X \times X \xrightarrow{r} [0,1]\Big) + \Big(Y \times Y \xrightarrow{s} [0,1]\Big) := \Big((X+Y) \times (X+Y) \xrightarrow{t} [0,1]\Big),$$

where:

$$\begin{aligned} t(\kappa_1 x_1, \kappa_1 x_2) &= r(x_1, x_2) & t(\kappa_2 y_1, \kappa_1 x_2) &= 0 \\ t(\kappa_1 x_1, \kappa_2 y_2) &= 0 & t(\kappa_2 y_1, \kappa_2 y_2) &= s(y_1, y_2). \end{aligned}$$

3 There is a parallel conjunction (tensor) relations:

$$\Big(X \times X \xrightarrow{r} [0,1]\Big) \otimes \Big(Y \times Y \xrightarrow{s} [0,1]\Big) := \Big((X \times Y) \times (X \times Y) \xrightarrow{t} [0,1]\Big),$$

where:

$$t((x_1, y_1), (x_2, y_2)) = r(x_1, x_2) \cdot s(y_1, y_2).$$

It has the truth relation from point (1) *as tensor unit.*

Proof. 1 Easy, see the proof of Lemma 10.5.2 (1).

2 Let $t \colon (X+Y) \times (X+Y) \to [0,1]$ be the coproduct relation defined above in point (2). The coprojection $\kappa_1 \colon X \to X+Y$ is a map $r \to t$ in **FRel**, and similarly for κ_2.

Next we show how to form cotuples in **FRel**. Let $f \colon (X,r) \to (Z,u)$ and $g \colon (Y,s) \to (Z,u)$ be two morphisms in **FRel** with the same codomain. The cotuple $[f,g] \colon X+Y \to Z$ is a morphism $(X,r)+(Y,s) \to (Z,u)$ in **FRel**, where the coproduct $(X,r)+(Y,s)$ is $(X+Y,t)$. Indeed:

$$\begin{aligned} t(\kappa_1 x_1, \kappa_1 x_2) = r(x_1, x_2) &\leq u(f(x_1), f(x_2)) \\ &= (u \circ ([f,g] \times [f,g]))(\kappa_1 x_1, \kappa_1 x_2) \\ t(\kappa_1 x_1, \kappa_2 y_2) = 0 &\leq u(f(x_1), g(y_2)) \\ &= (u \circ ([f,g] \times [f,g]))(\kappa_1 x_1, \kappa_2 y_2) \\ t(\kappa_2 y_1, \kappa_1 x_2) = 0 &\leq u(g(y_1), f(x_2)) \\ &= (u \circ ([f,g] \times [f,g]))(\kappa_2 y_1, \kappa_1 x_2) \\ t(\kappa_2 y_1, \kappa_2 y_2) = s(y_1, y_2) &\leq u(g(y_1), g(y_2)) \\ &= (u \circ ([f,g] \times [f,g]))(\kappa_2 y_1, \kappa_2 y_2). \end{aligned}$$

3 Left to the interested reader. ❑

In a next step we are going to lift the distribution functor $\mathcal{D} \colon \mathbf{Sets} \to \mathbf{Sets}$ to the category of fuzzy relations in such a way that the following diagram commutes.

$$\begin{tikzcd} \mathbf{FRel} \arrow[r, "FR(\mathcal{D})"] \arrow[d] & \mathbf{FRel} \arrow[d] \\ \mathbf{Sets} \arrow[r, "\mathcal{D}"] & \mathbf{Sets} \end{tikzcd} \tag{10.19}$$

For a relation $r\colon X \times X \to [0,1]$ there is a new relation $FR(\mathcal{D})(r)\colon \mathcal{D}(X) \times \mathcal{D}(X) \to [0,1]$ via a lifting that builds on classical work of Kantorovich. It uses couplings, see Section 3.1, and in particular the decoupling function $dcpl := \langle \mathcal{D}(\pi_1), \mathcal{D}(\pi_2)\rangle\colon \mathcal{D}(X \times X) \to \mathcal{D}(X) \times \mathcal{D}(X)$. It is defined on $\omega_1, \omega_2 \in \mathcal{D}(X)$ via validity:

$$FR(\mathcal{D})(r)\big(\omega_1, \omega_2\big) \;:=\; \bigwedge_{\tau \in dcpl^{-1}(\omega_1,\omega_2)} \tau \models r. \tag{10.20}$$

Originally, this definition emerged in a metric setting, where the relaton r is a metric on X, and produces a new metric on $\mathcal{D}(X)$. This goes back to Kantorovich and Rubinshtein (1958). The construction is also associated with Wasserstein and Monge. The history on this topic is a bit unclear clear. We refer to the bibliographic notes at the end of Chapter 6 of Villani (2009) for details.

The minimum (10.19) is actually achieved, via what is called the 'optimal' coupling, and this coupling can be computed. It is done via linear programming, originally introduced by Kantorovich, see Villani (2009); Matouěk and Gärtner (2006). In the sequel we occasionally claim that some coupling is optimal, without proof. It is then obtained via the linear programming approach. For an overview of applications (in computer science) of the Kantorovich distance, see Deng and Du (2009).

Example 10.6.3 Consider the set $X = \{a, b, c\}$ with discrete distance $d = Eq^{\perp}\colon X \times X \to [0,1]$, and with two distributions:

$$\omega_1 \;=\; \tfrac{1}{2}|a\rangle + \tfrac{1}{3}|b\rangle + \tfrac{1}{6}|c\rangle \qquad\qquad \omega_2 \;=\; \tfrac{1}{3}|a\rangle + \tfrac{1}{4}|b\rangle + \tfrac{5}{12}|c\rangle.$$

The Kantorovich distance $d(\omega_1, \omega_2)$ is $\frac{1}{4}$. It is achieved via the optimal coupling:

$$\tau \;=\; \tfrac{1}{3}\big|a,a\big\rangle + \tfrac{1}{6}\big|a,c\big\rangle + \tfrac{1}{4}\big|b,b\big\rangle + \tfrac{1}{12}\big|b,c\big\rangle + \tfrac{1}{6}\big|c,c\big\rangle.$$

This τ is indeed a coupling, with marginals $\tau[1,0] = \omega_1$ and $\tau[0,1] = \omega_2$. The Kantorovich distance arises as validity:

$$\tau \models d \;=\; \tau \models Eq^{\perp} \;=\; \sum_{x,y \in X,\, x \neq y} \tau(x,y) \;=\; \tfrac{1}{6} + \tfrac{1}{12} \;=\; \tfrac{1}{4}.$$

We can now define a more general lifting of probabilistic polynomial functors, by induction on the structure. In the context of bisimulations one use the equality relation on constants. Here we shall use its negation (inequality), that is, discrete distance. The same approach is used for apartness, instead of bisimulation, see (Geuvers and Jacobs, 2021, §5).

Definition 10.6.4 Let $F\colon \mathbf{Sets} \to \mathbf{Sets}$ be a probabilistic polynomial functor. The *relation lifting* of $FR(F)\colon \mathbf{FRel} \to \mathbf{FRel}$ of F is defined by induction on the structure of F.

- The identity functor id on **Sets** is lifted to the identity functor id on **FRel**, that is: $FP(id) := id$.
- For a set A, the constant functor $X \mapsto A$ on **Sets** is lifted to the inequality functor $(X, r) \mapsto (Eq^{\perp}\colon A \times A \to [0,1])$, as orthosupplement of the sharp equality relation Eq introduced in Definition 4.1.1 (4).
- The distribution functor $\mathcal{D}$ is lifted to $FR(\mathcal{D})$, as described in (10.19).
- Coproducts are lifted to coproducts: $FP(F_1 + F_2) := FP(F_1) + FP(F_2)$, see Lemma 10.5.2 (2).
- Products are lifted to parallel conjunctions, via: $FP(F_1 \times F_2) := FP(F_1) \otimes FP(F_2)$, now via Lemma 10.5.2 (3).

By construction, the following diagram commutes.

$$\begin{tikzcd} \mathbf{FRel} \arrow[r, "FR(F)"] \arrow[d] & \mathbf{FRel} \arrow[d] \\ \mathbf{Sets} \arrow[r, "F"] & \mathbf{Sets} \end{tikzcd} \tag{10.21}$$

Now let $c\colon X \to F(X)$ be a coalgebra of the functor F.

1 A *bisimulation* for c is coalgebra $c\colon r \to FR(F)(r)$ in the category **FRel** for the lifted functor $FR(F)$. It involves a fuzzy relation $r\colon X \times X \to [0,1]$ with satisfying:

$$r(x_1, x_2) \leq FR(F)(r)\big(c(x_1), c(x_2)\big).$$

2 A *behavioural metric* for the coalgebra c is a 1-bounded pseudo metric $d\colon X \times X \to [0,1]$ forming a fixed point:

$$d = FR(f)(d) \circ (c \times c). \tag{10.22}$$

In Baldan et al. (2018) it is required that this solution is least, with respect to the pointwise order on relation, but here we leave this open.

In the remainder of this section we elaborate examples of behavioural metrics for the craps game in Subsection 10.1.3 and Example 10.4.4.

Example 10.6.5 We start with the simplified craps game from Example 10.4.4, with set of positions $Y = \{\diamondsuit, \clubsuit, \spadesuit, \heartsuit\}$. We illustrate how to analytically compute the behavioural metric $d\colon Y \times Y \to [0,1]$. The following table already gives an overview.

d	$\diamondsuit$	$\clubsuit$	$\spadesuit$	$\heartsuit$
$\diamondsuit$	0			
$\clubsuit$	$\frac{419}{1980}$	0		
$\spadesuit$	$\frac{101}{660}$	$\frac{1}{10}$	0	
$\heartsuit$	$\frac{53}{396}$	$\frac{2}{11}$	$\frac{1}{11}$	0

(10.23)

For simplicity we have ommitted the symmetric cases. This d is an actual distance function.

