# 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: July 4, 2023

## Contents

Preface page v
1 Collections ..... 1
1.1 Notation ..... 3
1.2 Coefficients ..... 6
1.3 Cartesian products ..... 10
1.4 Lists ..... 14
1.5 Subsets ..... 26
1.6 Multisets ..... 37
1.7 Multisets in summations ..... 49
1.8 Coefficients of multisets ..... 57
1.9 Multiset partitions and the triangular prism ..... 68
1.10 Channels ..... 75
1.11 The role of category theory ..... 80
2 Discrete probability distributions ..... 88
2.1 Probability distributions ..... 89
2.2 Frequentist learning and stick breaking ..... 100
2.3 Parallel products ..... 112
2.4 Probabilistic channels ..... 124
2.5 String diagrams and Bayesian networks ..... 135
2.6 Draw distributions ..... 147
2.7 Convolution ..... 159
2.8 Divergence between distributions ..... 168
2.9 Exchangeability for positions and elements ..... 174
3 Drawing from an urn ..... 186
3.1 Zipping multisets ..... 188
3.2 Single draws ..... 199
3.3 The multinomial channel ..... 214
3.4 The hypergeometric channel ..... 229
3.5 The Pólya channel ..... 236
3.6 The parallel multinomial law: four definitions ..... 244
3.7 The parallel multinomial law: basic properties ..... 252
3.8 Parallel multinomials as law of monads ..... 267
3.9 Discrete Poisson point processes ..... 275
$4 \quad$ Observables and validity ..... 285
$4.1 \quad$ Validity ..... 287
4.2 The structure of observables ..... 302
4.3 Transformation of observables ..... 319
4.4 Validity and drawing ..... 328
4.5 Validity-based distances ..... 336
5 Variance and covariance ..... 351
5.1 Variance and shared-state covariance ..... 352
5.2 Draw distributions and their (co)variances ..... 363
5.3 Joint-state covariance and correlation ..... 373
5.4 Independence for random variables ..... 380
5.5 The law of large numbers, in weak form ..... 385
6 Updating distributions ..... 392
6.1 Update basics ..... 393
6.2 Examples of forward and backward inference ..... 404
6.3 Analysis of forward and backward inference ..... 422
6.4 Inference in Bayesian networks ..... 428
6.5 Hidden Markov models ..... 435
6.6 Updating draw distributions ..... 447
6.7 Discretisation, and coin bias learning ..... 453
6.8 Frequentist and Bayesian probability ..... 462
$7 \quad$ Daggers and disintegrations ..... 471
7.1 Bayesian inversion: the dagger of a channel ..... 472
7.2 Disintegration of joint distributions ..... 485
7.3 Disintegration, in general form ..... 493
$7.4 \quad$ Disintegration for learning with missing data ..... 502
7.5 Disintegration and inversion in machine learning ..... 507
7.6 Sufficient statistics ..... 519
7.7 Pearl's and Jeffrey's update rules ..... 526
7.8 Factorisation of joint states ..... 538
$7.9 \quad$ Categorical aspects of Bayesian inversion ..... 545
References ..... 553

## 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 [5] 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 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 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.

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 $\frac{3}{5}$.

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 ones. The ket notation $|-\rangle$ is just syntactic sugar that it used to separate the multiplicities $(2,3)$ from the elements $(R, B)$. 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 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 speci-
fication of programming languages, in computer security and privacy, in statebased 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, crucial ingredients (like distributions in expected values) 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 Theorem6.1.5. 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 computerassisted reasoning requires a clear syntax and semantics. For instance, it is hard to even express the above-mentioned Corollary 7.1.7 and Theorem 6.1.5 in standard probabilistic notation. 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 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.
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 from it. 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, 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 7.2.7.

Here is a characteristic illustration of our structure-based approach. A well known property of the Poisson distribution pois is commonly expressed as: if $X_{1} \sim \operatorname{pois}\left[\lambda_{1}\right]$ and $X_{2} \sim \operatorname{pois}\left[\lambda_{2}\right]$ then $X_{1}+X_{2} \sim \operatorname{pois}\left[\lambda_{1}+\lambda_{2}\right]$. 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 $\left[\lambda_{1}+\lambda_{2}\right]=$ pois $\left[\lambda_{1}\right]+\operatorname{pois}\left[\lambda_{2}\right]$. It says that the Poisson channel pois is a is a homomorphism of monoids, from non-negative reals $\mathbb{R}_{\geq 0}$ to distributions $\mathcal{D}(\mathbb{N})$ on the natural numbers, see Proposition 2.7.6for details. This result uses a 'convolution' commutative monoid structure on distributions whose underlying space is itself a commutative monoid. This monoid structure plays a role in many other situations, for instance in the fundamental distributive law that turns multisets of distributions into distributions 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 network), 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 transform both states and predicates. 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, inputs for successive learning steps, etc. The interplay between multisets and distributions, notably in learning, is a recurring theme.
3 States (distributions) are treated as separate from, and dual, to predicates. These predicates are the ingredients of an (implicit) probabilistic logic, with, for instance conjunction and negation operations. States are really different entities, 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 $[0,1]$. Along a channel one can transfer states forward, and predicates backward. A central notion is the validity of a predicate in a state, written as $\vDash$. It is standardly called expected value. Conditioning involves updating a state with a predicate.
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 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 - in presence of a 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 [146] 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, 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 nonmonotonid ${ }^{1}$ — a property that many logicians shy away from. This book does

[^0]aim to contribute towards bridging the divide mentioned by Pearl, by providing a mathematical basis for such a symbolic probabilistic logic, consisting of channels, states, predicates, transformations, 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. They recognised that taking probability distributions has the structure of a monad, which was published in the early 1980s in [63]. Roughly at the same time Dexter Kozen started the systematic investigation of probabilistic programming languages and logics, published in [116, 117]. 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 [25, 73]).
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. [33, 34, 36, 49 140, 170, 68, 8, 168].
- The compositional approach to Bayesian networks [29, 54] and to Bayesian reasoning [32, 96, 98].
- The use of categorical and diagrammatic methods in quantum foundations, including quantum probability, see [28] for an overview.
- The efforts to develop 'synthetic' probability theory via a categorical axiomatisation, see e.g. [57, 58, 161, 23, 82].

This book builds on these developments.
The intended audience consists of students and professionals - in mathematics, computer science, artificial intelligence 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 [174], but to provide clear and precise formalisations of the relevant structures. Mathematical abstraction (esp. categorical air guitar playing) is not a goal in itself (except maybe towards the end of Chapter 33: instead, the book tries to uncover relevant abstractions in concrete problems. It includes several basic algorithms, with a focus on the

[^1]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. 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 [7, 123], 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 lists, subsets, and multisets. 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).
Blindly drawing coloured balls from an urn is a basic model in discrete probability. Such draws are 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. They are typical for our 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 $X \rightarrow \mathbb{R}$ (including factors, predicates and events) that can be defined on sample spaces $X$, providing numerical information. The chapter concentrates on validity of obervables in states and on transformation of observables. 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. These 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 combination of a state and an observable, on the same underlying sample space. The statistical notions of variance and covariance are described in terms of of validity for such random variables in Chapter 5 This chapter distinguishes two forms of covariance, with a 'shared' or a 'joint' state, which satisfy different properties.

A very special technique in the area of probability theory is conditioning, also known as belief updating, 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. In Chapter 6 we formulate conditioning as an explicit operation, mapping a state $\omega$ and a predicate $p$ to a new updated state $\left.\omega\right|_{p}$. A key result is that the validity of $p$ in $\left.\omega\right|_{p}$ is higher than the validity of $p$ in the original state $\omega$. This means that we have learned from $p$ and adapted our state (of mind) from $\omega$ to $\left.\omega\right|_{p}$. This updating operation $\left.\omega\right|_{p}$ forms an action (of predicates on states) and satisfies Bayes' rule, in fuzzy form. The combination with forward and backward transformation along a channel leads to the 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, such as naive Bayesian classification and decision trees. Moreover, they form the basis of a new 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 [70, 26] that the human mind is seen as a Bayesian prediction engine that operates by reducing prediction errors.
One of the themes running through this book is how 'crossover' influence can be captured via channels - extracted from joint states 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 [109, Ch. 8].

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

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.

## 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 [23] 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 7, several additional chapters exist in unfinished form, for instance on statistical learning, probabilistic automata, causality and on continuous probability. They will be incorporated in due course.

Bart Jacobs, Nijmegen, May ??, 2023.
${ }^{2}$ 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 Recent versions of EfProb also allow calculations in fractional form.

## 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. 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, subsets and multisets are summarised in the table below.

|  | lists | subsets | multisets |
| :---: | :---: | :---: | :---: |
| 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.
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.

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 (1.48), 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 input $\rightarrow T$ (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 $X=\left\{x_{1}, \ldots, x_{n}\right\}$ we write $|X|=n$ for its number of elements. We call this number $|X|$ the size of the set $X$. Sometimes we also write $\operatorname{size}(X)=|X|$. 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 $\boldsymbol{n}=\{0,1, \ldots, n-1\}$ 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 $X \rightarrow \mathbf{2}$ can be identified with a predicate on the set $X$, or with a subset of $X$. Clearly, there is precisely one function $X \rightarrow \mathbf{1}$, namely the function that sends every element $x \in X$ to the sole element 0 in $\mathbf{1}=\{0\}$. In the other direction, functions $\mathbf{1} \rightarrow X$ 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:

$$
[n):=\{0,1, \ldots, n-1\}=n \quad \text { and } \quad(n]:=\{1,2, \ldots, n\} .
$$

We use the symbol := for definitions.
A function $f$ from a set $X$ to a set $Y$ is written as $f: X \rightarrow Y$ or as $X \xrightarrow{f} Y$. In that case we call $X$ the domain and $Y$ the codomain of the function. 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 inc: $\mathbb{N} \rightarrow \mathbb{N}$ with $\operatorname{inc}(x)=x+1$. For each set $X$ there is an identity function $i d_{X}: X \rightarrow X$, defined as $i d_{X}(x)=x$. When confusion is unlikely, we omit $X$ from $i d_{X}$ and simply write id: $X \rightarrow X$.

Given two functions $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ we use $g \circ f: X \rightarrow Z$ for the composite function ' $g$ after $f$ '. It is defined as $(g \circ f)(x)=g(f(x))$. Then $\operatorname{id}_{Y} \circ f=f=f \circ \mathrm{id}_{X}$. 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 $\rightsquigarrow$ and with their own composition $\odot$.

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


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 $\rightsquigarrow$ 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 $f: X \rightarrow Y$ and $g: Y \rightarrow X$ with $g \circ f=i d_{X}$ and $f \circ g=i d_{Y}$. 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 $f: X \xrightarrow{\cong} Y$ to indicate that $f$ is an isomorphism, with $f^{-1}: Y \xrightarrow{\cong} X$ 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 $X \xlongequal{\leftrightharpoons} X$.

We may write a surjective function as $X \rightarrow Y$ an an injective function as $X \mapsto Y$. A function that is surjective (injective) is also called a surjection (injection). A function of the form $f: X \rightarrow X$ 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 $r \circ s=i d$.
For an arbitrary function $f: X \rightarrow Y$ 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 $\varphi: X \rightarrow \mathbb{R}$ 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 $\varphi: X \rightarrow \mathbb{R}$ 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 $\mathbb{N} \rightarrow \mathbb{R}$. It may happen that the maximum exists, but still the argmax subset is empty, for instance for the function $\mathbb{N} \rightarrow \mathbb{R}$ 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 $\binom{n}{m}$, $\left(\binom{n}{m}\right),\left[\begin{array}{l}n \\ m\end{array}\right]$ and $\left\{\begin{array}{l}n \\ m\end{array}\right\}$. 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 relevant mathematical properties of these coefficients, for future reference.