We elaborate how to construct this table, starting with the case $d(\clubsuit, \spadesuit)$. We assume that d is an unknown behavioural distance. It must satisfy (10.22), which means in particular that $d(\clubsuit, \spadesuit) = \mathbf{FRel}(F)(d)\big(e(\clubsuit), e(\spadesuit)\big)$, for the functor $F(X) = \mathcal{D}(X + A)$, where $A = \{L, W\}$. We recall from Example 10.4.4 that the coalgebra $e\colon Y \to \mathcal{D}(Y + A)$ satisfies:

$$e(\clubsuit) = \tfrac{3}{4}|\clubsuit\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{12}|W\rangle \qquad\qquad e(\spadesuit) = \tfrac{13}{18}|\spadesuit\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{9}|W\rangle.$$

We have to lift the distance $d\colon Y \times Y \to [0,1]$ to $d'\colon (Y + A) \times (Y + A) \to [0,1]$ according to Lemma 10.6.2 (2), using the discrete distance $Eq^{\perp}$ on A. This gives:

$$d'(z_1, z_2) = \begin{cases} d(z_1, z_2) & \text{if } z_1, z_2 \in Y = \{\diamondsuit, \clubsuit, \spadesuit, \heartsuit\} \\ 0 & \text{if } z_1 = z_2 = L \text{ or } z_1 = z_2 = W \\ 1 & \text{otherwise.} \end{cases}$$

We can now write an equation for the (unknown) distance $x := d(\clubsuit, \spadesuit)$, namely:

$$\begin{aligned}
x &= \mathbf{FRel}(F)(d)\big(e(\clubsuit), e(\spadesuit)\big) \\
&= \mathbf{FRel}(\mathcal{D})(d')\big(e(\clubsuit), e(\spadesuit)\big) \\
&= \mathbf{FRel}(\mathcal{D})(d')\big(\tfrac{3}{4}|\clubsuit\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{12}|W\rangle,\ \tfrac{13}{18}|\spadesuit\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{9}|W\rangle\big) \\
&\overset{(*)}{=} \tfrac{13}{18}\big|\clubsuit, \spadesuit\big\rangle + \tfrac{1}{36}\big|\clubsuit, W\big\rangle + \tfrac{1}{6}\big|L, L\big\rangle + \tfrac{1}{12}\big|W, W\big\rangle \models d' \\
&= \tfrac{13}{18} \cdot d'(\clubsuit, \spadesuit) + \tfrac{1}{36} \cdot d'(\clubsuit, W) + \tfrac{1}{6} \cdot d'(L, L) + \tfrac{1}{12} \cdot d'(W, W) \\
&= \tfrac{13}{18} \cdot x + \tfrac{1}{36} \cdot 1.
\end{aligned}$$

Solving this equation yields the distance in the table (10.23), namely:

$$d(\clubsuit, \spadesuit) = x = \frac{1/36}{1 - 13/18} = \frac{1/36}{5/18} = \tfrac{1}{10}.$$

The marked equation $\overset{(*)}{=}$ holds because the distribution in $\mathcal{D}((Y+A)\times(Y+A))$ before the validity sign $\models$ is the optimal coupling of the distributions $e(\clubsuit)$ and $e(\spadesuit)$.

We do another case from the table (10.23), namely for $d(\diamondsuit, \clubsuit)$, assuming that we also know $d(\heartsuit, \clubsuit) = \frac{2}{11}$. We can now compute directly, without variable, using the optimal coupling:

$$\begin{aligned}
d(\diamondsuit, \clubsuit) &= \mathbf{FRel}(F)(d)\big(e(\diamondsuit), e(\clubsuit)\big) \\
&= \mathbf{FRel}(\mathcal{D})(d')\big(\tfrac{1}{6}|\clubsuit\rangle + \tfrac{2}{9}|\spadesuit\rangle + \tfrac{5}{18}|\heartsuit\rangle + \tfrac{1}{9}|L\rangle + \tfrac{2}{9}|W\rangle, \\
&\qquad\qquad \tfrac{3}{4}|\clubsuit\rangle + \tfrac{1}{6}|L\rangle + \tfrac{1}{12}|W\rangle\big) \\
&= \tfrac{1}{6}\big|\clubsuit, \clubsuit\big\rangle + \tfrac{2}{9}\big|\spadesuit, \clubsuit\big\rangle + \tfrac{5}{18}\big|\heartsuit, \clubsuit\big\rangle + \tfrac{1}{9}\big|L, L\big\rangle \\
&\qquad + \tfrac{1}{12}\big|W, \clubsuit\big\rangle + \tfrac{1}{18}\big|W, L\big\rangle + \tfrac{1}{12}\big|W, W\big\rangle \models d' \\
&= \tfrac{1}{6} \cdot d'(\clubsuit, \clubsuit) + \tfrac{2}{9} \cdot d'(\spadesuit, \clubsuit) + \tfrac{5}{18} \cdot d'(\heartsuit, \clubsuit) + \tfrac{1}{9} \cdot d'(L, L) \\
&\qquad + \tfrac{1}{12} \cdot d'(W, \clubsuit) + \tfrac{1}{18} \cdot d'(W, L) + \tfrac{1}{12} \cdot d'(W, W). \\
&= \tfrac{2}{9} \cdot \tfrac{1}{10} + \tfrac{5}{18} \cdot \tfrac{2}{11} + \tfrac{1}{12} \cdot 1 + \tfrac{1}{18} \cdot 1 = \tfrac{1}{45} + \tfrac{5}{99} + \tfrac{5}{36} = \tfrac{419}{1980}.
\end{aligned}$$

We see that calculating the behavioural metric is quite involved and happens via the calculation of optimal couplings — which can be done via linear programming, see Matouěk and Gärtner (2006). Alternatively, as suggested in Baldan et al. (2018), the behavioural metric can be obtained via approximation, as iteration twoards a fixed point. But then one does not get precise — in this case, fractional — outcomes, as in the above table (10.23).

We conclude with the behavioural pseudometric $X \times X \to [0, 1]$ for the set of positions $X = \{1, 4, 5, 6, 8, 9, 10\}$ for the original craps game from Subsection 10.1.3. We simply list the result, without computational details.

d	1	4	5	6	8	9	10
1	0						
4	$\frac{419}{1980}$	0					
5	$\frac{101}{660}$	$\frac{1}{10}$	0				
6	$\frac{53}{396}$	$\frac{2}{11}$	$\frac{1}{11}$	0			
8	$\frac{53}{396}$	$\frac{2}{11}$	$\frac{1}{11}$	0	0		
9	$\frac{101}{660}$	$\frac{1}{10}$	0	$\frac{1}{11}$	$\frac{1}{11}$	0	
10	$\frac{419}{1980}$	0	$\frac{1}{10}$	$\frac{2}{11}$	$\frac{2}{11}$	$\frac{1}{10}$	0

The most noteworthy outcomes are the equations:

$$d(4,10) = 0 \qquad d(5,9) = 0 \qquad d(6,8) = 0.$$

These pairs of positions with zero-distance are behaviourally indistinguishable.

Recall the homorphism of coalgebras $f\colon X \to Y$ from Example 10.4.4. It is not hard to see that it is an isometry, *i.e.* that it exactly preserves the behavioural distances: $d(f(x), f(x')) = d(x, x')$, for all $x, x' \in X$. The above three pairs of positions in X with distance zero form interesting cases, since they are mapped by f to the same elements in Y.

We observe that this map f is an isometry, but it is not injective. This is possible since the behavioural metric on X is a pseudometric.

The Kantorovich distance on probability distributions, see (10.20), forms a rich topic with many applications (see *e.g.* Deng and Du (2009)). The total variation distance is an instance. The laws of large urns and of large draws are described in Subsection 3.6.1 and Section 5.6 for the total variation distance. They hold in fact more generally, for arbitrary Kantorovich distances, see Jacobs (2025c). This Kantorovich distance can be described for multisets too, see *e.g.* Fritz and Perrone (2019). Using these Kantorovich distances, it is shown in Jacobs (2025c) that the multinomial, hypergeometric, and Pólya channels are isometries, as map $\mathcal{D}(X) \to \mathcal{D}(\mathcal{N}[K](X))$ and $\mathcal{N}[L](X) \to \mathcal{D}(\mathcal{N}[K](X))$. This shows once more that these draw channels are mathematically well-behaved operations.

Exercises

10.6.1 Consider the behaviour metrics table (10.23) in Example 10.6.5.

1 Prove that $d(\clubsuit, \heartsuit) = \frac{2}{11}$ and $d(\spadesuit, \heartsuit) = \frac{1}{11}$ via the (respective) couplings:

$$\begin{aligned}
&\tfrac{25}{36}\big|\clubsuit, \heartsuit\big\rangle + \tfrac{1}{18}\big|\clubsuit, W\big\rangle + \tfrac{1}{6}\big|L, L\big\rangle + \tfrac{1}{12}\big|W, W\big\rangle \\
&\tfrac{25}{36}\big|\spadesuit, \heartsuit\big\rangle + \tfrac{1}{36}\big|\spadesuit, W\big\rangle + \tfrac{1}{6}\big|L, L\big\rangle + \tfrac{1}{9}\big|W, W\big\rangle.
\end{aligned}$$

2 Prove also that $d(\diamondsuit, \spadesuit) = \frac{101}{660}$ and $d(\diamondsuit, \heartsuit) = \frac{53}{396}$ via the couplings:

$$\begin{aligned}
&\tfrac{1}{6}\big|\clubsuit, \spadesuit\big\rangle + \tfrac{2}{9}\big|\spadesuit, \spadesuit\big\rangle + \tfrac{5}{18}\big|\heartsuit, \spadesuit\big\rangle + \tfrac{1}{9}\big|L, L\big\rangle + \tfrac{1}{18}\big|W, \spadesuit\big\rangle + \tfrac{1}{18}\big|W, L\big\rangle + \tfrac{1}{9}\big|W, W\big\rangle \\
&\tfrac{1}{6}\big|\clubsuit, \heartsuit\big\rangle + \tfrac{2}{9}\big|\spadesuit, \heartsuit\big\rangle + \tfrac{5}{18}\big|\heartsuit, \heartsuit\big\rangle + \tfrac{1}{9}\big|L, L\big\rangle + \tfrac{1}{36}\big|W, \heartsuit\big\rangle + \tfrac{1}{18}\big|W, L\big\rangle + \tfrac{5}{36}\big|W, W\big\rangle.
\end{aligned}$$

10.6.2 Check in detail that the map of coalgebras (10.11) is an isometry — as claimed at the end of Example 10.6.5.

10.6.3 Recall the three limit results from Exercise 10.4.2 (1), starting in the positions $\clubsuit$, $\spadesuit$, and $\heartsuit$ of the (reduced) craps game.

1 Let's write D for the Kantorovich distance on the set of distributions $\mathcal{D}(\{L, W\})$, building on the discrete distance on $\{L, W\}$. Show that:

$$\begin{aligned}
D\Big(\tfrac{2}{3}|L\rangle + \tfrac{1}{3}|W\rangle,\ \tfrac{3}{5}|L\rangle + \tfrac{2}{5}|W\rangle\Big) &= \tfrac{1}{15} \\
D\Big(\tfrac{2}{3}|L\rangle + \tfrac{1}{3}|W\rangle,\ \tfrac{6}{11}|L\rangle + \tfrac{5}{11}|W\rangle\Big) &= \tfrac{4}{33} \\
D\Big(\tfrac{3}{5}|L\rangle + \tfrac{2}{5}|W\rangle,\ \tfrac{6}{11}|L\rangle + \tfrac{5}{11}|W\rangle\Big) &= \tfrac{3}{55}.
\end{aligned}$$

Prove these equations via the optimal couplings:

$$\begin{aligned}
&\tfrac{3}{5}\big|L, L\rangle + \tfrac{1}{15}\big|L, W\rangle + \tfrac{1}{3}\big|W, W\rangle \\
&\tfrac{6}{11}\big|L, L\rangle + \tfrac{4}{33}\big|L, W\rangle + \tfrac{1}{3}\big|W, W\rangle \\
&\tfrac{6}{11}\big|L, L\rangle + \tfrac{3}{55}\big|L, W\rangle + \tfrac{2}{5}\big|W, W\rangle.
\end{aligned}$$

2 Notice that these distances differ from the behavioural distances between the positions from which the distributions in the previous point are obtained via iteration — see Exercise 10.4.2 (1). Explain why this is the case.

References

Abramsky, S. 2016. Contextual semantics: From quantum mechanics to logic, databases, constraints, and complexity. Pages 23–50 of: *Contextuality from Quantum Physics to Psychology*. World Scientific Publishing Co.

Abramsky, S., and Jung, A. 1994. Domain Theory. Pages 1–168 of: Abramsky, S., Gabbay, Dov M., and Maibaum, T. (eds), *Handbook of Logic in Computer Science*, vol. 3. Oxford Univ. Press.

Adámek, J. 2005. Introduction to coalgebra. *Theory and Appl. of Categories*, **14(8)**, 157–199.

Adámek, J., Milius, S., and Moss, L. 2025. *Initial Algebras and Terminal Coalgebras. The Theory of Fixed Points of Functors*. Tracts in Theor. Comp. Sci., no. 62. Cambridge Univ. Press.

Aldous, D. 1985. Exchangeability and related topics. Pages 1–198 of: Hennequin, P. (ed), *École d'Été de Probabilités de Saint-Flour XIII — 1983*. Lect. Notes Math., no. 1117. Springer, Berlin.

Alfaro, L. de, Faella, M., and Stoelinga, M. 2009. Linear and branching system metrics. *IEEE Trans. on Software Eng.*, **35(2)**, 258–273.

Alfsen, E., and Shultz, F. 2001. *State spaces of operator algebras: basic theory, orientations and C^*-products*. Mathematics: Theory & Applications. Birkhauser Boston.

Allen, M., and Friston, K. 2018. From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, **195**, 2459–2482.

Anderson, C. 2008. The end of Theory. The Data Deluge makes the Scientific Method Obsolete. *Wired*, **16:07**. Available at https://www.wired.com/2008/06/pb-theory/.

Andrews, G. 1998. *The Theory of Partitions*. Cambridge Univ. Press.

Arbib, M. 1969. *Theories of Abstract Automata*. Englewood Cliffs, NJ: Prentice Hall.

Arimoto, S. 1972. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. on Information Theory*, **18**, 14–20.

Awodey, S. 2006. *Category Theory*. Oxford Logic Guides. Oxford Univ. Press.

Bacci, G., Bacci, G., Larsen, K., and Mardare, R. 2015. On the total variation distance of semi-Markov chains. Pages 185–199 of: Pitts, A. (ed), *Foundations of Software Science and Computation Structures*. Lect. Notes Comp. Sci., no. 9034. Springer, Berlin.

Bacci, G., Furber, R., Kozen, D., Mardare, R., Panangaden, P., and Scott, D. 2018. Boolean-Valued Semantics for the Stochastic λ-Calculus. Pages 669–678 of: Dawar, A., and Grädel, E. (eds), *Logic in Computer Science*. Computer Science Press.

Baier, C., and Katoen, J.-P. 2008. *Principles of Model Checking*. Cambridge, MA: MIT Press.

Bailer-Jones, C. 2017. *Practical Bayesian Inference*. Cambridge Univ. Press.

Baldan, P., Bonchi, F., Kerstan, H., and König, B. 2018. Coalgebraic Behavioral Metrics. *Logical Methods in Comp. Sci.*, **14(3)**.

Balke, A., and Pearl, J. 1994. Probabilistic Evaluation of Counterfactual Queries. Pages 230–237 of: Hayes-Roth, B., and Korf, R. (eds), *Proc. 12th Nat. Conf. on Artificial Intelligence*. AAAI Press / The MIT Press.

Barber, D. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge Univ. Press. Publicly available via http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Brml.HomePage.

Bareinboim, E., Correa, J., Ibeling, D., and Icard, T. 2022. *On Pearl's Hierarchy and the Foundations of Causal Inference*. Assoc. for Computing Machinery. Pages 507–556.

Barr, M., and Wells, Ch. 1985. *Toposes, Triples and Theories*. Springer, Berlin. Revised and corrected version available from URL: www.tac.mta.ca/tac/reprints/articles/12/tr12.pdf.

Barr, M., and Wells, Ch. 1990. *Category Theory for Computing Science*. Englewood Cliffs, NJ: Prentice Hall.

Bartels, F., Sokolova, A., and de Vink, E. 2004. A hierarchy of probabilistic system types. *Theor. Comp. Sci.*, **327(1-2)**, 3–22.

Barthe, G., and Hsu, J. 2021. Probabilistic Couplings from Program Logics. Pages 145–184 of: Barthe, G., Katoen, J.-P., and Silva, A. (eds), *Foundations of Probabilistic Programming*. Cambridge Univ. Press.

Baum, L., Petrie, T., Soules, G., and Weiss, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statistics*, **41**, 164–171.

Beck, J. 1969. Distributive laws. Pages 119–140 of: Eckman, B. (ed), *Seminar on Triples and Categorical Homology Theory*. Lect. Notes Math., no. 80. Springer, Berlin.

Bellman, R. 1957. *Dynamic Programming*. Princeton Univ. Press.

Benjamin, A., and Quinn, J. 2003. *Proofs That Really Count. The Art of Combinatorial Proof*. The Dolciani Mathematical Expositions, no. 27. Math. Assoc. of America.

Bernardo, J., and Smith, A. 2000. *Bayesian Theory*. John Wiley & Sons. Available via https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316870.

Billingsley, P. 1995. *Probability and Measure*. Wiley-Interscience, New York.

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Bishop, C., and Bishop, H. 2024. *Deep Learning. Foundations and Concepts*. Springer Nature.

Blahut, R. 1972. Computation of channel capacity and rate-distortion functions. *IEEE Trans. on Information Theory*, **18**, 460–473.

Blanchi, V., and Paquet, H. 2024. Element-free probability distributions and random partitions. In: *Logic in Computer Science*. Computer Science Press. Article No. 12.

Blei, D., Ng, A., and Jordan, M. 2003. Latent Dirichlet Allocation. *Journ. Machince Learning Research*, **3**, 993–1022.

Blitzstein, J., and Hwang, J. 2019. *Introduction to Probability*. 2nd edn. CRC Press.

Boisseau, G., and Sobociński, P. 2022. String Diagrammatic Electrical Circuit Theory. Pages 178–191 of: Kishida, K. (ed), *Applied Category Theory (ACT) 2021*. Elect. Proc. in Theor. Comp. Sci., no. 372.

Bonchi, F., König, B., and Petrişan, D. 2018. Up-To Techniques for Behavioural Metrics via Fibrations. Pages 17:1–17:17 of: Schewe, S., and Zhang, L. (eds), *Int. Conf. on Concurrency Theory*. LIPIcs, vol. 118. Schloss Dagstuhl.

Breugel, F. van, Hermida, C., Makkai, M., and Worrell, J. 2005. A behavioural pseudometric for probabilistic transition systems. *Theor. Comp. Sci.*, **331(1)**, 115–142.

Breugel, F. van, Hermida, C., Makkai, M., and Worrell, J. 2006. Approximating and computing behavioural distances in probabilistic transition systems. *Theor. Comp. Sci.*, **360(1)**, 373–385.

Breugel, F. van, Hermida, C., Makkai, M., and Worrell, J. 2007. Recursively defined metric spaces without contraction. *Theor. Comp. Sci.*, **380**, 143–163.

Bruza, P., and Abramsky, S. 2016. Probabilistic Programs: Contextuality and Relational Database Theory. Pages 163–174 of: *Quantum Interaction*.

Buscemi, F., and Scarani, V. 2021. Fluctuation theorems from Bayesian retrodiction. *Phys. Rev. E*, **103**, 052111.

Callen, H. 1985. *Thermodynamics and an Introduction to Thermostatistics*. 2nd ed. edn. New York: John Wiley & Sons.

Canonne, C. 2023. *A short note on an inequality between KL and TV*.

Capretta, V., Uustalu, T., and Vene, V. 2009. Corecursive Algebras: A Study of General Structured Corecursion. Pages 84–100 of: Oliveira, M. Vinicius Medeiros, and Woodcock, J. (eds), *Formal Methods: Foundations and Applications*. Lect. Notes Comp. Sci., no. 5902. Springer, Berlin.

Carboni, A., and Walters, R. 1987. Cartesian bicategories I. *Journ. of Pure & Appl. Algebra*, **49**(1-2), 11–32.

Casella, G., and Berger, R. 2002. *Statistical Inference*. 2nd edn. Duxbury Press.

Chan, H., and Darwiche, A. 2005. On the revision of probabilistic beliefs using uncertain evidence. *Artif. Intelligence*, **163**, 67–90.

Chang, C. 1958. Algebraic analysis of many-valued logics. *Trans. Amer. Math. Soc.*, **88**, 476–490.

Chen, S., and Liu, J. 1997. Statistical Applications of the Poisson-Binomial and conditional Bernoulli distributions. *Statistica Sinica*, **7**, 875–892.

Cheng, E. 2022. *The Joy of Abstraction. An Exploration of Math, Category Theory, and Life*. Cambridge Univ. Press.

Chib, S., and Greenberg, E. 1995. Understanding the Metropolis-Hastings Algorithm. *The Amer. Statistician*, **49(4)**, 327–335.