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

$$
n!:=1 \cdot 2 \cdot \ldots \cdot(n-1) \cdot n \quad \text { which means } \quad\left\{\begin{aligned}
0! & =1 \\
(n+1)! & =(n+1) \cdot n!.
\end{aligned}\right.
$$

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 two constructions.
Definition 1.2.1. Let numbers $n \geq m \geq 0$ be given.
1 The falling factorial ( $n)_{m}$ is defined as:

$$
(n)_{m}:=\frac{n!}{(n-m)!}=n \cdot(n-1) \cdot \ldots \cdot(n-m+1)
$$

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:

$$
\binom{n}{m}:=\frac{n!}{m!\cdot(n-m)!}=\frac{(n)_{m}}{m!} .
$$

3 For numbers $m_{1}, \ldots, m_{K} \in \mathbb{N}$ with $n=\sum_{i} m_{i}$ there is the multinomial coefficent given by:

$$
\binom{n}{m_{1} \ldots m_{K}}:=\frac{n!}{m_{1}!\cdot \ldots \cdot m_{K}!} .
$$

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 $0<m \leq n$,

$$
\begin{equation*}
\binom{n+1}{m}=\binom{n}{m}+\binom{n}{m-1} \tag{1.1}
\end{equation*}
$$

The following basic result will be useful later.
Lemma 1.2.2. Fix a number $m \in \mathbb{N}$. Then.

$$
\lim _{n \rightarrow \infty} \frac{\binom{n}{m}}{n^{m}}=\frac{1}{m!}
$$

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

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

Definition 1.2.3. Let $n \geq 1$ and $m \geq 0$.
1 The rising factorial $(n)^{m}$ is defined as:

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

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 $n^{\bar{m}}$ 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 $\left(\binom{n}{m}\right)=\frac{(n)^{m}}{m!}$.
It is well-known that $\binom{n}{m}$ 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

### 1.2.1 Stirling numbers

So-called Stirling numbers are used for counting various objects in combinatorics. They exists 'of the first kind' and 'of the second kind', written respectively as $\left[\begin{array}{c}n \\ m\end{array}\right]$ and $\left\{\begin{array}{l}n \\ m\end{array}\right\}$. 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 $m>0$.

$$
\left[\begin{array}{l}
0  \tag{1.2}\\
0
\end{array}\right]=1 \quad\left[\begin{array}{l}
0 \\
m
\end{array}\right]=\left[\begin{array}{l}
m \\
0
\end{array}\right]=0 \quad\left[\begin{array}{c}
n+1 \\
m
\end{array}\right]=n \cdot\left[\begin{array}{l}
n \\
m
\end{array}\right]+\left[\begin{array}{c}
n \\
m-1
\end{array}\right]
$$

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

Lemma 1.2.4. Let $n \geq 0$.
$1\left[\begin{array}{c}n \\ m\end{array}\right]=0$ when $m>n$;
$2\left[\begin{array}{l}n \\ n\end{array}\right]=1$;
$3\left[\begin{array}{c}n+1 \\ 1\end{array}\right]=n!$;
$4\left[\begin{array}{c}n \\ n-1\end{array}\right]=\binom{n}{2}$, when $n \geq 2$;
5 For $r \in \mathbb{R}$,

$$
\sum_{0 \leq m \leq n}\left[\begin{array}{l}
n \\
m
\end{array}\right] \cdot r^{m}=(r)^{n}=r \cdot(r+1) \cdot \ldots \cdot(r+n-1)
$$

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

We turn to Stirling numbers of the second kind, written as $\left\{\begin{array}{l}n \\ m\end{array}\right\}$, 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 $m>0$.

$$
\left\{\begin{array}{l}
0  \tag{1.3}\\
0
\end{array}\right\}=1 \quad\left\{\begin{array}{c}
m \\
0
\end{array}\right\}=\left\{\begin{array}{c}
0 \\
m
\end{array}\right\}=0 \quad\left\{\begin{array}{c}
n+1 \\
m
\end{array}\right\}=m \cdot\left\{\begin{array}{l}
n \\
m
\end{array}\right\}+\left\{\begin{array}{c}
n \\
m-1
\end{array}\right\}
$$

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

## Lemma 1.2.5.

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

$$
\sum_{0 \leq m \leq n}\left\{\begin{array}{l}
n \\
m
\end{array}\right\} \cdot(r)_{m}=r^{n}=r(r-1) \cdots(r-n+1)
$$

## 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 $n \in$ $\mathbb{N}$.
$1 \sum_{0 \leq i \leq n} i=\frac{n \cdot(n+1)}{2}$.
$2 \sum_{0 \leq i \leq n} i^{2}=\frac{n \cdot(n+1) \cdot(2 n+1)}{6}$.
These results can be generalised to $K$-ary sums, for $K \geq 1$.
$3 \sum_{0 \leq i_{1}, \ldots, i_{K} \leq n} i_{1}+\cdots+i_{K}=\frac{K \cdot n \cdot(n+1)^{K}}{2}$.
$4 \sum_{0 \leq i_{1}, \ldots, i_{K} \leq n}\left(i_{1}+\cdots+i_{K}\right)^{2}=\frac{K \cdot n \cdot(n+1)^{K} \cdot((3 K+1) \cdot n+2)}{12}$.
Proof. The first two items are standard, and easy to prove by induction on $n$. The last two items are consequences, obtained again by induction, now on $K$. Notice that sums start at 0 ; this is relevant for the multiple sums, in (3) and (4).

## Exercises

1.2.1 Prove Pascal's rule (1.1).
1.2.2 1 Use this rule (1.1) to prove:

$$
\sum_{0 \leq i \leq n}\binom{m+i}{m}=\binom{n+m+1}{m+1}
$$

2 Now use the previous equation and (1.1) to prove:

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

1.2.3 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:

$$
\binom{n}{n_{1}, \ldots, n_{K}}=\binom{n}{n_{1}} \cdot\binom{n-n_{1}}{n_{2}, \ldots, n_{K}}
$$

1.2.4 Prove the statements about Stirling numbers of the first kind that are listed in Lemma 1.2.4
1.2.5 Prove that for $n \geq m \geq 0$,

$$
\left[\begin{array}{l}
n \\
m
\end{array}\right]=\sum_{U \subseteq\{1, \ldots, n\},|U|=n-m} \prod_{k \in U} k-1 .
$$

1.2.6 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.
1.2.7 Prove the following equations for Stirling numbers of the second kind, for $n \geq 2$.
$1 \quad\left\{\begin{array}{c}n \\ n-1\end{array}\right\}=\binom{n}{2}$.
$2\left\{\begin{array}{l}n \\ 2\end{array}\right\}=2^{n-1}-1$.
1.2.8 Prove Proposition 1.2 .6

### 1.3 Cartesian products

This section briefly reviews some (standard) terminology and notation related to Cartesian products of sets.

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:

$$
X_{1} \times X_{2}:=\left\{\left(x_{1}, x_{2}\right) \mid x_{1} \in X_{1} \text { and } x_{2} \in X_{2}\right\} .
$$

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:

$$
X_{1} \times \cdots \times X_{n}:=\left\{\left(x_{1}, \ldots, x_{n}\right) \mid x_{1} \in X_{1}, \ldots, x_{n} \in X_{n}\right\} .
$$

The tuple $\left(x_{1}, \ldots, x_{n}\right)$ 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 $\Pi$, as:

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

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 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 $\mid$ - | 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 $\pi_{i}$ out of the product to the set $X_{i}$, as in:

$$
X_{1} \times \cdots \times X_{n} \xrightarrow{\pi_{i}} X_{i} \quad \text { given by } \quad \pi_{i}\left(x_{1}, \ldots, x_{n}\right):=x_{i} .
$$

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 $f_{1}: Y \rightarrow X_{1}, \ldots, f_{n}: Y \rightarrow X_{n}$, then we can form a new function $Y \rightarrow X_{1} \times \cdots \times X_{n}$, namely:

$$
Y \xrightarrow{\left\langle f_{1}, \ldots, f_{n}\right\rangle} X_{1} \times \cdots \times X_{n} \quad \text { via } \quad\left\langle f_{1}, \ldots, f_{n}\right\rangle(y):=\left(f_{1}(y), \ldots, f_{n}(y)\right) .
$$

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

$$
\begin{equation*}
\pi_{i} \circ\left\langle f_{1}, \ldots, f_{n}\right\rangle=f_{i} . \tag{1.4}
\end{equation*}
$$

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:

$$
\begin{equation*}
\left\langle f_{1}, \ldots, f_{n}\right\rangle \circ g=\left\langle f_{1} \circ g, \ldots, f_{n} \circ g\right\rangle \quad\left\langle\pi_{1}, \ldots, \pi_{n}\right\rangle=i d, \tag{1.5}
\end{equation*}
$$

where $g: Z \rightarrow Y$ 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 $f_{i}: X_{i} \rightarrow Y_{i}$. 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:

$$
f_{1} \times \cdots \times f_{n}=\left\langle f_{1} \circ \pi_{1}, \ldots, f_{n} \circ \pi_{n}\right\rangle
$$

so that:

$$
\left(f_{1} \times \cdots \times f_{n}\right)\left(x_{1}, \ldots, x_{n}\right)=\left(f_{1}\left(x_{1}\right), \ldots, f_{n}\left(x_{n}\right)\right) .
$$

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:

$$
X^{n}:=\underbrace{X \times \cdots \times X}_{n \text { times }}=\left\{\left(x_{1}, \ldots, x_{n}\right) \mid x_{i} \in X \text { for each } i\right\} .
$$

As special cases we have $X^{1}=X$ and $X^{0}=\mathbf{1}$. Since powers are special cases of Cartesian products, they come with projection functions $\pi_{i}: X^{n} \rightarrow X$ and tuple functions $\left\langle f_{1}, \ldots, f_{n}\right\rangle: Y \rightarrow X^{n}$ for $n$ functions $f_{i}: Y \rightarrow X$. Finally, for a function $f: X \rightarrow Y$ we write $f^{n}: X^{n} \rightarrow Y^{n}$ for the obvious $n$-fold parallelisation of $f$.

There is a copy function $\Delta[K]: X \rightarrow X^{K}=X \times \cdots \times X(K$ 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 $\Delta$ functions are alternatively called copiers or diagonals.
More generally, for two sets $X, Y$ we shall occasionally write:

$$
X^{Y}:=\{\text { functions } f: Y \rightarrow X\} .
$$

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 $\boldsymbol{n} \rightarrow X$ 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:

$$
\begin{equation*}
\frac{Z \times Y \xrightarrow{\stackrel{f}{\longrightarrow}} X}{Z \underset{g}{Y}} \tag{1.6}
\end{equation*}
$$

This means that for a function $f: Z \times Y \rightarrow X$ there is a corresponding function $\bar{f}: Z \rightarrow X^{Y}$, and vice-versa, for $g: Z \rightarrow X^{Y}$ there is a function $\bar{g}: Z \times Y \rightarrow X$, in such a way that $\overline{\bar{f}}=f$ and $\overline{\bar{g}}=g$. It is not hard to see that we can take $\bar{f}(z) \in X^{Y}$ to be the function $\bar{f}(z)(y)=f(z, y)$, for $z \in Z$ and $y \in Y$. Similarly, we use $\bar{g}(z, y)=g(z)(y)$.

The correspondence (1.6) is characteristic for so-called Cartesian closed categories.

## Exercises

1.3.1 Check what a tuple function $\left\langle\pi_{2}, \pi_{3}, \pi_{6}\right\rangle$ 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 $\left\langle f_{1}, \ldots, f_{n}\right\rangle$ is the unique function $h: Y \rightarrow X_{1} \times \cdots \times X_{n}$ with $\pi_{i} \circ h=f_{i}$ for each $i$.
1.3.3 Prove, using Equations (1.4) and (1.5) for tuples and projections, that:

$$
\left(g_{1} \times \cdots \times g_{n}\right) \circ\left(f_{1} \times \cdots \times f_{n}\right)=\left(g_{1} \circ f_{1}\right) \times \cdots \times\left(g_{n} \circ f_{n}\right) .
$$

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


Check that $\pi_{i} \circ \Delta=i d$, for each $1 \leq i \leq K$.
1.3.5 Define functions in both directions, using tuples and projections, that yield isomorphisms:

$$
X \times Y \cong Y \times X \quad 1 \times X \cong X \quad X \times(Y \times Z) \cong(X \times Y) \times Z
$$

Try to use Equations (1.4) and (1.5) to prove these isomorphisms, without reasoning with elements.
1.3.6 Similarly, show that exponents satisfy:

$$
X^{1} \cong X \quad 1^{Y} \cong 1 \quad(X \times Y)^{Z} \cong X^{Z} \times Y^{Z} \quad X^{Y \times Z} \cong\left(X^{Y}\right)^{Z}
$$

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

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

by:

$$
\operatorname{zip}[K]\left(\left(x_{1}, \ldots, x_{K}\right),\left(y_{1}, \ldots, y_{K}\right)\right):=\left(\left(x_{1}, y_{1}\right), \ldots,\left(x_{K}, y_{K}\right)\right) .
$$

1 Show that zip is an isomorphism, with inverse function unzip $[K]:=$ $\left\langle\left(\pi_{1}\right)^{K},\left(\pi_{2}\right)^{K}\right\rangle$.

2 Show that the following diagram commutes.


### 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.
For an arbitrary set $X$ we write $\mathcal{L}(X)$ for the set of all finite lists $\left[x_{1}, \ldots, x_{n}\right]$ 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 (-).