Cho, K., and Jacobs, B. 2017. The EfProb Library for Probabilistic Calculations. In: Bonchi, F., and König, B. (eds), *Conference on Algebra and Coalgebra in Computer Science (CALCO 2017)*. LIPIcs, vol. 72. Schloss Dagstuhl.

Cho, K., and Jacobs, B. 2019. Disintegration and Bayesian inversion via string diagrams. *Math. Struct. in Comp. Sci.*, **29(7)**, 938–971.

Cho, K., Jacobs, B., Westerbaan, A., and Westerbaan, B. 2015. *An Introduction to Effectus Theory*. see https://arxiv.org/abs/1512.05813.

Cignoli, R., D'Ottaviano, I, and Mundici, D. 2000. *Algebraic Foundations of Many-Valued Reasoning*. Trends in Logic, vol. 7. Springer.

Clark, A. 2016. *Surfing Uncertainty. Prediction, Action, and the Embodied Mind*. Oxford Univ. Press.

Clerc, F., Dahlqvist, F., Danos, V., and Garnier, I. 2017. Pointless learning. Pages 355–369 of: Esparza, J., and Murawski, A. (eds), *Foundations of Software Science and Computation Structures*. Lect. Notes Comp. Sci., no. 10203. Springer, Berlin.

Coecke, B., and Kissinger, A. 2016. *Picturing Quantum Processes. A First Course in Quantum Theory and Diagrammatic Reasoning*. Cambridge Univ. Press.

Coecke, B., and Spekkens, R. 2012. Picturing classical and quantum Bayesian inference. *Synthese*, **186(3)**, 651–696.

Conway, J.B. 1990. *A Course in Functional Analysis*. 2nd edn. Graduate Texts in Mathematics 96. Springer.

Conway, J.H., and Guy, R. 1996. *Famous Families of Numbers: Bell Numbers and Stirling Numbers*. Springer New York. Pages 91–126.

Cover, T., and Thomas, J. 2006. *Elements of Information Theory*. 2nd edn. Wiley-Interscience.

Cox, R. 1961. *The Algebra of Probable Inference*. Baltimore, MD: Johns Hopkins Univ. Press.

Crane, H. 2016. The Ubiquitous Ewens Sampling Formula. *Statistical Science*, **31(1)**, 1–19.

Csiszár, I. 1975. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, **3(1)**, 146–158.

Culbertson, J., and Sturtz, K. 2014. A Categorical Foundation for Bayesian Probability. *Appl. Categorical Struct.*, **22(4)**, 647–662.

Dahlqvist, F., and Kozen, D. 2020. Semantics of higher-order probabilistic programs with conditioning. Pages 57:1–57:29 of: *Princ. of Programming Languages*. ACM Press.

Dahlqvist, F., Danos, V., and Garnier, I. 2016. Robustly Parameterised Higher-Order Probabilistic Models. Pages 23:1–23:15 of: Desharnais, J., and Jagadeesan, R. (eds), *Int. Conf. on Concurrency Theory*. LIPIcs, vol. 59. Schloss Dagstuhl.

Dahlqvist, F., Parlant, L., and Silva, A. 2018. Layer by Layer – Combining Monads. Pages 153–172 of: Fischer, B., and Uustalu, T. (eds), *Theoretical Aspects of Computing*. Lect. Notes Comp. Sci., no. 11187. Springer, Berlin.

Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Zhi-Xuan, T., Wing, J., and Tenenbaum, J. 2024. *Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems*. https://arxiv.org/abs/2405.06624.

Danos, V., and Ehrhard, T. 2011. Probabilistic coherence spaces as a model of higher-order probabilistic computation. *Information & Computation*, **209(6)**, 966–991.

Danos, V., Desharnais, J., Laviolette, F., and Panangaden, P. 2006. Bisimulation and cocongruence for probabilistic systems. *Information & Computation*, **204**, 503–523.

Darwiche, A. 2009. *Modeling and Reasoning with Bayesian Networks*. Cambridge Univ. Press.

Dash, S., and Staton, S. 2020. A Monad for Probabilistic Point Processes. In: Spivak, D., and Vicary, J. (eds), *Applied Category Theory Conference*. Elect. Proc. in Theor. Comp. Sci.

Dash, S., and Staton, S. 2021. Monads for measurable queries in probabilistic databases. In: Sokolova, A. (ed), *Math. Found. of Programming Semantics*.

David, F. 1962. *Games, Gods and Gambling. The history of probability from the earliest times to the Newtonian era*. Griffin, London. Available via https://archive.org/details/gamesgodsgambling-david-1962.

Davies, E., and Lewis, J. 1970. An Operational Approach to Quantum Probability. *Communic. Math. Physics*, **17**, 239–260.

Dawid, A. 1987. The Difficulty About Conjunction. *Journ. Royal Statistical Soc.*, **36(2-3)**, 91–97.

de Finetti, B. 1930. Funzione caratteristica di un fenomeno aleatorio. *Memorie della R. Accademia Nazionale dei Lincei*, **IV, fasc. 5**, 86–113. Available at www.brunodefinetti.it/Opere/funzioneCaratteristica.pdf.

de Finetti, B. 2017. *Theory of Probability: A critical introductory treatment*. Wiley.

Dempster, A., Laird, N., and Rubin, D. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journ. Royal Statistical Soc.*, **39(1)**, 1–38.

Deng, Y., and Du, W. 2009. The Kantorovich Metric in Computer Science: A Brief Survey. Pages 73–82 of: Baier, C., and di Pierro, A. (eds), *Quantitative Aspects of Programming Languages*. Elect. Notes in Theor. Comp. Sci., no. 253(3). Elsevier, Amsterdam.

Derkinderen, V., Manhaeve, R., Zuidberg Dos Martires, P., and De Raedt, L. 2024. Semirings for probabilistic and neuro-symbolic logic programming. *Int. Journ. of Approximate Reasoning*, **171**, 109130.

Desharnais, J., Edalat, A., and Panangaden, P. 2002. Bisimulation for labelled Markov processes. *Information & Computation*, **179(2)**, 163–193.

Desharnais, J., Gupta, V., Jagadeesan, R., and Panangaden, P. 2004. Metrics for labelled Markov processes. *Theor. Comp. Sci.*, **318**, 232–354.

Diaconis, P., and Zabell, S. 1982. Updating subjective probability. *Journ. American Statistical Assoc.*, **77**, 822–830.

Diaconis, P., and Zabell, S. 1983. *Some Alternatives to Bayes' Rule*. Tech. rept. 339. Stanford Univ., Dept. of Statistics.

Dice, N. Espinosa, Kaye, M., Ahmed, H., and Montañez, G. 2021. A Probabilistic Theory of Abductive Reasoning. Pages 562–571 of: *Int. Conf. on Agents and Artificial Intelligence*. SciTePress.

Dietrich, F., List, C., and Bradley, R. 2016. Belief revision generalized: A joint characterization of Bayes' and Jeffrey's rules. *Journ. of Economic Theory*, **162**, 352–371.

Dijkstra, E. 1976. *A Discipline of Programming*. Englewood Cliffs, NJ: Prentice Hall.

Dijkstra, E., and Scholten, C. 1990. *Predicate Calculus and Program Semantics*. Berlin: Springer.

Dill, K., and Bromberg, S. 2010. *Molecular Driving Forces. Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*. 2nd rev. edn. New York: Garland Science.

Do, C., and Batzoglou, S. 2008. What is the expectation maximization algorithm? *Nature Biotechnology*, **26**, 897–899.

Donsker, M., and Varadhan, S. 1983. Asymptotic evaluation of certain Markov process expectations for large time. IV. *Communic. Pure and Applied Math.*, **36(2)**, 183–212.

Druzdzel, M., and Simon, H. 1993. Causality in Bayesian Belief Networks. Pages 3–11 of: Heckerman, D., and Mamdani, E. (eds), *Uncertainty in Artif. Intelligence*. Morgan Kaufmann.

Dunn, J. 1967. Characterization of the bivariate negative binomial distribution. *Jour. Arkansas Acad. of Science*, **21**.

Dvurečenskij, A., and Pulmannová, S. 2000. *New Trends in Quantum Structures*. Dordrecht: Kluwer Acad. Publ.

Eddy, S. 2004. What is a hidden Markov model? *Nature Biotechnology*, **22(10)**, 1315–1316.

Ehrhard, T., Tasson, C., and Pagani, M. 2014. Probabilistic coherence spaces are fully abstract for probabilistic PCF. Pages 309–320 of: Jagannathan, S., and Sewell, P. (eds), *Princ. of Programming Languages*. ACM Press.

Eilenberg, S. 1974. *Automata, Languages and Machines*. Academic Press. 2 volumes.

Eppendahl, A. 1999. Coalgebra-to-Algebra Morphisms. Pages 42–49 of: Hofmann, M., Rosolini, G., and Pavlovic, D. (eds), *Conf. on Category Theory and Computer Sci.* Elect. Notes in Theor. Comp. Sci., no. 29. Elsevier, Amsterdam.

Erwig, M., and Walkingshaw, E. 2009. A DSL for Explaining Probabilistic Reasoning. Pages 335–359 of: Taha, W. (ed), *Domain-Specific Languages*. Lect. Notes Comp. Sci., no. 5658. Springer, Berlin.

et al., C. Reijnen. 2020. Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study. *PLoS Medic*, **17(5)**, e1003111.

Ewens, W. 1972. The sampling theory of selectively neutral alleles. *Theoret. Population Biology*, **3**, 87–112.

Feller, W. 1970. *An Introduction to Probability Theory and Its applications*. 3rd rev. edn. Vol. I. Wiley.

Fenton, N., Neil, M., and Berger, D. 2016. Bayes and the Law. *Ann. Review of Statistics and its Applications*, **3**, 51–77.

Feys, F., Hansen, H., and Moss, L. 2018. Long-Term Values in Markov Decision Processes, (Co)Algebraically. Pages 78–99 of: Cîrstea, C. (ed), *Coalgebraic Methods in Computer Science (CMCS 2018)*. Lect. Notes Comp. Sci., no. 11202. Springer, Berlin.

Fierens, D., van den Broeck, G., Renkens, J., Shterionov, D., Gutmann, B., Thon, I., Janssens, G., and Raedt, L. De. 2015. Inference and learning in probabilistic logic programs using weighted Boolean formulas. *Theory & Practice of Logic Programming*, **15(3)**, 358–401.

Fisher, R. 1922. On the mathematical foundations of theoretical statistics. *Philos. Trans. Royal Soc.*, **222**, 594–604.

Fong, B. 2012. *Causal Theories: A Categorical Perspective on Bayesian Networks*. M.Phil. thesis, Univ. of Oxford.

Foulis, D., and Bennett, M. 1994. Effect Algebras and Unsharp Quantum Logics. *Found. Physics*, **24(10)**, 1331–1352.

Friston, K. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, **11(2)**, 127–138.

Fritz, T. 2020. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Advances in Math.*, **370**, 107239.

Fritz, T., and Klingler, A. 2023. The d-Separation Criterion in Categorical Probability. *Journal of Machine Learning Research*, **24(46)**, 1–49.

Fritz, T., and Perrone, P. 2019. A probability monad as the colimit of spaces of finite samples. *Theory and Appl. of Categories*, **34(7)**, 170–220.

Fritz, T., and Perrone, P. 2020. Monads, Partial Evaluations, and Rewriting. Pages 129–148 of: Johann, P. (ed), *Math. Found. of Programming Semantics*. Elect. Notes in Theor. Comp. Sci., no. 352. Elsevier, Amsterdam.

Fritz, T., and Rischel, E. 2020. Infinite products and zero-one laws in categorical probability. *Compositionality*, **2(3)**.

Fritz, T., Klingler, A., McNeely, D., Shah-Mohammed, A., and Wang, Y. 2024. *Hidden Markov Models and the Bayes Filter in Categorical Probability*. https://arxiv.org/abs/2401.14669.

Fujii, S., Katsumata, S., and Melliès, P. 2016. Towards a Formal Theory of Graded Monads. Pages 513–530 of: Jacobs, B., and Löding, C. (eds), *Foundations of Software Science and Computation Structures*. Lect. Notes Comp. Sci., no. 9634. Springer, Berlin.

Garner, R. 2023. Hypernormalisation in an abstract setting. *Information & Computation*, **293**, 105044.

Geiger, D., T., and Pearl, J. 1990. Identifying independence in Bayesian networks. *Networks*, **20**, 507–534.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. 2014. *Bayesian Data Analysis*. 3nd edn. CRC Press.

Geuvers, H., and Jacobs, B. 2021. Relating Apartness and Bisimulation. *Logical Methods in Comp. Sci.*, **17(3)**.

Gibbs, A., and Su, F. 2002. On Choosing and Bounding Probability Metrics. *Int. Statistical Review*, **70(3)**, 419–435.

Gigerenzer, G., and Hoffrage, U. 1995. How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review*, **102(4)**, 684–704.

Giry, M. 1982. A categorical approach to probability theory. Pages 68–85 of: Banaschewski, B. (ed), *Categorical Aspects of Topology and Analysis*. Lect. Notes Math., no. 915. Springer, Berlin.

Glabbeek, R. van, Smolka, S., and Steffen, B. 1995. Reactive, Generative, and Stratified Models of Probabilistic Processes. *Information & Computation*, **121**, 59–80.

Goguen, J., and Burstall, R. 1992. Institutions: Abstract Model Theory for Specification and Programming. *Journ. ACM*, **39(1)**, 95–146.

Goodman, N., and Stuhlmüller, A. 2014. *The Design and Implementation of Probabilistic Programming Languages*. https://dippl.org.

Goodman, N., and Tenenbaum, J. 2016. *Probabilistic Models of Cognition*. http://probmods.org.

Gordon, A., Henzinger, T., Nori, A., and Rajamani, S. 2014. Probabilistic Programming. Pages 167–181 of: *Future of Software Engineering*. ACM.

Goy, A., and Petrişan, D. 2020. Combining probabilistic and non-deterministic choice via weak distributive laws. Pages 454–464 of: Hermanns, H., Zhang, L., Kobayashi, N., and Miller, D. (eds), *Logic in Computer Science*. Computer Science Press.

Graham, R., Knuth, D., and Patashnik, O. 1994. *Concrete Mathematics*. 2nd edn. Addison-Wesley.

Granas, A., and Dugundji, J. 2023. *Fixed Point Theory*. Monographs in Math. Springer.

Griffiths, T., and Tenenbaum, J. 2007. From mere coincidences to meaningful discoveries. *Cognition*, **103**, 180–226.

Griffiths, T., Kemp, C., and Tenenbaum, J. 2008. Bayesian models of cognition. Pages 59–100 of: Sun, R. (ed), *Cambridge Handbook of Computational Cognitive Modeling*. Cambridge Univ. Press.

Griffiths, T., Chater, N., and Tenenbaum, J. 2024. *Bayesian Models of Cognition. Reverse Engineering the Mind*. MIT Press.

Grinstead, C., and Snell, J. 1998. *Introduction to Probability*. Amer. Math. Soc. Available via stats.libretexts.org.

Gu, T., and Zanasi, F. 2019. A Coalgebraic Perspective on Probabilistic Logic Programming. Pages 10:1–10:21 of: Roggenbach, M., and Sokolova, A. (eds), *Conference on Algebra and Coalgebra in Computer Science (CALCO 2019)*. LIPIcs, vol. 139. Schloss Dagstuhl.

Gudder, S. 1998. Morphisms, tensor products and σ-effect algebras. *Reports on Math. Phys.*, **42**, 321–346.

Halpern, J. 2003. *Reasoning about Uncertainty*. Cambridge, MA: MIT Press.

Hayhoe, M., Alajaji, F., and Gharesifard, B. 2017. A Pólya urn-based model for epidemics on networks. Pages 358–363 of: *American Control Conference*.

Heunen, C., Kammar, O., Staton, S., and Yang, H. 2017. A convenient category for higher-order probability theory. Pages 1–12 of: *Logic in Computer Science*. IEEE Computer Society.

Hino, W., Hasuo, H. Kobayashi I., and Jacobs, B. 2016. Healthiness from Duality. In: *Logic in Computer Science*. Computer Science Press.

Hohwy, J. 2013. *The Predictive Mind*. Oxford Univ. Press.

Hyland, M., and Power, J. 2007. The Category Theoretic Understanding of Universal Algebra: Lawvere Theories and Monads. Pages 437–458 of: Cardelli, L., Fiore, M., and Winskel, G. (eds), *Computation, Meaning, and Logic: Articles dedicated to Gordon Plotkin*. Elect. Notes in Theor. Comp. Sci., no. 172. Elsevier, Amsterdam.

Jacobs, B. 1999. *Categorical Logic and Type Theory*. Amsterdam: North Holland.

Jacobs, B. 2010. Convexity, duality, and effects. Pages 1–19 of: Calude, C., and Sassone, V. (eds), *IFIP Theoretical Computer Science 2010*. IFIP Adv. in Inf. and Comm. Techn., no. 82(1). Springer, Boston.

Jacobs, B. 2015. New Directions in Categorical Logic, for Classical, Probabilistic and Quantum Logic. *Logical Methods in Comp. Sci.*, **11(3)**.

Jacobs, B. 2016a. *Introduction to Coalgebra. Towards Mathematics of States and Observations*. Tracts in Theor. Comp. Sci., no. 59. Cambridge Univ. Press.

Jacobs, B. 2016b. *Introduction to Coalgebra. Towards Mathematics of States and Observations*. Cambridge Univ. Press, to appear.

Jacobs, B. 2017a. Hyper Normalisation and Conditioning for Discrete Probability Distributions. *Logical Methods in Comp. Sci.*, **13(3:17)**.

Jacobs, B. 2017b. A Note on Distances between Probabilistic and Quantum distributions. In: Silva, A. (ed), *Math. Found. of Programming Semantics*. Elect. Notes in Theor. Comp. Sci. Elsevier, Amsterdam.

Jacobs, B. 2017c. A recipe for State and Effect Triangles. *Logical Methods in Comp. Sci.*, **13(2)**.

Jacobs, B. 2019. The Mathematics of Changing one's Mind, via Jeffrey's or via Pearl's update rule. *Journ. of Artif. Intelligence Research*, **65**, 783–806.

Jacobs, B. 2021a. From Multisets over Distributions to Distributions over Multisets. In: *Logic in Computer Science*. Computer Science Press.

Jacobs, B. 2021b. Learning from What's Right and Learning from What's Wrong. Pages 116–133 of: Sokolova, A. (ed), *Math. Found. of Programming Semantics*. Elect. Proc. in Theor. Comp. Sci., no. 351.

Jacobs, B. 2021c. Multinomial and Hypergeometric distributions in Markov Categories. Pages 98–115 of: Sokolova, A. (ed), *Math. Found. of Programming Semantics*. Elect. Proc. in Theor. Comp. Sci., no. 351.

Jacobs, B. 2022a. *Basic Combinatorics and Sufficient Statistics for Mutations on Multiple Datatypes*. Researchers.One, https://researchers.one/articles/22.11.00003.

Jacobs, B. 2022b. Partitions and Ewens Distributions in element-free Probability Theory. In: *Logic in Computer Science*. Computer Science Press.

Jacobs, B. 2022c. A Reconstruction of Ewens' Sampling Formula via Lists of Coins. Pages 339–357 of: Jansen, N., Stoelinga, M., and van den Bos, P. (eds), *A Journey from Process Algebra via Timed Automata to Model Learning. Essays Dedicated to Frits Vaandrager on the Occasion of His 60th Birthday*. Lect. Notes Comp. Sci., no. 13560. Springer, Berlin.

Jacobs, B. 2022d. Stick Breaking, in Coalgebra and Probability. Pages 176–193 of: Hansen, H., and Zanasi, F. (eds), *Coalgebraic Methods in Computer Science (CMCS 2022)*. Lect. Notes Comp. Sci., no. 13225. Springer, Berlin.

Jacobs, B. 2022e. Urns & Tubes. *Compositionality*, **4(4)**.

Jacobs, B. 2023a. A principled approach to Expectation Maximisation and Latent Dirichlet Allocation using Jeffrey's update rule. Pages 256–273 of: Hansen, H., Scedrov, A., and de Queiroz, R. (eds), *Worksh. on Logic, Language, Inf. & Comp.* Lect. Notes Comp. Sci., no. 13923.

Jacobs, B. 2023b. Sufficient Statistics and Split Idempotents in Discrete Probability Theory. In: Hsu, J., and Tasson, Ch. (eds), *Math. Found. of Programming Semantics*. Elect. Notes in Theor. Inform. & Comp. Sci., no. 1.

Jacobs, B. 2024. *Getting Wiser from Multiple Data: Probabilistic Updating according to Jeffrey and Pearl*. https://arxiv.org/abs/2405.12700.