Thus, the set of lists over $X$ can be defined as a (disjoint) union of all powers of $X$, as in:

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

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: $\mathcal{L}(\mathbf{1}) \cong \mathbb{N}$.
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 $f: X \rightarrow Y$ 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 $\mathcal{L}(f): \mathcal{L}(X) \rightarrow \mathcal{L}(Y)$ for the map-list function, defined as:

$$
\mathcal{L}(f)\left(\left[x_{1}, \ldots, x_{n}\right]\right):=\left[f\left(x_{1}\right), \ldots, f\left(x_{n}\right)\right] .
$$

In this way identity maps and compositions are preserved:

$$
\mathcal{L}(i d)=\text { id } \quad \mathcal{L}(g \circ f)=\mathcal{L}(g) \circ \mathcal{L}(f) .
$$

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 $\mathcal{L}\left(\pi_{i}\right)$, where $\pi_{i}$ is a projection map. For instance, let $\ell \in \mathcal{L}(X \times Y)$ be a list of tuples of the form $\ell=\left[\left(x_{1}, y_{1}\right), \ldots,\left(x_{n}, y_{n}\right)\right]$. The first marginal $\mathcal{L}\left(\pi_{1}\right)(\ell) \in$ $\mathcal{L}(X)$ is then computed as:

$$
\begin{aligned}
\mathcal{L}\left(\pi_{1}\right)(\ell) & =\mathcal{L}\left(\pi_{1}\right)\left(\left[\left(x_{1}, y_{1}\right), \ldots,\left(x_{n}, y_{n}\right)\right]\right) \\
& =\left[\pi_{1}\left(x_{1}, y_{1}\right), \ldots, \pi_{1}\left(x_{n}, y_{n}\right)\right] \\
& =\left[x_{1}, \ldots, x_{n}\right] .
\end{aligned}
$$

### 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 $M \times$ $M \rightarrow M$, 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 $a, b, c \in M$,

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

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 $f: M \rightarrow N$ is called a homomorphism of monoids if $f$ preserves the unit and binary operation, in the sense that:

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

Diagrammatically we can express the second equation as:


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 $f:(\mathbb{N}, 0,+) \rightarrow(\mathbb{N}, 1, \cdot)$.

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 $+: \mathcal{L}(X) \times \mathcal{L}(X) \rightarrow$ $\mathcal{L}(X)$ as binary operation:

$$
\left[x_{1}, \ldots, x_{n}\right]+\left[y_{1}, \ldots, y_{m}\right]:=\left[x_{1}, \ldots, x_{n}, y_{1}, \ldots, y_{m}\right] .
$$

This monoid $(\mathcal{L}(X),[],+)$ is neither commutative nor idempotent.
2 For each function $f: X \rightarrow Y$ the associated map $\mathcal{L}(f): \mathcal{L}(X) \rightarrow \mathcal{L}(Y)$ is a homomorphism of monoids.

Thus, lists are monoids via concatenation. But there is more to say: lists are free monoids. We shall occasionally make use of this basic property and so we like to make it explicit. We shall encounter similar freeness properties for other collection types.

Each element $x \in X$ yields a singleton list, written as unit $(x):=[x] \in \mathcal{L}(X)$. The resulting function unit : $X \rightarrow \mathcal{L}(X)$ 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 $f: X \rightarrow M$. Then there is a unique homomorphism of monoids $\bar{f}:(\mathcal{L}(X),[],+-) \rightarrow(M, 0,+)$ with $\bar{f} \circ$ unit $=f$.

The homomorphism $\bar{f}$ 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.


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

$$
\bar{f}([])=0 \quad \text { and } \quad \bar{f}([x])=f(x)
$$

Further, on an list $\left[x_{1}, \ldots, x_{n}\right]$ of length $n \geq 2$ we necessarily have:

$$
\begin{aligned}
\bar{f}\left(\left[x_{1}, \ldots, x_{n}\right]\right) & =\bar{f}\left(\left[x_{1}\right]+\cdots+\left[x_{n}\right]\right) \\
& =\bar{f}\left(\left[x_{1}\right]\right)+\cdots+\bar{f}\left(\left[x_{n}\right]\right) \\
& =f\left(x_{1}\right)+\cdots+f\left(x_{n}\right) .
\end{aligned}
$$

Thus, there is only one way to define $\bar{f}$. By construction, this $\bar{f}: \mathcal{L}(X) \rightarrow M$ 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 $\alpha: M \times X \rightarrow X$ satisfying:

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

for all $a, b \in M$ and $x \in X$.
(Sometimes monoid actions occur of the form $X \times M \rightarrow X$, 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 $f: X \rightarrow Y$ satisfing:

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

This last equation corresponds to commutation of the following diagram.


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 $s \cdot\left(s^{\prime} \cdot v\right)=\left(s \cdot s^{\prime}\right) \cdot v$. Also, as we shall see, probabilistic updating can be described via monoid actions. The action map $\alpha: M \times X \rightarrow X$ 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.

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 unit: $X \rightarrow \mathcal{L}(X)$, given by $\operatorname{unit}(x):=[x]$. There is also a 'flatten' function which turns a list of lists into a list by removing inner brackets. This function is written as flat : $\mathcal{L}(\mathcal{L}(X)) \rightarrow$ $\mathcal{L}(X)$. It is defined as:

$$
\operatorname{flat}\left(\left[\left[x_{11}, \ldots, x_{1 n_{1}}\right], \ldots,\left[x_{k 1}, \ldots, x_{k n_{k}}\right]\right]\right):=\left[x_{11}, \ldots, x_{1 n_{1}}, \ldots, x_{k 1}, \ldots, x_{k n_{k}}\right]
$$

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 $f: X \rightarrow Y$ one has:

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

Equivalently, the following two diagrams commute.



2 One further has:

$$
\text { flat } \circ \text { unit }=\text { id }=\text { flat } \circ \mathcal{L}(\text { unit }) \quad \text { flat } \circ \text { flat }=\text { flat } \circ \mathcal{L}(\text { flat }) .
$$

These two equations can equivalently be expressed via commutation of:


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

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

Next, for the second item we take an arbitrary list $\left[x_{1}, \ldots, x_{n}\right] \in \mathcal{L}(X)$. Then:

$$
\begin{aligned}
(\text { flat } \circ \text { unit })\left(\left[x_{1}, \ldots, x_{n}\right]\right) & =\operatorname{flat}\left(\left[\left[x_{1}, \ldots, x_{n}\right]\right]\right) \\
& =\left[x_{1}, \ldots, x_{n}\right] \\
(\text { flat } \circ \mathcal{L}(\text { unit }))\left(\left[x_{1}, \ldots, x_{n}\right]\right) & =\operatorname{flat}\left(\left[\operatorname{unit}\left(x_{1}\right), \ldots, \operatorname{unit}\left(x_{n}\right)\right]\right) \\
& =\operatorname{flat}\left(\left[\left[x_{1}\right], \ldots,\left[x_{n}\right]\right]\right) \\
& =\left[x_{1}, \ldots, x_{n}\right] .
\end{aligned}
$$

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 1.11 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 $\alpha: \mathcal{L}(X) \rightarrow X$ satisfying $\alpha \circ$ unit $=$ id and $\alpha \circ$ flat $=\alpha \circ$ $\mathcal{L}(\alpha)$, as in:


2 Let $\left(M_{1}, u_{1},+_{1}\right)$ and $\left(M_{2}, u_{2},+_{2}\right)$ be two monoids, with corresponding $\mathcal{L}$ algebras $\alpha_{1}: \mathcal{L}\left(M_{1}\right) \rightarrow M_{1}$ and $\alpha_{2}: \mathcal{L}\left(M_{2}\right) \rightarrow M_{2}$. A function $f: M_{1} \rightarrow M_{2}$ is then a homomorphism of monoids if and only if the diagram

$$
\begin{array}{cc}
\mathcal{L}\left(M_{1}\right) \xrightarrow{\mathcal{L}(f)} & \mathcal{L}\left(M_{2}\right)  \tag{1.10}\\
\alpha_{1} \downarrow \\
M_{1} \xrightarrow{\substack{2}} & \begin{array}{c}
\alpha_{2} \\
M_{2} .
\end{array}
\end{array}
$$

commutes.

This result says that instead of giving a binary operation + with an identity element $u$ we can give a single operation $\alpha$ 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 $(\boxed{1.9}$. We shall see that same diagrams also work for other types of monoids (and collection types).

Proof. 1 If $(X, u,+)$ is a monoid, we can define $\alpha: \mathcal{L}(X) \rightarrow X$ in one go as $\alpha\left(\left[x_{1}, \ldots, x_{n}\right]\right):=x_{1}+\cdots+x_{n}$. 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 (1.9). From a more abstract perspective we define $\alpha: \mathcal{L}(X) \rightarrow X$ via freeness (1.8).

In the other direction, assume an $\mathcal{L}$-algebra $\alpha: \mathcal{L}(X) \rightarrow X$. 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 (1.9):

$$
\begin{aligned}
x+u=\alpha([x, \alpha([])]) & \stackrel{\boxed{11.9}=}{=} \alpha([\alpha(\operatorname{unit}(x)), \alpha([])]) \\
& =\alpha(\mathcal{L}(\alpha)([[x],[]])) \\
& \stackrel{[1.9}{-} \alpha(\operatorname{flat}([[x],[]]))=\alpha([x]) \stackrel{[1.9}{=} x .
\end{aligned}
$$

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

$$
\begin{aligned}
x+(y+z)=\alpha([x, \alpha([y, z])]) & \stackrel{\boxed{1.9}}{=} \alpha([\alpha(\text { unit }(x)), \alpha([y, z])]) \\
& =\alpha(\mathcal{L}(\alpha)([[x],[y, z]])) \\
& \stackrel{\text { I1.9) }}{=} \alpha(\text { flat }([[x],[y, z]])) \\
& =\alpha([x, y, z]) \\
& =\alpha(\text { flat }([[x, y],[z]])) \\
& \stackrel{\text { I1.9 }}{=} \alpha(\mathcal{L}(\alpha)([[x, y],[z]])) \\
& =\alpha([\alpha([x, y]), \alpha(\text { unit }(z))]) \\
& \stackrel{\text { I.9) }}{=} \alpha([\alpha([x, y]), z])=(x+y)+z .
\end{aligned}
$$

2 Now let $f: M_{1} \rightarrow M_{2}$ be a homomorphism of monoids. Diagram 1.10)
then commutes:

$$
\begin{aligned}
\left(\alpha_{2} \circ \mathcal{L}(f)\right)\left(\left[x_{1}, \ldots, x_{n}\right]\right) & =\alpha_{2}\left(\left[f\left(x_{1}\right), \ldots, f\left(x_{n}\right)\right]\right) \\
& =f\left(x_{1}\right)+\cdots+f\left(x_{n}\right) \\
& =f\left(x_{1}+\cdots+x_{n}\right) \quad \text { since } f \text { is a homomorphism } \\
& =\left(f \circ \alpha_{1}\right)\left(\left[x_{1}, \ldots, x_{n}\right]\right) .
\end{aligned}
$$

In the other direction, if 1.10 commutes for a function $f: M_{1} \rightarrow M_{2}$, then $f$ is a homomorphism of monoids, since:

$$
f\left(u_{1}\right)=f\left(\alpha_{1}([])\right) \stackrel{1.10}{=} \alpha_{2}(\mathcal{L}(f)([]))=\alpha_{2}([])=u_{2} .
$$

Similarly one checks that sums are preserved:

$$
\begin{aligned}
f\left(x+{ }_{1} y\right)=f\left(\alpha_{1}([x, y])\right) & \stackrel{\boxed{11.10}}{=} \alpha_{2}(\mathcal{L}(f)([x, y])) \\
& =\alpha_{2}([f(x), f(y)])=f(x)+{ }_{2} f(y) .
\end{aligned}
$$

We see that the algebraic structure $(M, u,+)$ on the set $M$ is expressed as an algebra $\alpha: \mathcal{L}(M) \rightarrow M$, 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 shall use the length $\|\ell\| \in \mathbb{N}$ of a list $\ell$, see Exercise 1.4 .4 for more details, and also the sum and product of a list of natural numbers, defined as:

$$
\begin{aligned}
\operatorname{sum}\left(\left[n_{1}, \ldots, n_{k}\right]\right) & :=n_{1}+\cdots+n_{k}=\sum_{i} n_{i} \\
\operatorname{prod}\left(\left[n_{1}, \ldots, n_{k}\right]\right) & :=n_{1} \cdot \ldots \cdot n_{k}=\prod_{i} n_{i} .
\end{aligned}
$$

See also 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: $\mathcal{L}\left(\mathbb{N}_{>0}\right) \rightarrow \mathbb{N}$ of the above sum and product functions.

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

$$
\operatorname{sum}^{-1}(N):=\left\{\ell \in \mathcal{L}\left(\mathbb{N}_{>0}\right) \mid \operatorname{sum}(\ell)=N\right\} .
$$

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

$$
\begin{equation*}
[1,1,1,1],[1,1,2],[1,2,1],[2,1,1],[2,2],[1,3],[3,1],[4] . \tag{1.11}
\end{equation*}
$$

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 (1.11).
Here is a first, easy counting result.
Lemma 1.4.7. For $N \in \mathbb{N}_{>0}$, the subset $\operatorname{sum}^{-1}(N) \subseteq \mathcal{L}\left(\mathbb{N}_{>0}\right)$ has $2^{N-1}$ elements, that is, $\left|\operatorname{sum}^{-1}(N)\right|=2^{N-1}$.

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 $K \in \mathbb{N}$,

$$
\begin{equation*}
\sum_{0 \leq k \leq K} \frac{1}{2^{k}}=\frac{2^{K+1}-1}{2^{K}} \tag{*}
\end{equation*}
$$

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

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

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 this result is postponed. It involves rather complex probability distributions on multiset partitions, see Corollary ??. A proof using distributions on coin lists may be found in [85]. We are not aware of a (more) elementary proof.

Theorem 1.4.8. For each $N \in \mathbb{N}_{>0}$,

$$
\begin{equation*}
\sum_{\ell \in \operatorname{sum}^{-1}(N)} \frac{1}{\|\ell\|!\cdot \operatorname{prod}(\ell)}=1 . \tag{1.12}
\end{equation*}
$$

At this stage we only give an example, for $N=4$, using the corresponding lists $\ell$ in (1.11). The associated sum (1.12) is illustrated below.

| $[1,1,1,1]$ | $[1,1,2]$ | $[1,2,1]$ | $[2,1,1]$ | $[2,2]$ | $[1,3]$ | $[3,1]$ | $[4]$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\frac{I}{4!\cdot 1}$ | $\frac{1}{3!\cdot 2}$ | $\frac{1}{3!\cdot 2}$ | $\frac{1}{3!\cdot 2}$ | $\frac{1}{2!\cdot 4}$ | $\frac{1}{2!\cdot 3}$ | $\frac{1}{2!\cdot 3}$ | $\frac{1}{1!\cdot 4}$ |
| $\\|$ | $\\|$ | $\\|$ | $\\|$ | $\\|$ | $\\|$ | $\\|$ | $\\|$ |
| $\frac{1}{24}$ | $\frac{1}{12}$ | $\frac{1}{12}$ | $\frac{1}{12}$ | $\frac{1}{8}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{4}$ |
|  |  |  |  |  |  |  |  |

Obviously, elements in a list are ordered. Thus, in 1.11 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:

$$
\begin{equation*}
\operatorname{supp}[K]\left(\left[x_{1}, \ldots, x_{K}\right]\right):=\left\{x_{1}, \ldots, x_{K}\right\} . \tag{1.13}
\end{equation*}
$$

For instance, $\operatorname{supp}[5](\langle a, a, b, c, b\rangle)=\{a, b, c\}$. Clearly, this support map ignores the multiplicity of occurrences of elements in its input list. We can describe support as a function $\operatorname{supp}[K]: X^{K} \rightarrow \mathcal{P}(X)$, where $\mathcal{P}$ is used for powerset. The parameter $K$ in $\operatorname{supp}[K]$, 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 $\left\{\begin{array}{l}K \\ n\end{array}\right\}$ 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:

$$
\left|\operatorname{supp}[K]^{-1}(U)\right|=|U|!\cdot\left\{\begin{array}{c}
K  \tag{1.14}\\
|U|
\end{array}\right\} .
$$

For instance, for the set $U=\{a, b\}$ and number $K=4$ there are $|U|!\cdot\left\{\left\{\begin{array}{c}K \\ |U|\end{array}\right\}=\right.$ $2!\cdot\left\{\begin{array}{l}4 \\ 2\end{array}\right\}=2 \cdot 7=14$ lists of length 4 with support $U$, namely:
$[a, a, a, b][a, a, b, a][a, a, b, b][a, b, a, a][a, b, a, b][a, b, b, a][a, b, b, b]$ $[b, a, a, a][b, a, a, b][b, a, b, a][b, a, b, b][b, b, a, a][b, b, a, b][b, b, b, a]$.

## Exercises

1.4.1 Let $X=\{a, b, c\}$ and $Y=\{u, v\}$ be sets with a function $f: X \rightarrow Y$ 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:

- $\ell_{1}+\ell_{2}$
- $\ell_{2}+\ell_{1}$
- $\ell_{1}+\ell_{1}$
- $\ell_{1}+\left(\ell_{2}+\ell_{1}\right)$
- $\left(\ell_{1}+\ell_{2}\right)+\ell_{1}$
- $\mathcal{L}(f)\left(\ell_{1}\right)$
- $\mathcal{L}(f)\left(\ell_{2}\right)$
- $\mathcal{L}(f)\left(\ell_{1}\right)+\mathcal{L}(f)\left(\ell_{2}\right)$
- $\mathcal{L}(f)\left(\ell_{1}+\ell_{2}\right)$.
1.4.2 We write $\log$ for the logarithm function with some base $b>0$, so that $\log (x)=y$ iff $x=b^{y}$. Verify that the logarithm function $\log$ is a map of monoids:

$$
\left(\mathbb{R}_{>0}, 1, \cdot\right) \xrightarrow{\log }(\mathbb{R}, 0,+)
$$

Often the log function is used to simplify a computation, by turning multiplications into additions. Then one uses that $\log$ is precisely this homomorphism of monoids. (An additonal useful property is that log is monotone: it preserves the order.)
1.4.3 Consider the 'root of squares' function $r$ s: $\mathbb{R}_{>0} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}$ defined by:

$$
r s(x, y)=\sqrt{x^{2}+y^{2}}
$$

Show that this yields a commutative semigroup structure on $\mathbb{R}_{>0}$.
1.4.4 Define a length function $\|-\|: \mathcal{L}(X) \rightarrow \mathbb{N}$ 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 $X \rightarrow \mathbf{1}$ 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 $\mathcal{L}(\mathcal{L}(X)) \rightarrow \mathcal{L}(X)$ can be described in terms of concatenation + as:

$$
\operatorname{flat}\left(\left[\ell_{1}, \ldots, \ell_{n}\right]\right)=\ell_{1}+\cdots+\ell_{n}
$$

2 Now consider the correspondence of Proposition 1.4.6 (1). Conclude that the algebra $\alpha: \mathcal{L}(\mathcal{L}(X)) \rightarrow \mathcal{L}(X)$ associated with the monoid ( $\mathcal{L}(X)$, [], +) from Lemma 1.4.2, (1) is flat.
1.4.6 Check Equation 1.12 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:


1 Describe these maps explicitly on a sequence $\left[n_{1}, \ldots, n_{k}\right]$ 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:

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

In other words, the following diagram commutes.


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. We use the natural numbers $\mathbb{N}$ as additive monoid. Check that there is a monoid action $\mathbb{N} \times M \rightarrow$ $M$, given by $(n, a) \mapsto n \cdot a=a+\cdots+a, n$ times.
1.4.9 Let $X$ be an arbitrary set.

1 Show that the set $X^{X}$ of functions $X \rightarrow X$ forms a monoid, via composition, with the identity function as identity element.
2 Let $M$ be a monoid. For a function $\alpha: X \times M \rightarrow X$ define $\bar{\alpha}: M \rightarrow$ $X^{X}$ by $\alpha(a)(x)=\alpha(x, a)$. Show that $\alpha$ is a monoid action if and only if $\bar{\alpha}$ is a map of monoids.
3 Let $X \times Y \rightarrow X$ be an arbitrary function. Turn it into a monoid action $X \times Y^{\star} \rightarrow X$, 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:

$$
\begin{align*}
\mathcal{P}_{*}(X) & :=\{U \in \mathcal{P}(X) \mid U \neq \emptyset\} \\
\mathcal{P}_{\text {fin }}(X) & :=\{U \in \mathcal{P}(X) \mid U \text { is finite }\} \\
\mathcal{P}[K](X) & :=\left\{U \in \mathcal{P}_{\text {fin }}(X)| | U \mid=K\right\}  \tag{1.15}\\
\mathcal{P}_{s}[K] & :=\left\{U \in \mathcal{P}_{\text {fin }}(X)|1 \leq|U| \leq K\} .\right.
\end{align*}
$$

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

First of all, $\mathcal{P}$ is a functor: it works both on sets and on functions. Given a function $f: X \rightarrow Y$ we can define a new function $\mathcal{P}(f): \mathcal{P}(X) \rightarrow \mathcal{P}(Y)$ 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:

$$
\mathcal{P}(\text { id })=\text { id } \quad \mathcal{P}(g \circ f)=\mathcal{P}(g) \circ \mathcal{P}(f) .
$$

We can use functoriality for marginalisation: for a subset (relation) $R \subseteq X \times Y$ on a product set we get its first marginal $\mathcal{P}\left(\pi_{1}\right)(R) \in \mathcal{P}(X)$ as the subset:

$$
\mathcal{P}\left(\pi_{1}\right)(R)=\left\{\pi_{1}(z) \mid z \in R\right\}=\left\{\pi_{1}(x, y) \mid(x, y) \in R\right\}=\{x \mid \exists y .(x, y) \in R\} .
$$

The next topic is the monoid structure on powersets. The first result is an analogue of Lemma 1.4 .2 and 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 $\mathcal{P}(f): \mathcal{P}(X) \rightarrow \mathcal{P}(Y)$ is a map of monoids, for $f: X \rightarrow Y$.
Both properties also hold for $\mathcal{P}_{\text {fin }}$ instead of $\mathcal{P}$.
Next we define unit and flatten maps for subsets, much like for lists. The function unit: $X \rightarrow \mathcal{P}(X)$ sends an element to a singleton subset: unit $(x):=$ $\{x\}$. The flatten function flat: $\mathcal{P}(\mathcal{P}(X)) \rightarrow \mathcal{P}(X)$ is given by union: for $A \subseteq$ $\mathcal{P}(X)$,

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

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

## Lemma 1.5.2.

1 For each function $f: X \rightarrow Y$ the 'naturality' diagrams


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



### 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 (1.13), given by $\operatorname{supp}\left(\left[x_{1}, \ldots, x_{n}\right]\right)=\left\{x_{1}, \ldots, x_{n}\right\}$. It forms a function $\operatorname{supp}: \mathcal{L}(X) \rightarrow \mathcal{P}_{\text {fin }}(X)$.
There is no systematic way to go in the other direction, via a function $\mathcal{P}_{\text {fin }}(X) \rightarrow \mathcal{L}(X)$ 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 2.4.13 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: $\mathcal{L}(X) \rightarrow \mathcal{P}_{\text {fin }}(X)$, for a fixed set $X$.

1 It is a map of monoids $(\mathcal{L}(X),[],+) \rightarrow\left(\mathcal{P}_{\text {fin }}(X), \emptyset, \cup\right)$.
2 It is natural, in the sense that for $f: X \rightarrow Y$ one has:


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



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

$$
\begin{aligned}
\left(\mathcal{P}_{\text {fin }}(f) \circ \operatorname{supp}\right)\left(\left[x_{1}, \ldots, x_{n}\right]\right) & =\mathcal{P}_{\text {fin }}(f)\left(\operatorname{supp}\left(\left[x_{1}, \ldots, x_{n}\right]\right)\right) \\
& =\mathcal{P}_{\text {fin }}(f)\left(\left\{x_{1}, \ldots, x_{n}\right\}\right) \\
& =\left\{f\left(x_{1}\right), \ldots, f\left(x_{n}\right)\right\} \\
& =\operatorname{supp}\left(\left[f\left(x_{1}\right), \ldots, f\left(x_{n}\right)\right]\right) \\
& =\operatorname{supp}\left(\mathcal{L}(f)\left(\left[x_{1}, \ldots, x_{n}\right]\right)\right) \\
& =(\operatorname{supp} \circ \mathcal{L}(f))\left(\left[x_{1}, \ldots, x_{n}\right]\right) .
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
& (\text { flat } \circ \operatorname{supp} \circ \mathcal{L}(\operatorname{supp}))\left(\left[\left[x_{11}, \ldots, x_{1 n_{1}}\right], \ldots,\left[x_{k 1}, \ldots, x_{k n_{k}}\right]\right]\right) \\
& =(\cup \circ \operatorname{supp})\left(\left[\operatorname{supp}\left(\left[x_{11}, \ldots, x_{1 n_{1}}\right]\right), \ldots, \operatorname{supp}\left(\left[x_{k 1}, \ldots, x_{k n_{k}}\right]\right)\right]\right) \\
& =(\cup \circ \operatorname{supp})\left(\left[\left\{x_{11}, \ldots, x_{1 n_{1}}\right\}, \ldots,\left\{x_{k 1}, \ldots, x_{k n_{k}}\right\}\right]\right) \\
& =\bigcup\left(\left\{\left\{x_{11}, \ldots, x_{1 n_{1}}\right\}, \ldots,\left\{x_{k 1}, \ldots, x_{k n_{k}}\right\}\right\}\right) \\
& =\left\{x_{11}, \ldots, x_{1 n_{1}}, \ldots, x_{k 1}, \ldots, x_{k n_{k}}\right\} \\
& =\operatorname{supp}\left(\left[x_{11}, \ldots, x_{1 n_{1}}, \ldots, x_{k 1}, \ldots, x_{k n_{k}}\right]\right) \\
& =(\operatorname{supp} \circ \text { flat })\left(\left[\left[x_{11}, \ldots, x_{1 n_{1}}\right], \ldots,\left[x_{k 1}, \ldots, x_{k n_{k}}\right]\right]\right) .
\end{aligned}
$$

### 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 $\mathcal{P}_{\text {fin }}(X)$ 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 $f: X \rightarrow M$ between them. Then there is a unique homomorphism of monoids $\bar{f}:\left(\mathcal{P}_{\text {fin }}(X), \emptyset, \cup\right) \rightarrow(M, 0,+)$ with $\bar{f} \circ$ unit $=f$. We represent this situation in the diagram below.