Jacobs, B. 2025a. Commutativity of Probabilistic Belief Revision. *Frontiers in Cognition*, **4**.

Jacobs, B. 2025b. *Discrete Boltzmann distributions via multisets and their coefficients*. https://arxiv.org/abs/2507.05719v1.

Jacobs, B. 2025c. Drawing with distance. *Logical Methods in Comp. Sci.*, **21(2)**.

Jacobs, B. 2025d. A fresh look at Bivariate Bionomial Distributions. In: *Math. Found. of Programming Semantics*.

Jacobs, B., and Staton, S. 2020. De Finetti's construction as a categorical limit. Pages 90–111 of: Petrişan, D., and Rot, J. (eds), *Coalgebraic Methods in Computer Science (CMCS 2020)*. Lect. Notes Comp. Sci., no. 12094. Springer, Berlin.

Jacobs, B., and Stein, D. 2023a. Counting and Matching. Pages 28:1–28:15 of: Klin, B., and Pimentel, E. (eds), *Computer Science Logic*. LIPIcs, vol. 252. Schloss Dagstuhl.

Jacobs, B., and Stein, D. 2023b. Overdrawing Urns using Categories of Signed Probabilities. Pages 172–189 of: *Applied Category Th.* Elect. Proc. in Theor. Comp. Sci., no. 397.

Jacobs, B., and Stein, D. 2023c. Pearl's and Jeffrey's Update as Modes of Learning in Probabilistic Programming. *Elect. Notes in Theor. Inform. & Comp. Sci.*, **Volume 3 - Proceedings of MFPS XXXIX**.

Jacobs, B., and Széles, M. 2025. Drawing and Recolouring. Pages 8:1–8:15 of: Cîrstea, C., and Knapp, A. (eds), *Conference on Algebra and Coalgebra in Computer Science (CALCO 2025)*. LIPIcs, vol. 342. Schloss Dagstuhl.

Jacobs, B., and Westerbaan, A. 2020. Distances between States and between Predicates. *Logical Methods in Comp. Sci.*, **16(1)**. See https://lmcs.episciences.org/6154.

Jacobs, B., and Zanasi, F. 2016. A predicate/state transformer semantics for Bayesian learning. Pages 185–200 of: Birkedal, L. (ed), *Math. Found. of Programming Semantics*. Elect. Notes in Theor. Comp. Sci., no. 325. Elsevier, Amsterdam.

Jacobs, B., and Zanasi, F. 2017. A Formal Semantics of Influence in Bayesian Reasoning. Pages 21:1–21:14 of: Larsen, K., Bodlaender, H., and Raskin, J.-F. (eds), *Math. Found. of Computer Science*. LIPIcs, vol. 83. Schloss Dagstuhl.

Jacobs, B., and Zanasi, F. 2021. The Logical Essentials of Bayesian Reasoning. Pages 295–331 of: Barthe, G., Katoen, J.-P., and Silva, A. (eds), *Foundations of Probabilistic Programming*. Cambridge Univ. Press.

Jacobs, B., Silva, A., and Sokolova, A. 2015. Trace semantics via determinization. *Journ. of Computer and System Sci.*, **81(5)**, 859–879.

Jacobs, B., Mandemaker, J., and Furber, R. 2016. The Expectation Monad in Quantum Foundations. *Information & Computation*, **250**, 87–114.

Jacobs, B., Kissinger, A., and Zanasi, F. 2021a. Causal inference via string diagram surgery: A diagrammatic approach to interventions and counterfactuals. *Math. Struct. in Comp. Sci.*, **31(5)**, 553–574.

Jacobs, B., Levy, P., and Rot, J. 2021b. Steps and Traces. *Journ. of Logic and Computation*, **31(6)**, 1482–1525.

Jacobs, B., Széles, M., and Stein, D. 2025. Compositional Inference for Bayesian Networks and Causality. In: *Math. Found. of Programming Semantics*.

Jaynes, E. 1982. Some applications and extensions of the De Finetti representation theorem. Pages 31–42 of: Goel, P., and Zellner, A. (eds), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*. Studies in Bayesian Econometrics and Statistics, vol. 6. North Holland.

Jaynes, E. 2003. *Probability Theory: the Logic of Science*. Cambridge Univ. Press.

Jeffrey, R. 1968. Probable Knowledge. Pages 166–190 of: Lakatos, I. (ed), *The Problem of Inductive Logic*. Studies in Logic and the Foundations of Mathematics, vol. 51. Elsevier, Amsterdam.

Jeffrey, R. 1983. *The Logic of Decision*. 2nd rev., first published in 1965 edn. The Univ. of Chicago Press.

Jensen, F., and Nielsen, T. 2007. *Bayesian Networks and Decision Graphs*. 2nd rev. edn. Statistics for Engineering and Information Science. Springer.

Jewell, N. 2009. *Statistics for Epidemiology*. Texts in Statistical Science Series. Chapman and Hall.

Jimenez, L., Heskes, T., and Stoelinga, M. 2021. Fault Trees, Decision Trees, And Binary Decision Diagrams: A Systematic Comparison. Pages 673–680 of: Castanier, B., Cepin, M., Bigaud, D., and Berenguer, C. (eds), *Proc. 31st Europ. Safety and Reliability Conference*.

Johnson, N., and Kotz, S. 1977. *Urn models and their application: An approach to modern discrete probability theory*. John Wiley.

Johnson, N., Kotz, S., and Balakrishnan, N. 1997. *Discrete Multivariate Distributions*. John Wiley & Sons.

Johnson, N., Kemp, A., and Kotz, S. 2005. *Univariate Discrete Distributions*. 3rd edn. John Wiley & Sons.

Johnstone, P. 1982. *Stone Spaces*. Cambridge Studies in Advanced Mathematics, no. 3. Cambridge Univ. Press.

Jones, C. 1989. *Probabilistic Non-determinism*. Ph.D. thesis, Edinburgh Univ.

Joyce, P. 1998. Partition structures and sufficient statistics. *Journ. of Applied Probability*, **35(3)**, 622–632.

Jung, A., and Tix, R. 1998. The Troublesome Probabilistic Powerdomain. Pages 70–91 of: Edalat, A., Jung, A., Keimel, K., and Kwiatkowska, M. (eds), *Comprox III, Third Workshop on Computation and Approximation*. Elect. Notes in Theor. Comp. Sci., no. 13. Elsevier, Amsterdam.

Jurafsky, D., and Martin, J. 2018. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall. Third Edition, available at https://web.stanford.edu/~jurafsky/slp3/.

Kaelbling, L., Littman, M., and Moore, A. 1996. Reinforcement Learning: A Survey. *Journ. of Artif. Intelligence Research*, **4**, 237–285.

Kallenberg, O. 2021. *Foundations of Modern Probability*. Probability Theory and Stochastic Modelling, no. 99. Springer.

Kamenica, E., and Gentzkow, M. 2011. Bayesian Persuasion. *American Economic Review*, **101(6)**, 2590–2615.

Kantorovich, L., and Rubinshtein, G. 1958. On a space of totally additive functions. *Vestnik Leningrad Univ.*, **13**, 52–59.

Keimel, K., and Plotkin, G. 2017. Mixed powerdomains for probability and nondeterminism. *Logical Methods in Comp. Sci.*, **13(1)**.

Khiatani, D., and Ghose, U. 2017. Weather forecasting using Hidden Markov Model. Pages 220–225 of: *Int. Conf. on Computing and Communication Techn. for Smart Nation*.

Kingman, J. 1978a. Random partitions in population genetics. *Proc. Royal Soc.*, **Series A, 361**, 1–20.

Kingman, J. 1978b. The representation of partition structures. *Journ. London Math. Soc.*, **18(2)**, 374–380.

Knuth, D., and Yao, A. 1976. The complexity of nonuniform random number generation. Pages 357–428 of: Traub, J. (ed), *Algorithms and Complexity: New Directions and Recent Results*. Academic Press, New York.

Kock, A. 1971a. Bilinearity and cartesian closed monads. *Math. Scand.*, **29**, 161–174.

Kock, A. 1971b. Closed categories generated by commutative monads. *Journ. Austr. Math. Soc.*, **XII**, 405–424.

Kock, A. 1995. Monads for which structures are adjoint to units. *Journ. of Pure & Appl. Algebra*, **104**, 41–59.

Kock, A. 2012. Commutative monads as a theory of distributions. *Theory and Appl. of Categories*, **26(4)**, 97–131.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models. Principles and Techniques*. Cambridge, MA: MIT Press.

Komendantskaya, E., and Power, J. 2011. Coalgebraic Semantics for Derivations in Logic Programming. Pages 268–282 of: Klin, B., and Cîrstea, C. (eds), *Conference on Algebra and Coalgebra in Computer Science (CALCO 2011)*. Lect. Notes Comp. Sci., no. 6859. Springer, Berlin.

Kozen, D. 1981. Semantics of probabilistic programs. *Journ. Comp. Syst. Sci*, **22(3)**, 328–350.

Kozen, D. 1985. A probabilistic PDL. *Journ. Comp. Syst. Sci*, **30(2)**, 162–178.

Kozen, D. 2009. *Optimal Coin Flipping*. Manuscript.

Kozen, D., and Silva, A. 2024. Multisets and Distributions. Pages 168–187 of: Capretta, V., Krebbers, R., and Wiedijk, F. (eds), *Logics and Type Systems in Theory and Practice*. Lect. Notes Comp. Sci., no. 14560. Springer, Berlin.

Kullback, S. 1959. *Information Theory and Statistics*. John Wiley & Sons.

Lachman, D. 2024. The Category of ω-Effect Algebras: Tensor Product and ω-Completion. *Order*.

Lambek, J. 1968. A fixed point theorem for complete categories. *Math. Zeitschr.*, **103**, 151–161.

Lang, L., Baudot, P., Quax, R., and Forré, P. 2025. Information Decomposition Diagrams Applied beyond Shannon Entropy: A Generalization of Hu's Theorem. *Compositionality*, **7**.

Larrañaga, P., and Moral, S. 2011. Probabilistic graphical models in artificial intelligence. *Applied Soft Computing*, **11(2)**, 1511–1528.

Larsen, K., and Skou, A. 1991. Bisimulation through probabilistic testing. *Information & Computation*, **94**, 1–28.

Lau, P., Koo, T., and Wu, C. 2020. Spatial Distribution of Tourism Activities: A Pólya Urn Process Model of Rank-Size Distribution. *Journ. of Travel Research*, **59(2)**, 231–246.