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

$$
\bar{f}\left(\left\{x_{1}, \ldots, x_{n}\right\}\right):=f\left(x_{1}\right)+\cdots+f\left(x_{n}\right), \quad \text { with special case } \quad \bar{f}(\emptyset)=0 .
$$

The order in the above sum $f\left(x_{1}\right)+\cdots+f\left(x_{n}\right)$ does not matter since $M$ is commutative. The function $\bar{f}$ 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 $\mathcal{P}_{\text {fin }}$-algebra $\alpha: \mathcal{P}_{\text {fin }}(X) \rightarrow X$, namely so that the diagrams


commute.
2 Let $\left(M_{1}, u_{1},+_{1}\right)$ and $\left(M_{2}, u_{2},+_{2}\right)$ be two commutative idempotent monoids, with corresponding $\mathcal{P}_{\text {fin }}$-algebras $\alpha_{1}: \mathcal{P}_{\text {fin }}\left(M_{1}\right) \rightarrow M_{1}$ and $\alpha_{2}: \mathcal{P}_{\text {fin }}\left(M_{2}\right) \rightarrow$ $M_{2}$. A function $f: M_{1} \rightarrow M_{2}$ is a map of monoids if and only if the rectangle

$$
\begin{gather*}
\mathcal{P}_{\text {fin }}\left(M_{1}\right) \xrightarrow{\mathcal{P}_{\text {fin }}(f)} \mathcal{P}_{\text {fin }}\left(M_{2}\right)  \tag{1.18}\\
\begin{array}{c}
\alpha_{1} \downarrow \\
M_{1} \\
\\
\\
\\
\\
\\
M_{2}
\end{array}
\end{gather*}
$$

commutes.
Proof. This works very much like in the proof of Proposition 1.4.6. If $(X, u,+)$ is a monoid, we define $\alpha: \mathcal{P}_{\text {fin }}(X) \rightarrow X$ by freeness as $\alpha\left(\left\{x_{1}, \ldots, x_{n}\right\}\right):=x_{1}+$ $\cdots+x_{n}$. In the other direction, given $\alpha: \mathcal{P}_{\text {fin }}(X) \rightarrow X$ 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 $\mathcal{P}_{\text {fin }}$. One can also consider algebras $\mathcal{P}(X) \rightarrow X$ for the (general) powerset functor $\mathcal{P}$. Such algebras turn the set $X$ into a complete lattice, see [10, 104, 127] 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 $\operatorname{extr}_{1}(R): X \rightarrow \mathcal{P}(Y)$ and extr $2(R): Y \rightarrow \mathcal{P}(X)$. The are given by:

$$
\operatorname{extr}_{1}(R)(x):=\{y \in Y \mid(x, y) \in R\} \quad \operatorname{extr}_{2}(R)(y):=\{x \in X \mid(x, y) \in R\}
$$

In fact, one can easily reconstruct the relation $R$ from extr ${ }_{1}(R)$, and also from extr $_{2}(R)$, via:

$$
R=\left\{(x, y) \mid y \in \operatorname{extr}_{1}(R)(x)\right\}=\left\{(x, y) \mid x \in \operatorname{extr}_{2}(R)(y)\right\} .
$$

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:

$$
\begin{equation*}
\mathcal{P}(Y)^{X} \cong \mathcal{P}(X \times Y) \cong \mathcal{P}(X)^{Y} . \tag{1.19}
\end{equation*}
$$

Functions of the form $X \rightarrow \mathcal{P}(Y)$ will later be called 'channels' from $X$ to $Y$, see Section 1.10 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 $S P(X)$ for the set of set partitions of $X$ :

$$
\begin{align*}
S P(X):=\{P \subseteq \mathcal{P}(X) \mid \cup & P=X \text { and } \forall B \in P . B \neq \emptyset \text { and }  \tag{1.20}\\
& \left.\forall B, B^{\prime} \in P . B \neq B^{\prime} \Rightarrow B \cap B^{\prime}=\emptyset\right\} .
\end{align*}
$$

The subsets $B \in P$ of $X$, for a partition $P \in S P(X)$, are often called blocks.
2 For a number $K \geq 1$ we simply write $S P(K)=S P(\{1, \ldots, K\})$ 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 $\operatorname{size}[K]: S P(K) \rightarrow(K]$, where $\operatorname{size}[K](P):=|P|$ is the number of blocks in a partition $P \in S P(K)$. The parameter $K$ in size $[K]$ is omitted if it is clear from the context.

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

$$
\{\{1\},\{2\},\{3\}\} \quad\{\{1,2\},\{3\}\} \quad\{\{1,3\},\{2\}\} \quad\{\{1\},\{2,3\}\} \quad\{\{1,2,3\}\} .
$$

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

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

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

Proposition 1.5.7. Let $K \geq 1$ and $1 \leq n \leq K$. The number of set partitions in $S P(K)$ of size $n$ is given by the Stirling number $\left\{\begin{array}{l}K \\ n\end{array}\right\}$ of the second kind:

$$
\left|\operatorname{size}[K]^{-1}(n)\right|=\left\{\begin{array}{l}
K \\
n
\end{array}\right\} \quad \text { where } \quad \operatorname{size}[K]: S P(K) \longrightarrow\{1, \ldots, K\}
$$

Proof. We show, by induction on $K \geq 1$, that the size expression $\left|\operatorname{size}[K]^{-1}(n)\right|$ satisfies the recurrence equations (1.3), 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:

$$
\left|\operatorname{size}[1]^{-1}(1)\right|=1=\left\{\begin{array}{l}
1 \\
1
\end{array}\right\} .
$$

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

$$
\left|\operatorname{size}[K+1]^{-1}(1)\right|=1=\left\{\begin{array}{c}
K+1 \\
1
\end{array}\right\}, \quad \text { by Lemma 1.2.5 (2) } .
$$

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

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

Thus we can perform the induction step as follows.

$$
\begin{aligned}
\left|\operatorname{size}[K+1]^{-1}(n)\right| & =n \cdot\left|\operatorname{size}[K]^{-1}(n)\right|+\left|\operatorname{size}[K]^{-1}(n-1)\right| \\
& \stackrel{\text { (IH) }}{=} n \cdot\left\{\begin{array}{l}
K \\
n
\end{array}\right\}+\left\{\begin{array}{c}
K \\
n-1
\end{array}\right\} \\
& \stackrel{\text { I.3) }}{=}\left\{\begin{array}{c}
K+1 \\
n
\end{array}\right\} .
\end{aligned}
$$

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 [3, 152, 92].

Definition 1.5.8. Fix an arbitrary set $X$ and a number $K \geq 1$. There is a match function $\operatorname{mat}[K]: X^{K} \rightarrow S P(K)$ defined as:

$$
\begin{equation*}
\operatorname{mat}[K]\left(x_{1}, \ldots, x_{K}\right):=\left\{\left\{i \mid x_{i}=x_{1}\right\}, \ldots,\left\{i \mid x_{i}=x_{K}\right\}\right\} . \tag{1.21}
\end{equation*}
$$

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:

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

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.


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

$$
\begin{equation*}
\left|\operatorname{mat}^{-1}(P)\right|=(n)_{|P|}=\frac{n!}{(n-|P|)!} . \tag{1.23}
\end{equation*}
$$

Proof. 1 The number of blocks of $\operatorname{mat}(\vec{x})$ 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 (1.21).
2 By an elementary counting argument: given a set partition $P \in S P(X)$ 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 $r \in \mathbb{R}$,

$$
\begin{equation*}
\sum_{P \in S P(X)} r^{|P|} \cdot \prod_{B \in P}(|B|-1)!=(r)^{|X|}=\prod_{0 \leq i<|X|} r+i . \tag{1.24}
\end{equation*}
$$

In particular, for $r=1$,

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

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 (1.24) 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 1.24 holds for $Y$.

$$
\begin{aligned}
& \sum_{P \in S P(X)} r^{|P|} \cdot \prod_{B \in P}(|B|-1)! \\
& =\sum_{Q \in S P(Y)} r^{|Q \cup\{\{x\}\}|} . \prod_{B \in Q \cup\{\{x\}\}}(|B|-1)! \\
& +\sum_{Q \in S P(Y)} \sum_{C \in Q} r^{|(Q \backslash\{C\}) \cup\{C \cup\{x\}\}|} . \prod_{B \in(Q \backslash\{C\}) \cup\{C \cup\{x\}\}}(|B|-1)! \\
& =\sum_{Q \in S P(Y)} r^{|Q|+1} \cdot \prod_{B \in Q}(|B|-1)!+\sum_{Q \in S P(Y)} \sum_{C \in Q} r^{|Q|} \cdot|C| \cdot \prod_{B \in Q}(|B|-1)! \\
& \stackrel{(\mathrm{IH})}{=} r \cdot\left(\prod_{0 \leq i<|Y|} r+i\right)+|Y| \cdot\left(\prod_{0 \leq i<|Y|} r+i\right) \quad \text { since } \sum_{C \in Q}|C|=|Y| \\
& =(r+|Y|) \cdot\left(\prod_{0 \leq i<|Y|} r+i\right) \\
& =\prod_{0 \leq i<|X|} r+i .
\end{aligned}
$$

## Remark 1.5.11.

1 As a consequence of Proposition 1.5 .7 we can determine the number of set partitions of $K$ as:

$$
|S P(K)|=\sum_{1 \leq n \leq K}\left|\operatorname{size}[K]^{-1}(n)\right|=\left\{\begin{array}{c}
K \\
1
\end{array}\right\}+\cdots+\left\{\begin{array}{l}
K \\
K
\end{array}\right\} .
$$

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 $S P(1), S P(2), S P(3), \ldots$

2 Eearlier we skipped the proof of Lemma 1.4.9. 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 $\operatorname{supp}(\ell)=U$. Well, if $\ell$ contains $n$ different elements, then the associated set partition $\operatorname{mat}(\ell) \in$ $S P(K)$ will have $n$ blocks. Proposition 1.5 .7 tells that there are $\left\{\begin{array}{l}K \\ n\end{array}\right\}$ such set partitions. There are $n$ ! many lists $\ell$ with elements from $U$ that match to a particular set partition. Hence in total we have $n!\cdot\left\{\begin{array}{l}K \\ n\end{array}\right\}=|U|!\cdot\left\{\begin{array}{c}K \\ |U|\end{array}\right\}$ many lists, as claimed in Lemma 1.4.9

## Exercises

1.5.1 Continuing Exercise 1.4.1, compute:

- $\operatorname{supp}\left(\ell_{1}\right)$
- $\operatorname{supp}\left(\ell_{2}\right)$
- $\operatorname{supp}\left(\ell_{1}+\ell_{2}\right)$
- $\operatorname{supp}\left(\ell_{1}\right) \cup \operatorname{supp}\left(\ell_{2}\right)$
- $\operatorname{supp}\left(\mathcal{L}(f)\left(\ell_{1}\right)\right)$
- $\mathcal{P}_{\text {fin }}(f)\left(\operatorname{supp}\left(\ell_{1}\right)\right)$.
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 $\neg: \mathcal{P}(X) \rightarrow \mathcal{P}(X)$, given by:

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

is a homomorphism of monoids between $(\mathcal{P}(X), \emptyset, \cup)$ and $(\mathcal{P}, X, \cap)$, in both directions. In fact, if forms an isomorphism of monoids, since $\neg \neg U=U$.
2 Prove that the intersections monoid structure is not preserved by maps $\mathcal{P}(f): \mathcal{P}(X) \rightarrow \mathcal{P}(Y)$.
Hint: Look at preservation of the unit $X \in \mathcal{P}(X)$.
1.5.3 1 Check that in general,

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

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:

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

Check that $\mathcal{P}$ does restrict to a functor $\mathcal{P}_{s}[K]$.
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 $P \in S P(X)$;
- equivalence relations $E \subseteq X \times X$.
- equivalence classes of surjective functions $X \rightarrow A$, for some set $A$ where $h: X \rightarrow A$ and $k: X \rightarrow B$ are equivalent if there is an isomorphism $j: A \xlongequal[Э]{\cong} B$ with $j \circ h=k$.
2 Let $P, Q \in S P(X)$ 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 $S P(X)$ defined in the previous item is a partial order. It has a least and greatest element; find out what they are.
4 Let $f: X \rightarrow Y$ be a surjective function. Show that it gives rise to a function:


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: $X^{K} \rightarrow$ $\mathcal{P}_{s}[K](X)$, for each set $X$ and number $K \geq 1$. Check that these maps are natural in $X$, in the sense that for each function $f: X \rightarrow Y$ the following rectangle commutes.

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:

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

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

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

1.5.8 Prove that Stirling numbers of the first kind arise not only via subsets, as in Exercise 1.2.5, but also via set partitions, as in:

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

### 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. 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 statistical 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); in that case, the order of the words is ignored. 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.
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). 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 solutations may occur multiple times. For instance $x^{3}-7 x^{2}+16 x-12=(x-2)(x-2)(x-3)$ has as multiset of solutions $2|2\rangle+1|3\rangle$.

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. 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 $\mathrm{C}_{2} \mathrm{H}_{4} \mathrm{O}_{2}$ 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:

$$
n_{1}\left|x_{1}\right\rangle+\cdots+n_{k}\left|x_{k}\right\rangle \quad \text { where } \quad n_{i} \in \mathbb{N} \text { and } x_{i} \in X
$$

This expression is a formal sum, not an actual sum (for instance in $\mathbb{R}$ ). We may write it as $\sum_{i} n_{i}\left|x_{i}\right\rangle$. 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 brackets (if any) in a sum do not matter.

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 $\varphi: X \rightarrow \mathbb{N}$ that has finite support, where the support $\operatorname{supp}(\varphi) \subseteq X$ is the subset $\operatorname{supp}(\varphi)=\{x \in X \mid \varphi(x) \neq 0\}$ where $\varphi$ is non-zero. Notice that the set $X$ may be infinite, but each multiset $\varphi$ over $X$ is required to have finite support $\operatorname{supp}(\varphi) \subseteq X$.

For each element $x \in X$ the number $\varphi(x) \in \mathbb{N}$ indicates how often $x$ occurs in the multiset $\varphi$. Such a function $\varphi$ can also be written as a formal sum $\sum_{x} \varphi(x)|x\rangle$, where $x$ ranges over $\operatorname{supp}(\varphi)$.
For instance, the multiset $9|b\rangle+2|a\rangle$ over $A=\{a, b, c\}$ corresponds to the function $\varphi: A \rightarrow \mathbb{N}$ 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 $\varphi$ 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:

$$
\mathcal{M}(X):=\left\{\varphi: X \rightarrow \mathbb{R}_{\geq 0} \mid \operatorname{supp}(\varphi) \text { is finite }\right\} .
$$

The functions in $\mathcal{M}(X)$ may be called mass functions, as in [161]. 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:

$$
\begin{aligned}
\mathcal{M}_{*}(X) & :=\{\varphi \in \mathcal{M}(X) \mid \operatorname{supp}(\varphi) \text { is non-empty }\} \\
& =\left\{\varphi: X \rightarrow \mathbb{R}_{\geq 0} \mid \operatorname{supp}(\varphi) \text { is finite and non-empty }\right\} .
\end{aligned}
$$

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:

$$
\mathcal{M}[K](X):=\{\varphi \in \mathcal{M}(X) \mid\|\varphi\|=K\} \quad \text { where } \quad\|\varphi\|:=\sum_{x} \varphi(x)
$$

This size number $\|\varphi\|$ gives the total number of elements in the multiset $\varphi$.
5 Finally, when the set $X$ is finite, we use special notation for the subsets of (natural) multisets $\varphi$ with full support, that is with $\operatorname{supp}(\varphi)=X$.

$$
\begin{aligned}
\mathcal{M}_{f s}(X) & :=\{\varphi \in \mathcal{M}(X) \mid \operatorname{supp}(\varphi)=X\} \\
\mathcal{N}_{f s}(X) & :=\{\varphi \in \mathcal{N}(X) \mid \operatorname{supp}(\varphi)=X\} .
\end{aligned}
$$

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 $f: X \rightarrow Y$ we define $\mathcal{M}(f): \mathcal{M}(X) \rightarrow$ $\mathcal{M}(Y)$ 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 $f: X \rightarrow Y$ defines the transformation of colours. It tells that a ball of colour $x \in X$ in $\varphi$ will be repainted with colour $f(x) \in Y$. Notice that two different colours $x, x^{\prime} \in X$ may have the same colour, after repainting, namely when $f(x)=f\left(x^{\prime}\right)$.

Definition 1.6.2. For a function $f: X \rightarrow Y$ 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:

$$
\mathcal{M}(f)\left(\sum_{i} r_{i}\left|x_{i}\right\rangle\right):=\sum_{i} r_{i}\left|f\left(x_{i}\right)\right\rangle
$$

Equivalently, using function notation:

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

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 $\sum_{i} r_{i}\left|f\left(x_{i}\right)\right\rangle$ it may happen that $f\left(x_{i_{1}}\right)=f\left(x_{i_{2}}\right)$ for $x_{i_{1}} \neq x_{i_{2}}$, so that $r_{i_{1}}$ and $r_{i_{2}}$ are added together. Thus, the support of $\mathcal{M}(f)\left(\sum_{i} r_{i}\left|x_{i}\right\rangle\right)$ may have fewer elements than the support of $\sum_{i} r_{i}\left|x_{i}\right\rangle$, but the sum of all multiplicities is the same in $\mathcal{M}(f)\left(\sum_{i} r_{i}\left|x_{i}\right\rangle\right)$ and $\sum_{i} r_{i}\left|x_{i}\right\rangle$, see Exercise 1.7 .2 below.

Applying $\mathcal{M}$ to a projection function $\pi_{i}: X_{1} \times \cdots \times X_{n} \rightarrow X_{i}$ yields a function $\mathcal{M}\left(\pi_{i}\right)$, from the set $\mathcal{M}\left(X_{1} \times \cdots \times X_{n}\right)$ of multisets over a product to the set $\mathcal{M}\left(X_{i}\right)$ of multisets over a component. This $\mathcal{M}\left(\pi_{i}\right)$ 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.

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:

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

These sums restrict to $\mathcal{N}(X)$.
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 $r \cdot(-): \mathcal{M}(X) \rightarrow \mathcal{M}(X)$ preserves the sums $(\mathbf{0},+)$ from the previous item, and is thus a map of monoids.
3 For each $f: X \rightarrow Y$, the function $\mathcal{M}(f): \mathcal{M}(X) \rightarrow \mathcal{M}(Y)$ 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 supp: $\mathcal{M}(X) \rightarrow \mathcal{P}_{\text {fin }}(X)$ is a homomorphism of monoids and is natural in $X$. The latter means that for each function $f: X \rightarrow Y$ the following rectangle commutes.


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 ketnotation. 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.

| $\mathbf{0}$ | $\mathbf{1}$ | $\mathbf{2}$ | $\mathbf{3}$ | $\mathbf{4}$ | $\mathbf{5}$ | $\mathbf{6}$ | $\mathbf{7}$ | $\mathbf{8}$ | $\mathbf{9}$ | $\mathbf{1 0}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 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:


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.


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

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

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.


Next, consider the two-dimensional table (1.28) 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 |  |

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 $\tau$ over the product set / space $B \times M$, that is, as an element $\tau \in \mathcal{N}(B \times M)$, namely:

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

Such a multiset can be plotted in three dimensions as:


We see that Table (1.28) contains 'totals' in its vertical and horizontal margins. They can be obtained from the multiset $\tau$ as marginals, using the functoriality of $\mathcal{N}$. This works as follows. Applying the natural multiset functor $\mathcal{N}$ to the two projections $\pi_{1}: B \times M \rightarrow B$ and $\pi_{2}: B \times M \rightarrow M$ yields marginal distributions on $B$ and $M$, namely:

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

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.

### 1.6.2 Unit and flatten for multisets

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

$$
\text { flat } \left.\left.\left.\left.\left(\frac{1}{3}|2| a\right\rangle+2|c\rangle\right\rangle+5|1| b\right\rangle+\frac{1}{6}|c\rangle\right\rangle\right)=\frac{2}{3}|a\rangle+5|b\rangle+\frac{3}{2}|c\rangle .
$$

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

$$
\operatorname{flat}\left(\sum_{i} r_{i}\left|\varphi_{i}\right\rangle\right):=\sum_{x \in X}\left(\sum_{i} r_{i} \cdot \varphi_{i}(x)\right)|x\rangle .
$$

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 $f: X \rightarrow Y$ the two rectangles


commute.
2 The next two diagrams also commute.


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 $f: X \rightarrow M$ has a unique extension to a homomorphism of monoids $\bar{f}:(\mathcal{N}(X), \mathbf{0},+) \rightarrow(M, 0,+)$ with $\bar{f} \circ$ unit $=f$. The diagram below captures the situation, where the dashed arrow is used for uniqueness.


Proof. One defines:

$$
\bar{f}\left(n_{1}\left|x_{1}\right\rangle+\cdots+n_{k}\left|x_{k}\right\rangle\right):=n_{1} \cdot f\left(x_{1}\right)+\cdots+n_{k} \cdot f\left(x_{k}\right) .
$$

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 $\alpha: \mathcal{N}(X) \rightarrow X$ making the two diagrams below commute.



2 Let $\left(M_{1}, u_{1},+_{1}\right)$ and $\left(M_{2}, u_{2},+_{2}\right)$ be two commutative monoids, with corresponding $\mathcal{N}$-algebras $\alpha_{1}: \mathcal{N}\left(M_{1}\right) \rightarrow M_{1}$ and $\alpha_{2}: \mathcal{N}\left(M_{2}\right) \rightarrow M_{2}$. A function $f: M_{1} \rightarrow M_{2}$ is a map of monoids if and only if the rectangle
commutes.
Proof. Analogously to the proof Proposition 1.4.6 if $(X, u,+)$ is a commutative monoid, we define $\alpha: \mathcal{N}(X) \rightarrow X$ by turning formal sums into actual sums: $\alpha\left(\sum_{i} n_{i}\left|x_{i}\right\rangle\right):=\sum_{i} n_{i} \cdot x_{i}$, see (the proof of) Proposition 1.6.5. In the other direction, given $\alpha: \mathcal{N}(X) \rightarrow X$ 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:

$$
\begin{equation*}
\mathcal{M}(Y)^{X} \cong \mathcal{M}(X \times Y) \cong \mathcal{M}(X)^{Y} . \tag{1.32}
\end{equation*}
$$

This is analogous to 1.19 for powerset.
How does this work in detail? Suppose we have an arbitary multiset / table $\sigma \in \mathcal{M}(X \times Y)$. From $\sigma$ one can extract a function extr ${ }_{1}(\sigma): X \rightarrow \mathcal{M}(Y)$, and also extr $2(\sigma): Y \rightarrow \mathcal{M}(X)$, via:

$$
\operatorname{extr}_{1}(\sigma)(x)=\sum_{y \in Y} \sigma(x, y)|y\rangle \quad \operatorname{extr}_{2}(\sigma)(y)=\sum_{x \in X} \sigma(x, y)|x\rangle
$$

Notice that we are - conveniently - mixing ket and function notation for multisets. Conversely, $\sigma$ can be reconstructed from extr ${ }_{1}(\sigma)$, and also from $\operatorname{extr}_{2}(\sigma)$, via $\sigma(x, y)=\operatorname{extr}_{1}(\sigma)(x)(y)=\operatorname{extr}_{2}(\sigma)(y)(x)$.
Functions of the form $X \rightarrow \mathcal{M}(Y)$ will also be used as channels from $X$ to $Y$, see Section 1.10. That's why we often speak about 'channel extraction'.
As illustration, we apply extraction to the medicine - blood pressure Table 1.28 described as the multiset $\tau \in \mathcal{M}(B \times M)$. It gives rise to two channels extr ${ }_{1}(\tau): B \rightarrow \mathcal{M}(M)$ and extr $2(\tau): M \rightarrow \mathcal{M}(B)$. Explicitly:

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

We see that this extracted function captures the two rows of Table 1.28. Similarly we get the columns via the second extracted function:

$$
\begin{aligned}
& \operatorname{extr}_{2}(\tau)(0)=10|L\rangle+5|H\rangle \\
& \operatorname{extr}_{2}(\tau)(1)=35|L\rangle+10|H\rangle \\
& \operatorname{extr}_{2}(\tau)(2)=25|L\rangle+15|H\rangle
\end{aligned}
$$

## 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 $\mathcal{M}\left(\pi_{1}\right)(\varphi) \in \mathcal{M}(X)$ and $\mathcal{M}\left(\pi_{2}\right)(\varphi) \in \mathcal{M}(Y)$.
1.6.3 Let $P \subseteq \mathbb{N}_{>0}$ be the subset of prime numbers. Describe an essential homomorphism property of prime number factorisation pnf as a function:

$$
\mathbb{N}_{>0} \xrightarrow{p n f} \mathcal{N}(P)
$$

Describe concretely pnf(2023).
1.6.4 Consider the chemical equation for burning methane:

$$
\mathrm{CH}_{4}+2 \mathrm{O}_{2} \longrightarrow \mathrm{CO}_{2}+2 \mathrm{H}_{2} \mathrm{O}
$$

Check that there is an underlying equation of multisets:

$$
\begin{aligned}
& \text { flat }(1|1| C\rangle+4|H\rangle\rangle+2|2| O\rangle\rangle) \\
& \quad=\text { flat }(1|1| C\rangle+2|O\rangle\rangle+2|2| H\rangle+1|O\rangle\rangle)
\end{aligned}
$$

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:

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

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 $\sum_{i} n_{i} x^{i}$ with natural numbers as coefficients - where $x$ is some placeholder variable.
2 Check that addition of multisets corresponds to addition of poynomials.
1.6.7 $\quad$ Check that $\mathcal{M}(i d)=$ 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: $\mathcal{P}_{\text {fin }}(X) \rightarrow \mathcal{N}(X)$ given by $\operatorname{units}\left(\left\{x_{1}, \ldots, x_{n}\right\}\right):=\sum_{i} 1\left|x_{i}\right\rangle$. 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: $\mathcal{N}(\mathcal{N}(X)) \rightarrow$ $\mathcal{N}(X)$.
2 Notice also that this flatten map is a homomorphism of monoids.
1.6.12 Verify that the support map supp: $\mathcal{M}(X) \rightarrow \mathcal{P}_{\text {fin }}(X)$ commutes with
extraction functions, in the sense that the following diagram commutes.