Lauritzen, S. 1996. *Graphical models*. Oxford: Oxford Univ. Press.

Lauritzen, S., and Spiegelhalter, D. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journ. Royal Statistical Soc.*, **50(2)**, 157–224.

Lavore, E. Di, and Roman, M. 2023. Evidential Decision Theory via Partial Markov Categories. In: *Logic in Computer Science*. Computer Science Press.

Lawvere, F. 1962. *The category of probabilistic mappings*. Unpublished manuscript, see ncatlab.org/nlab/files/lawvereprobability1962.pdf.

Leinster, T. 2014. *Basic Category Theory*. Cambridge Studies in Advanced Mathematics. Cambridge Univ. Press. Available online via https://arxiv.org/abs/1612.09375.

Leinster, T. 2021. *Entropy and Diversity: The Axiomatic Approach*. Cambridge Univ. Press. Available online via https://arxiv.org/abs/2012.02113.

Libkin, L., and Wong, L. 1993. Some Properties of Query Languages for Bags. Pages 97–114 of: Beeri, C., Ohori, A., and Shasha, D. (eds), *Database Programming Languages*. Workshops in Computing. Springer, Berlin.

Lindvall, T. 1987. *Lectures on the Coupling Method*. New York: John Wiley & Sons.

Lopuhaä-Zwakenberg, M. 2024. Fault Tree Reliability Analysis via Squarefree Polynomials. Pages 39–49 of: Mayo, F., Pires, L., and Seidewitz, E. (eds), *Proc. 12th Int. Conf. on Model-Based Software and Systems Engineering*. Scitepress.

Lorenz, R., and Tull, S. 2023. *Causal models in string diagrams*. https://arxiv.org/abs/2304.07638.

Mackay, D. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge Univ. Press.

Mahmoud, H. 2008. *Pólya Urn Models*. Chapman and Hall.

Manes, E. 1974. *Algebraic Theories*. Berlin: Springer.

Manes, E., and Mulry, P. 2007. Monad compositions I: general constructions and recursive distributive laws. *Theory and Appl. of Categories*, **18**, 172–208.

Manning, C., and Schütze, H. 2003. *Foundations of statistical natural language processing*. MIT Press.

Mardare, R., Panangaden, P., and Plotkin, G. 2016. Quantitative Algebraic Reasoning. In: *Logic in Computer Science*. Computer Science Press.

Matouěk, J., and Gärtner, B. 2006. *Understanding and Using Linear Programming*. Berlin: Springer Verlag.

Mac Lane, S. 1971. *Categories for the Working Mathematician*. Berlin: Springer.

Mac Lane, S. 1986. *Mathematics: Form and Function*. Berlin: Springer.

Mcareavey, K., Bauters, K., Hong, J., and Liu, W. 2017. The event calculus in probabilistic logic programming with annotated disjunctions. Pages 105–113 of: Bidoit, M., and Dauchet, M. (eds), *Proc. Int. J. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*.

McIver, A., and Morgan, C. 2004. *Abstraction, refinement and proof for probabilistic systems*. Monographs in Comp. Sci. Springer.

McIver, A., Morgan, C., Smith, G., Espinoza, B., and Meinicke, L. 2014. Abstract Channels and Their Robust Information-Leakage Ordering. Pages 83–102 of: Abadi, M., and Kremer, S. (eds), *Princ. of Security and Trust*. Lect. Notes Comp. Sci., no. 8414. Springer, Berlin.

McIver, A., Morgan, C., and Rabehaja, T. 2015. Abstract Hidden Markov Models: A Monadic Account of Quantitative Information Flow. Pages 597–608 of: *Logic in Computer Science*. Computer Science Press.

McLachlan, G., and Krishnan, T. 2007. *The EM Algorithm and Extensions*. 2nd edn. Probability and Statistics. John Wiley & Sons.

Milius, S., Pattinson, D., and Schröder, L. 2015. Generic Trace Semantics and Graded Monads. Pages 253–269 of: Moss, L., and Sobocinski, P. (eds), *Conference on Algebra and Coalgebra in Computer Science (CALCO 2015)*. LIPIcs, vol. 35. Schloss Dagstuhl.

Mitchell, T. 1997. *Machine Learning*. McGraw-Hill.

Moggi, E. 1991. Notions of Computation and Monads. *Information & Computation*, **93(1)**, 55–92.

Moore, A. 2002. *Markov Systems, Markov Decision Processes, and Dynamic Programming*. Tutorial slides, see `www.cs.cmu.edu/~./awm/tutorials/mdp09.pdf`.

Mundici, D. 2011. *Advanced Łukasiewicz calculus and MV-algebras*. Trends in Logic, vol. 35. Springer.

Murphy, K. 2012. *Machine Learning. A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Murphy, K. 2023. *Probabilistic Machine Learning: Advanced Topics*. Cambridge, MA: MIT Press.

Nagel, R. 1974. Order unit and base norm spaces. Pages 23–29 of: Hartkämper, A., and Neumann, H. (eds), *Foundations of Quantum Mechanics and Ordered Linear Spaces*. Lect. Notes Physics, no. 29. Springer, Berlin.

Nielsen, M., and Chuang, I. 2000. *Quantum Computation and Quantum Information*. Cambridge Univ. Press.

Olmedo, F., Gretz, F., Kaminski, B. Lucien, Katoen, J-P., and McIver, A. 2018. Conditioning in Probabilistic Programming. *ACM Trans. on Prog. Lang. & Syst.*, **40(1)**, 4:1–4:50.

Ozawa, M. 1984. Quantum measuring processes of continuous observables. *Journ. Math. Physics*, **25**, 79–87.

Pan, R., Peng, Y., and Ding, Z. 2006. Belief Update in Bayesian Networks Using Uncertain Evidence. Pages 441–444 of: *Int. Conf. on Tools with Artificial Intelligence*.

Panangaden, P. 1998. The Category of Markov Kernels. Pages 171–187 of: Baier, C., Huth, M., Kwiatkowska, M., and Ryan, M. (eds), *Workshop on Probabilistic Methods in Verification*. Elect. Notes in Theor. Comp. Sci., no. 21. Elsevier, Amsterdam.

Panangaden, P. 2015. Probabilistic bisimulation. *ACM SIGLOG News*, **2(3)**, 72–84.

Parikh, R. 1966. On Context-Free Languages. *Journ. ACM*, **13(4)**, 570–581.

Parr, T., Pezzulo, G., and Friston, K. 2022. *Active Inference. The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, MA: MIT Press.

Paulsen, V., and Tomforde, M. 2009. Vector Spaces with an order unit. *Indiana Univ. Math. Journ.*, **58-3**, 1319–1359.

Pearl, J. 1982. Reverend Bayes on inference engines: a distributed hierarchical approach. Pages 133–136 of: *AAAI-82: Sec. Nat. Conf. on Artificial Intelligence*. AAAI Press.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Graduate Texts in Mathematics 118. Morgan Kaufmann.

Pearl, J. 1989. Probabilistic Semantics for Nonmonotonic Reasoning: A Survey. Pages 505–516 of: Brachman, R., Levesque, H., and Reiter, R. (eds), *First Intl. Conf. on Principles of Knowledge Representation and Reasoning*. Morgan Kaufmann.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, **82**, 669–710.

Pearl, J. 2009. *Causality. Models, Reasoning, and Inference*. 2nd edn. Cambridge Univ. Press.

Pearl, J., and Mackenzie, D. 2019. *The Book of Why*. Penguin Books.

Pearl, J., and Verma, T. 1991. A Theory of Inferred Causation. Pages 441–452 of: Allen, J., Fikes, R., and Sandewall, E. (eds), *Principles of Knowledge Representation and Reasoning (KR'91)*. Morgan Kaufmann.

Pearl, J., Glymour, M., and Jewell, N. 2016. *Causal Inference in Statistics*. Wiley.

Perrone, P. 2024a. Markov Categories and Entropy. *IEEE Trans. on Information Theory*, **70(3)**, 1671–1692.

Perrone, P. 2024b. *Starting Category Theory*. Singapore: World Scientific.

Piedeleu, R., and Zanasi, F. (eds). 2025. *An Introduction to String Diagrams for Computer Scientists*. Elements in Applied Category Theory. Cambridge Univ. Press.

Piedeleu, R., Torres-Ruiz, M., Silva, A., and Zanasi, F. 2025. A Complete Axiomatisation of Equivalence for Discrete Probabilistic Programming. In: *EU Symp. on Programming*.

Pierce, B. 1991. *Basic Category Theory for Computer Scientists*. Cambridge, MA: MIT Press.

Pishro-Nik, H. 2014. *Introduction to probability, statistics, and random processes*. Kappa Research LLC. Available at https://www.probabilitycourse.com.

Pitman, J. 1995. Exchangeable and partially exchangeable random partitions. *Probability Th. and related fields*, **102(2)**, 145–158.

Pitman, J. 2006. *Combinatorial Stochastic Processes: École d'Été de Probabilités de Saint-Flour XXXII — 2002*. Lect. Notes Math., no. 1875. Springer.

Plotkin, G. 1983. *Domains*. The "Pisa" notes. Available from https://homepages.inf.ed.ac.uk/gdp/publications/Domains_a4.ps.

Poole, D. 2008. The Independent Choice Logic and Beyond. Pages 222–243 of: Raedt, L. De, Frasconi, P., Kersting, K., and Muggleton, S. (eds), *Probabilistic Inductive Logic Programming: Theory and Applications*. Springer, Berlin.

Poole, D., and Mackworth, A. 2017. *Artificial Intelligence. Foundations of Computational Agents*. 2nd edn. Cambridge Univ. Press. Publicly available via https://www.cs.ubc.ca/~poole/aibook/2e/html/ArtInt2e.html.

Pouget, A., Beck, J., Ma, W., and Latham, P. 2013. Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, **16**, 1170–1178.

Prescher, D. 2005. *A Tutorial on the Expectation-Maximization Algorithm Including Maximum-Likelihood Estimation and EM Training of Probabilistic Context-Free Grammars*. See https://arxiv.org/abs/cs/0412015.

Pulmannová, S., and Gudder, S. 1998. Representation theorem for convex effect algebras. *Commentationes Mathematicae Universitatis Carolinae*, **39(4)**, 645–659.

Puterman, M. 1994. *Markov Decision Processes. Discrete Stochastic Dynamic Programming*. New Jersey: John Wiley and Sons.