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

$$
\operatorname{extr}_{1}(\operatorname{supp}(\tau))(x)=\operatorname{supp}\left(\operatorname{extr}_{1}(\tau)(x)\right) .
$$

Here we use that $\operatorname{supp}^{X}(f):=\operatorname{supp} \circ f$, so that $\operatorname{supp}^{X}(f)(x)=$ $\operatorname{supp}(f(x))$.
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 $r \cdot(-): M \rightarrow M$, for each $r \in \mathbb{R}_{\geq 0}$, forming a homomorphism of monoids. It is like a vector space, not over all reals, but only over the non-negative reals. Homomorphisms of cones preserve such scalar multiplications.

Let $X$ be a set and $(M, 0,+)$ a cone, with a function $f: X \rightarrow M$. Prove that there is a unique homomorphism of cones $\bar{f}: \mathcal{M}(X) \rightarrow M$ with $\bar{f} \circ$ 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.
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 $\varphi$. For instance, we can look at the size of its support $|\operatorname{supp}(\varphi)| \in \mathbb{N}$, 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 $\varphi$ is a natural multiset. Below we will introduce several more such numbers for natural multisets $\varphi$, namely $\varphi \rrbracket$ and ( $\varphi$ ), and later on also a binomial coefficient $\binom{\psi}{\varphi}$.

## Definition 1.7.1.

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

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

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 $K \in \mathbb{N}$ we use and order $\leq_{K}$ with $K$ as subscript for:

$$
\varphi \leq_{K} \psi \Longleftrightarrow\|\varphi\|=K \text { and } \varphi \leq \psi
$$

3 For a collection of numbers $r=\left(r_{x}\right)_{x \in X}$ we write:

$$
r^{\varphi}:=\prod_{x \in X} r_{x}^{\varphi(x)}=\vec{r}^{\varphi} .
$$

The latter vector notation is appropriate in a situation with a particular order.
4 The factorial $\varphi \rrbracket$ of a natural multiset $\varphi \in \mathcal{N}(X)$ is the product of the factorial of its multiplicities:

$$
\begin{equation*}
\varphi \rrbracket:=\prod_{x \in \operatorname{supp}(\varphi)} \varphi(x)! \tag{1.33}
\end{equation*}
$$

5 The multiset coefficient $(\varphi)$ is defined as:

$$
(\varphi):=\frac{\|\varphi\|!}{\varphi!}=\frac{\|\varphi\|!}{\prod_{x} \varphi(x)!}=\binom{\|\varphi\|}{\varphi\left(x_{1}\right), \ldots, \varphi\left(x_{n}\right)} .
$$

The multinomial coefficient from Definition 1.2.1 (3) is used in the latter formulation; it assumes that the support of $\varphi$ is somehow ordered as $\left[x_{1}, \ldots, x_{n}\right]$.

For instance,

$$
(3|R\rangle+2|B\rangle) \square=3!\cdot 2!=12 \quad \text { and } \quad(3|R\rangle+2|B\rangle)=\frac{5!}{12}=10
$$

In Proposition 1.7 .2 below we make precise how the multinomial coefficient ( $\varphi$ ) in item (5) counts the number of lists that 'accumulate to' the multiset $\varphi$.

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 ( $\varphi$ ). It has our preference.
For a multiset $\varphi$ we have already used the 'support' definition $\operatorname{supp}(\varphi)=$ $\{x \mid \varphi(x) \neq 0\}$. This yields a map supp: $\mathcal{M}(X) \rightarrow \mathcal{P}_{\text {fin }}(X)$, which is wellbehaved, in the sense that it is natural and preserves the monoid structures on $\mathcal{M}(X)$ and $\mathcal{P}_{\text {fin }}(X)$, see Lemma 1.6.3 (4).

We have also seen a support map from lists $\mathcal{L}$ to finite powerset $\mathcal{P}_{\text {fin }}$. This support map factorises through multisets, as described with a new function acc in the following triangle.


The 'accumulator' map acc: $\mathcal{L}(X) \rightarrow \mathcal{N}(X)$ 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. [143, 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 $\ell \in \mathcal{L}(X)$,

$$
\begin{equation*}
\operatorname{acc}\left(x_{1}, \ldots, x_{n}\right):=1\left|x_{1}\right\rangle+\cdots+1\left|x_{n}\right\rangle . \tag{1.35}
\end{equation*}
$$

Since multiple terms $1|x\rangle$ add up we have $\operatorname{acc}(\ell)(x)=n$ if and only if the element $x$ occurs $n$ times in the list $\ell$.
In an example:

$$
\operatorname{acc}(a, b, a, b, c, b, b)=2|a\rangle+4|b\rangle+1|c\rangle .
$$

The above diagram 1.34 reflects an earlier informal statement, namely that multisets are somehow in between lists and subsets.
A basic question is: how many (ordered) sequences of coloured balls give rise to a specific urn with balls? More technically, given a natural multiset $\varphi$, how many lists $\ell$ statisfy $\operatorname{acc}(\ell)=\varphi$ ? In yet another form, what is the size $\left|\operatorname{acc}^{-1}(\varphi)\right|$ of the inverse image? This is where the multiset coefficient $(\varphi)$ from Definition 1.7.1 (5) comes in.

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

$$
\begin{equation*}
X^{K} \xrightarrow{\operatorname{acc}[K]} \mathcal{N}[K](X) . \tag{1.36}
\end{equation*}
$$

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

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

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

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

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

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

Multinomial coefficients satisfy the following recurrence relations.

$$
\begin{equation*}
\binom{K-1}{k_{1}-1, \ldots, k_{n}}+\cdots+\binom{K-1}{k_{1}, \ldots, k_{n}-1}=\binom{K}{k_{1}, \ldots, k_{n}} \tag{1.37}
\end{equation*}
$$

for multinomial coefficients. A snappy re-formulation, for a natural multiset $\varphi$, is:

$$
\begin{equation*}
\sum_{x \in \operatorname{supp}(\varphi)}(\varphi-1|x\rangle)=(\varphi) . \tag{1.38}
\end{equation*}
$$

Multinomial coefficients, see Definition 1.2.1(3), are useful, for instance in the Multinomial Theorem (see e.g. [159]):

$$
\begin{equation*}
\left(r_{1}+\cdots+r_{n}\right)^{K}=\sum_{k_{i}, \sum_{i} k_{i}=K}\binom{K}{k_{1}, \ldots, k_{n}} \cdot r_{1}^{k_{1}} \cdot \ldots \cdot r_{n}^{k_{n}} . \tag{1.39}
\end{equation*}
$$

An equivalent formulation using multisets is:

$$
\begin{align*}
\left(r_{1}+\cdots+r_{n}\right)^{K} & =\sum_{\varphi \in \mathcal{M}[K](\{1, \ldots, n\})}(\varphi) \cdot \vec{r}^{\varphi} \\
& =\sum_{\varphi \in \mathcal{M}[K](\{1, \ldots, n\})}(\varphi) \cdot \prod_{1 \leq i \leq n} r_{i}^{\varphi(i)} . \tag{1.40}
\end{align*}
$$

Remark 1.7.3. The above formulation (1.40) 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 $\mathcal{M}_{ \pm}(X)=\{\psi: X \rightarrow \mathbb{R} \mid \psi$ 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 $\psi \in \mathcal{M}_{ \pm}(X)$,

$$
\begin{equation*}
\|\psi\|^{K}=\sum_{\varphi \in \mathcal{M}[K](X)}(\varphi) \cdot \prod_{x \in X} \psi(x)^{\varphi(x)} . \tag{1.41}
\end{equation*}
$$

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 $r \in[0,1)$,

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

2 As a special case,

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

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

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

4 For $r_{1}, \ldots, r_{m} \in[0,1]$ with $\sum_{i} r_{i}<1$,

$$
\sum_{n_{1}, \ldots, n_{m} \geq 0}\binom{K+\sum_{i} n_{i}}{K, n_{1}, \ldots, n_{m}} \cdot \prod_{i} r_{i}^{n_{i}}=\frac{1}{\left(1-\sum_{i} r_{i}\right)^{K+1}} .
$$

Equivalently,

$$
\sum_{\varphi \in \mathcal{N}(\{1, \ldots, m\})}\binom{K+\|\varphi\|}{K} \cdot(\varphi) \cdot \vec{r}^{\varphi}=\frac{1}{\left(1-\sum_{i} r_{i}\right)^{K+1}}
$$

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 $s_{n} \rightarrow \frac{1}{1-r}$ as $n \rightarrow \infty$.
3 We choose to use the first item, but there are other ways to prove this result, see Exercise 1.7.15

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

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

$$
\begin{aligned}
& \sum_{n_{1}, \ldots, n_{m} \geq 0}\binom{K+\sum_{i} n_{i}}{K, n_{1}, \ldots, n_{m}} \cdot \prod_{i} r_{i}^{n_{i}} \\
= & \sum_{n \geq 0} \sum_{n_{1}, \ldots, n_{m}, \Sigma_{i} n_{i}=n}\binom{K+n}{K, n_{1}, \ldots, n_{m}} \cdot \prod_{i} r_{i}^{n_{i}} \\
= & \sum_{n \geq 0} \sum_{n_{1}, \ldots, n_{m}, \sum_{i} n_{i}=n}\binom{K+n}{K} \cdot\binom{n}{n_{1}, \ldots, n_{m}} \cdot \prod_{i} r_{i}^{n_{i}} \\
= & \sum_{n \geq 0}\binom{K+n}{K} \cdot \sum_{n_{1}, \ldots, n_{m}, \sum_{i} n_{i}=n}\binom{n}{n_{1}, \ldots, n_{m}} \cdot \prod_{i} r_{i}^{n_{i}} \\
& \stackrel{1.39}{=} \sum_{n \geq 0}\binom{K+n}{K} \cdot\left(\sum_{i} r_{i}\right)^{n} \\
= & \frac{1}{\left(1-\sum_{i} r_{i}\right)^{K+1}}, \quad \text { by item (1). }
\end{aligned}
$$

## Exercises

1.7.1 Consider the function $f:\{a, b, c\} \rightarrow\{0,1\}$ 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:

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

Conclude that $(\varphi) \neq(\mathcal{N}(f)(\varphi))$, in general.
2 Now take $\psi=2|0\rangle+3|1\rangle \in \mathcal{N}(\{0,1\})$.

- Compute ( $\psi$ ).
- Show that there are four multisets $\varphi_{1}, \varphi_{2}, \varphi_{3}, \varphi_{4} \in \mathcal{N}(\{a, b, c\})$ with $\mathcal{M}(f)\left(\varphi_{i}\right)=\psi$, for each $i$.
- Check that $\left.\left.\left.\left.\mathbf{(} \psi) \neq \mathbf{(} \varphi_{1}\right)+\mathbf{(} \varphi_{2}\right)+\mathbf{(} \varphi_{3}\right)+\mathbf{(} \varphi_{4}\right)$.

What is the general formulation now?
1.7.2 Check that:

1 the size map $\|-\|: \mathcal{M}(X) \rightarrow \mathbb{R}_{\geq 0}$ 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 Show that for natural multisets $\varphi, \psi \in \mathcal{N}(X)$,

$$
\varphi \leq \psi \Longleftrightarrow \exists \varphi^{\prime} \in \mathcal{N}(X) \cdot \varphi+\varphi^{\prime}=\psi .
$$

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

1 Show that:

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

2 Derive the recurrence relation 1.38 from this equation.
1.7.6 Prove the multiset formulation 1.40 of the Multinomial Theorem, by induction on $K$.
1.7.7 Let $K \geq 0$ and let set $X$ have $n \geq 1$ elements. Prove that:

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

This generalises the well known sum-formula for binomial coefficients: $\sum_{0 \leq k \leq K}\binom{K}{k}=2^{K}$, for $n=2$. Hint: Use $n=1+\cdots+1$ in (1.40).
1.7.8 Fix $N \geq 0$ and consider the obvious addition function $\operatorname{sum}_{N}: \mathbb{N}^{N} \rightarrow$ $\mathbb{N}$ given by $\operatorname{sum}_{N}\left(k_{1}, \ldots, k_{N}\right)=\sum_{i} k_{i}$.

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

$$
\left|\left(\operatorname{sum}_{N}\right)^{-1}(m)\right|=N^{m} .
$$

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

$$
\left|\left(\operatorname{sum}_{N}^{*}\right)^{-1}(m)\right|=\binom{m-1}{N-1} .
$$

1.7.9 Let $\varphi$ be a natural multiset. Show that:
$1 \mathbf{( \varphi )}=1 \Longleftrightarrow \operatorname{supp}(\varphi)$ is a singleton;
$2(\varphi)=\|\varphi\|!\Longleftrightarrow \forall x \in \operatorname{supp}(\varphi) . \varphi(x)=1 \Longleftrightarrow \varphi$ consists of singletons.
1.7.10 Show that $\|\operatorname{acc}(\ell)\|=\|\ell\|$, using the length $\|\ell\|$ of a list $\ell$ from Exercise 1.4.4
1.7.11 Consider the $K$-ary accumulator function (1.36, for $K>0$. Check that acc is invariant under transposition, in the sense that for each permutation $\pi$ of the positions $\{1, \ldots, K\}$ one has:

$$
\operatorname{acc}\left(x_{1}, \ldots, x_{K}\right)=\operatorname{acc}\left(x_{\pi(1)}, \ldots, x_{\pi(K)}\right)
$$

1.7.12 1 Check that the accumulator map acc: $\mathcal{L}(X) \rightarrow \mathcal{N}(X)$ is a homomorphism of monoids.
2 Check also that it arises via freeness of lists, as extension of the function unit ${ }^{\mathcal{N}}: X \rightarrow \mathcal{N}(X)$, following Proposition 1.4.3.