Quinlan, J. 1986. Induction of Decision Trees. *Machine Learning*, **1(1)**, 81–106.

Rabin, M. 1963. Probabilistic Automata. *Information & Control*, **6(3)**, 230–245.

Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, **77(2)**, 257–286.

Raedt, L., and Kimmig, A. 2015. Probabilistic (logic) programming concepts. *Machine Learning*, **100**, 5–47.

Raedt, L., Kimmig, A., and Toivonen, H. 2007. ProbLog: a probabilistic prolog and its application in link discovery. Pages 2468–2473 of: *Int. Joint Conf. on Artif. Intelligence*. AAAI Press.

Rao, C. 1955. Estimation and tests of significance in factor analysis. *Pychometrika*, **20(2)**, 93–111.

Rao, C. 1973. *Linear Statistical Inference and its Applications*. 2nd edn. Probability and Statistics. John Wiley & Sons.

Rao, R., and Ballard, D. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, **2**, 79–87.

Rice, J. 2006. *Mathematical Statistics and Data Analysis*. 3nd edn. Duxbury Press.

Ross, S. 2007. *Introduction to Probability Models*. 9th edn. Academic Press.

Ross, S. 2018. *A first course in probability*. 10th edn. Pearson Education.

Rota, G.-C., and Baclawski, K. 1979. *An Introduction To Probability And Random Processes*. Available at https://archive.org/details/GianCarlo_Rota_and_Kenneth_Baclawski__An_Introduction_to_Probability_and_Random_Processes.

Rubner, Y., Tomasi, C., and Guibas, L. 2000. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. Journ. of Computer Vision*, **40**, 99–121.

Russell, S., and Norvig, P. 2003. *Artificial Intelligence. A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.

Rutten, J. 2000. Universal Coalgebra: a Theory of Systems. *Theor. Comp. Sci.*, **249**, 3–80.

Sangiorgi, D. 2011. *Introduction to Bisimulation and Coinduction*. Cambridge Univ. Press.

Sato, T., and Kameya, Y. 2008. New Advances in Logic-Based Probabilistic Modeling by PRISM. Pages 118–155 of: Raedt, L. De, Frasconi, P., Kersting, K., and Muggleton, S. (eds), *Probabilistic Inductive Logic Programming: Theory and Applications*. Springer, Berlin.

Ścibior, A., Kammar, O., Vákár, M., Staton, S., Yang, H., Cai, Y., Ostermann, K., Moss, S., Heunen, C., and Ghahramani, Z. 2018. Denotational Validation of Higher-order Bayesian Inference. Pages 60:1–60:29 of: *Princ. of Programming Languages*. ACM Press.

Scozzafava, P. 1993. Uniform distribution and sum modulo *m* of independent random variables. *Statistics & Probability Letters*, **18(4)**, 313–314.

Selinger, P. 2011. A survey of graphical languages for monoidal categories. Pages 289–355 of: Coecke, B. (ed), *New Structures in Physics*. Lect. Notes Physics, no. 813. Springer, Berlin.

Selvin, S. 1975. A problem in probability (letter to the editor). *Amer. Statistician*, **29(1)**, 67.

Sethuraman, J. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Simons, H. 2011. *An Introduction to Category Theory*. Cambridge Univ. Press.

Simpson, E. 1951. The interpretation of interaction in contingency tables. *Journ. Royal Statistical Soc.*, **13**, 238–241.

Singh, D., Ibrahim, A., Yohanna, T., and Singh, J. 2008. A Systematization of the Fundamentals of Multisets. *Lecturas Matemáticas*, **29**, 33–48.

Skorski, M. 2022. *Handy Formulas for Binomial Moments*. See https://www.arxiv.org/abs/2012.06270.

Sloman, S. 2005. *Causal Models. How People Think about the World and Its Alternatives*. Oxford Univ. Press.

Smith, R., Friston, K., and Whyte, C. 2022. A step-by-step tutorial on active inference and its application to empirical data. *Journ. Math. Psychology*, **107**, 102632.

Snell, J., and Vanderbei, R. 1995. Three bewitching paradoxes. Pages 355–370 of: Snell, J. (ed), *Topics in Contemporary Probability and its Applications*. CRC Press.

Sokolova, A. 2011. Probabilistic systems coalgebraically: A survey. *Theor. Comp. Sci.*, **412(38)**, 5095–5110.

Spiegelhalter, D. 2019. *The Art of Statistics. How to learn from data*. New York: Basic Books.

Spiegelhalter, D., Dawid, A., Lauritzen, S., and Cowell, R. 1993. Bayesian Analysis in Expert Systems. *Statistical Science*, **8(3)**, 219–247.

Spirtes, P. 2010. Introduction to causal inference. *Journal of Machine Learning Review*, **11**, 1643–1662.

Spirtes, P., Glymour, C., and Scheines, R. 2000. *Causation, Prediction, and Search*. 2nd rev. edn. Cambridge, MA: MIT Press.

Sprunger, D., Katsumata, S., Dubut, J., and Hasuo, I. 2018. Fibrational Bisimulations and Quantitative Reasoning. Pages 190–213 of: Cîrstea, C. (ed), *Coalgebraic Methods in Computer Science (CMCS 2018)*. Lect. Notes Comp. Sci., no. 11202. Springer, Berlin.

Staton, S. 2011. Relating coalgebraic notions of bisimulation. *Logical Methods in Comp. Sci.*, **7(1:13)**, 1–21.

Staton, S. 2021. Probabilistic Programs as Measures. Pages 43–74 of: Barthe, G., Katoen, J.-P., and Silva, A. (eds), *Foundations of Probabilistic Programming*. Cambridge Univ. Press.

Staton, S., and Summer, N. 2022. Quantum de Finetti Theorems as Categorical Limits, and Limits of State Spaces of C*-algebras. Pages 400–414 of: Gogioso, S., and Hoban, M. (eds), *Quantum Programming Languages (QPL 2022)*. Elect. Proc. in Theor. Comp. Sci., no. 394.

Staton, S., Yang, H., Heunen, C., Kammar, O., and Wood, F. 2016. Semantics for probabilistic programming: higher-order functions, continuous distributions, and soft constraints. In: *Logic in Computer Science*. Computer Science Press.

Stein, A. 2018. Behavioral Probability. Pages 45–72 of: Teitelbaum, J., and Zeiler, K. (eds), *Research Handbook on Behavioral Law and Economics*. Edward Elgar Publishing.

Stone, M. 1949. Postulates for the barycentric calculus. *Ann. Math.*, **29**, 25–30.

Suhov, Y., and Kelbert, M. 2005. *Probability and Statistics by Example: Volume 1, Basic Probability and Statistics*. Cambridge Univ. Press.

Sutton, R., and Barto, A. 2005. *Reinforcement Learning. An Introduction*. 2nd edn. Cambridge, MA: MIT Press.

Tang, W., and Tang, F. 2023. The Poisson Binomial Distribution — Old & New. *Statistical Science*, **38(1)**, 108–119.

Tavaré, S. 2021. The magical Ewens sampling formula. *Bull. London Math. Soc.*, **53**, 1563–1582.

Taylor, D., and Kokshoorn, B. 2023. *Forensic DNA Trace Evidence Interpretation. Activity Level Propositions and Likelihood Ratios*. CRC Press.

Thrun, S., Burgard, W., and Fox, D. 2005. *Probabilistic Robotics*. Cambridge, MA: MIT Press.

Tierney, L. 1994. Markov Chains for Exploring Posterior Distributions. *Ann. Stat.*, **22(4)**, 1701–1728.

Tijms, H. 2007. *Understanding Probability: Chance Rules in Everyday Life*. 2nd edn. Cambridge Univ. Press.

Tull, S., Kleiner, J., and Smithe, T. St Clere. 2023. *Active Inference in String Diagrams: A Categorical Account of Predictive Processing and Free Energy*.

Tull, S., Lorenz, R., Clark, S., Khan, I., and Coecke, B. 2024. *Towards Compositional Interpretability for XAI*.

Tversky, A., and Kahneman, D. 1982. Evidential impact of base rates. Pages 153–160 of: Kahneman, D., Slovic, P., and Tversky, A. (eds), *Judgement under uncertainty: Heuristics and biases*. Cambridge Univ. Press.

Valtorta, M., Kim, Y.-G., and Vomlel, J. 2002. Soft evidential update for probabilistic multiagent systems. *Int. Journ. of Approximate Reasoning*, **29(1)**, 71–106.

Varacca, D. 2003. *Probability, Nondeterminism and Concurrency: Two Denotational Models for Probabilistic Computation*. Ph.D. thesis, Univ. Aarhus. BRICS Dissertation Series, DS-03-14.

Varacca, D., and Winskel, G. 2006. Distributing probability over non-determinism. *Math. Struct. in Comp. Sci.*, **16**, 87–113.

Verma, T., and Pearl, J. 1988. Causal networks: semantics and expressiveness. Pages 69–78 of: Shachter, R., Levitt, T., Kanal, L., and Lemmer, J. (eds), *Uncertainty in Artif. Intelligence*. North-Holland.

Villani, C. 2009. *Optimal Transport — Old and New*. Berlin Heidelberg: Springer.

Walton, C. 2023. *Spies. The Epic Intelligence War Between East*. Simon & Schuster.

Westerbaan, A., Westerbaan, B., and van de Wetering, J. 2020. A characterisation of ordered abstract probabilities. In: *Logic in Computer Science*. Computer Science Press.

Wingate, D., Stuhlmüller, A., and Goodman, N. 2011. Lightweight Implementations of Probabilistic Programming Languages Via Transformational Compilation. Pages 770–778 of: Gordon, G., Dunson, D., and Dudík, M. (eds), *Proc. of Machine Learning Research*.

Witten, I., Frank, E., and Hall, M. 2011. *Data Mining – Practical Machine Learning Tools and Techniques*. Amsterdam: Elsevier.

Yoon, B. 2009. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, **10(6)**, 402–415.

Zhang, Y., and Amin, N. 2022. Reasoning about 'reasoning about reasoning': Semantics and contextual equivalence for probabilistic programs with nested queries and recursion. In: *Princ. of Programming Languages*. ACM Press.

Zwart, M. 2020. *On the Non-Compositionality of Monads via Distributive Laws*. Ph.D. thesis, Dep. of Comp. Science, Univ. of Oxford.

Zwart, M., and Marsden, D. 2019. No-Go Theorems for Distributive Laws. Pages 1–13 of: *Logic in Computer Science*. Computer Science Press.