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

1.7.13 Prove that the following diagram commutes, where swap is the obvious map, comparable to transposition of a matrix.


Hint: Use, from the previous exercise, that accumulation is a homorphism of monoids.
1.7.14 Let $n \geq 1$ and $r \in(0,1)$. Show that:

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

1.7.15 Elaborate the details of the following two (alternative) proofs of the equation in Theorem 1.7.4 (3).
1 Use the derivate $\frac{d}{d r}$ 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.

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

1.7.16 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 $f\left(x_{1}, \ldots, x_{n}\right)$ and a natural multiset $\varphi \in \mathcal{N}(\{1, \ldots, n\})$ write:

$$
\partial^{\varphi} f:=\left(\partial x_{1}\right)^{\varphi(1)} \cdots\left(\partial x_{n}\right)^{\varphi(n)} f .
$$

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

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

### 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 $\psi, \varphi$ 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:

$$
\begin{align*}
\binom{\psi}{\varphi} & :=\frac{\psi \rrbracket}{\varphi \rrbracket \cdot(\psi-\varphi) \rrbracket} \\
& =\frac{\prod_{x} \psi(x)!}{\left(\prod_{x} \varphi(x)!\right) \cdot\left(\prod_{x}(\psi(x)-\varphi(x))!\right)}=\prod_{x \in \operatorname{supp}(\psi)}\binom{\psi(x)}{\varphi(x)} . \tag{1.42}
\end{align*}
$$

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

$$
\begin{equation*}
\binom{n}{\varphi}:=\frac{n!}{\varphi!\cdot(n-\|\varphi\|)!} . \tag{1.43}
\end{equation*}
$$

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 a generalisation of Vandermonde's formula. We reformulate it for numbers $B, G \in \mathbb{N}$ and $K \leq B+G$. It says:

$$
\begin{equation*}
\binom{B+G}{K}=\sum_{b \leq B, g \leq G, b+g=K}\binom{B}{b} \cdot\binom{G}{g} . \tag{1.44}
\end{equation*}
$$

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:

$$
\sum_{\varphi \leq K \psi}\binom{\psi}{\varphi}=\binom{L}{K} \quad \text { so that } \quad \sum_{\varphi \leq K \psi} \frac{\binom{\psi}{\varphi}}{\binom{L}{K}}=1 \text {. }
$$

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 3.4 later on.

Proof. We use induction on the number of elements in $\operatorname{supp}(\psi)$. We go through some initial values explicitly. If the number of elements in $\operatorname{supp}(\psi)$ 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 $\operatorname{supp}(\psi)$ 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 $\operatorname{supp}(\psi)=\{x, y\}$, corresponds the ordinary form of Vandermonde's formula (1.44). We briefly show how this equation (1.44) 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 $K>0$.

$$
\begin{aligned}
& \sum_{b \leq B, g \leq G+1, b+g=K}\binom{B}{b} \cdot\binom{G+1}{g} \\
& =\binom{B}{K} \cdot\binom{G+1}{0}+\binom{B}{K-1} \cdot\binom{G+1}{1}+\cdots+\binom{B}{K-G} \cdot\binom{G+1}{G}+\binom{B}{K-G-1} \cdot\binom{G+1}{G+1} \\
& \stackrel{1.1]}{=}\binom{B}{K} \cdot\binom{G}{0}+\binom{B}{K-1} \cdot\binom{G}{1}+\binom{B}{K-1} \cdot\binom{G}{0} \\
& +\cdots+\binom{B}{K-G} \cdot\binom{G}{G}+\binom{B}{K-G} \cdot\binom{G}{G-1}+\binom{B}{K-G-1} \cdot\binom{G}{G} \\
& =\sum_{b \leq B, g \leq G, b+g=K}\binom{B}{b} \cdot\binom{G}{g}+\sum_{b \leq B, g \leq G, b+g=K-1}\binom{B}{b} \cdot\binom{G}{g} \\
& \stackrel{(\mathrm{IH})}{=}\binom{B+G}{K}+\binom{B+G}{K-1} \\
& \stackrel{[1.1]}{-}\binom{B+G+1}{K} \text {. }
\end{aligned}
$$

For the induction step, let $\operatorname{supp}(\psi)=\left\{x_{1}, \ldots, x_{n}, y\right\}$, for $n \geq 2$. Writing $\ell=\psi(y), L^{\prime}=L-\ell$ and $\psi^{\prime}=\psi-\ell|y\rangle \in \mathcal{N}\left[L^{\prime}\right]\left(\left\{x_{1}, \ldots, x_{n}\right\}\right)$ gives:

$$
\begin{aligned}
& \sum_{\varphi \leq K}\binom{\psi}{\varphi}=\sum_{\varphi \leq K} \prod_{x}\binom{\psi(x)}{\varphi(x)}=\sum_{n \leq \ell} \sum_{\varphi \leq K-n} \psi^{\prime} \\
&\binom{\ell}{n} \cdot \prod_{i}\binom{\psi^{\prime}\left(x_{i}\right)}{\varphi\left(x_{i}\right)} \\
& \stackrel{(\mathrm{IH})}{=} \sum_{n \leq \ell, K-n \leq L-\ell}\binom{\ell}{n} \cdot\binom{L-\ell}{K-n} \stackrel{\text { II.44] }}{=}\binom{L}{K} .
\end{aligned}
$$

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 $K \leq L=\|\psi\|$.

1 For $\varphi \leq_{K} \psi$ one has:

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

2 Now:

$$
\mathbf{(} \psi)=\sum_{\varphi \leq \kappa \psi} \mathbf{( \varphi )} \cdot(\psi-\varphi) .
$$

The earlier equation 1.38 is a special case, for $K=1$.

Proof. 1 Because:

$$
\begin{aligned}
\frac{(\varphi) \cdot(\psi-\varphi)}{\mathbf{( \psi )}} & =\frac{K!}{\varphi!} \cdot \frac{(L-K)!}{(\psi-\varphi) \rrbracket} \cdot \frac{\psi \rrbracket}{L!} \\
& =\frac{K!\cdot(L-K)!}{L!} \cdot \frac{\psi \rrbracket}{\varphi \rrbracket \cdot(\psi-\varphi) \rrbracket}=\frac{\binom{\psi}{\varphi}}{\binom{L}{K}}
\end{aligned}
$$

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

$$
\left.\left.\sum_{\varphi \leq \kappa \psi}(\varphi) \cdot(\psi-\varphi)=(\psi) \cdot \frac{\sum_{\varphi \leq \kappa \psi}\binom{\psi}{\varphi}}{\binom{L}{K}}=\mathbf{(} \psi\right) \cdot \frac{\binom{L}{K}}{\binom{L}{K}}=\mathbf{(} \psi\right) .
$$

### 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. [159]. 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 $\psi, \varphi$ be natural multisets over the same set $X$, where $\psi$ is non-zero and $\operatorname{supp}(\varphi) \subseteq \operatorname{supp}(\psi)$.

$$
\begin{equation*}
\left(\binom{\psi}{\varphi}\right):=\prod_{x \in \operatorname{supp}(\psi)}\left(\binom{\psi(x)}{\varphi(x)}\right)=\prod_{x \in \operatorname{supp}(\psi)}\binom{\psi(x)+\varphi(x)-1}{\varphi(x)} . \tag{1.45}
\end{equation*}
$$

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:

$$
\begin{array}{lll}
3|a\rangle, & 3|b\rangle, & 3|c\rangle, \\
2|b\rangle+1|c\rangle, & 1|a\rangle+2|c\rangle, & 1|b\rangle+2|a\rangle+1|c\rangle,
\end{array} \quad 1|a\rangle+2|b\rangle, \quad 1|a\rangle+1|b\rangle+1|c\rangle .
$$

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 (1.44).

Lemma 1.8.5. Fix numbers $B \geq 1$ and $G \geq 1$. For all $K$ on has:

$$
\begin{equation*}
\left(\binom{B+G}{K}\right)=\sum_{0 \leq k \leq K}\left(\binom{B}{k}\right) \cdot\left(\binom{G}{K-k}\right) . \tag{1.46}
\end{equation*}
$$

In particular:

$$
\begin{equation*}
\binom{B+K}{K}=\left(\binom{B+1}{K}\right)=\sum_{0 \leq k \leq K}\left(\binom{B}{k}\right)=\sum_{0 \leq k \leq K}\left(\binom{B}{K-k}\right) . \tag{1.47}
\end{equation*}
$$

Proof. The second equation (1.47) 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 1.8.11) below.

$$
\begin{equation*}
\left(\binom{n}{K+1}\right)+\left(\binom{n+1}{K}\right)=\left(\binom{n+1}{K+1}\right) . \tag{*}
\end{equation*}
$$

We shall prove the first equation 1.46 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 (1.46) for $B=1$, by induction on $K$.
- When $K=0$ both sides in 1.46 are equal to 1 .
- Assume Equation (1.46) holds for $K$ (and $B=1$ ).

$$
\begin{aligned}
\sum_{0 \leq k \leq K+1}\left(\binom{1}{k}\right) \cdot\left(\binom{G}{(K+1)-k}\right)=\sum_{0 \leq k \leq K+1}\left(\binom{G}{K-(k-1)}\right) & =\left(\binom{G}{K+1}\right)+\sum_{0 \leq \ell \leq K}\left(\binom{G}{K-\ell}\right) \\
& \stackrel{(\mathrm{IH})}{=}\left(\binom{G}{K+1}\right)+\left(\binom{G+1}{K}\right) \\
& \stackrel{(*)}{=}\left(\binom{G+1}{K+1}\right) .
\end{aligned}
$$

- Now assume Equation (1.46) 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 1.46 are equal to 1 .
- Now assume that Equation (1.46) holds for $K$, and for $B$. Then:

$$
\begin{aligned}
& \sum_{0 \leq k \leq K+1}\left(\binom{B+1}{k}\right) \cdot\left(\binom{G}{(K+1)-k}\right) \\
& =\left(\binom{G}{K+1}\right)+\sum_{0 \leq k \leq K}\left(\binom{B+1}{k+1}\right) \cdot\left(\binom{G}{K-k}\right) \\
& \stackrel{(*)}{=}\left(\binom{G}{K+1}\right)+\sum_{0 \leq k \leq K}\left[\left(\binom{B}{k+1}\right)+\left(\binom{B+1}{k}\right)\right] \cdot\left(\binom{G}{K-k}\right) \\
& =\left(\binom{G}{K+1}\right)+\sum_{0 \leq k \leq K}\left(\binom{B}{k+1}\right) \cdot\left(\binom{G}{K-k}\right)+\sum_{0 \leq k \leq K}\left(\binom{B+1}{k}\right) \cdot\left(\binom{G}{K-k}\right) \\
& \stackrel{(\mathrm{IH}, \mathrm{~K})}{=} \sum_{0 \leq k \leq K+1}\left(\binom{B}{k}\right) \cdot\left(\binom{G}{(K+1)-k}\right)+\left(\binom{(B+1)+G}{K}\right) \\
& \stackrel{(\mathrm{IH}, \mathrm{~B})}{=}\left(\binom{B+G}{K+1}\right)+\left(\left(\binom{(B+1)+G}{K}\right)\right. \\
& \stackrel{(*)}{=}\left(\binom{(B+1)+G}{K+1}\right) .
\end{aligned}
$$

We now get the double-bracket analogue of Lemma 1.8.2

Proposition 1.8.6. Let $\psi$ be a non-empty natural multiset. Write $X=\operatorname{supp}(\psi)$ and $L=\|\psi\|$. Then, for each $K \in \mathbb{N}$,

$$
\sum_{\varphi \in \mathcal{N}[K](X)}\left(\binom{\psi}{\varphi}\right)=\left(\binom{L}{K}\right) \quad \text { so } \quad \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\left(\binom{\psi}{\varphi}\right)}{\left(\binom{L}{K}\right)}=1 \text {. }
$$

The fractions in this equation will show up later in so-called Pólya distributions, see Subsection 2.6.2 and Section 3.5. These fractions capture the probability of drawing a multiset $\varphi$ from an urn $\psi$ 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 $\psi$, 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 $\operatorname{supp}(\psi)=X \cup\{y\}$ where $y \notin X$ and $X$ is not empty. Write:

$$
L=\|\psi\| \quad \ell=\psi(y)>0 \quad \psi^{\prime}=\psi-\ell|y\rangle \quad L^{\prime}=L-\ell>0 .
$$

By construction $X=\operatorname{supp}\left(\psi^{\prime}\right)$ and $L^{\prime}=\left\|\psi^{\prime}\right\|$. Now:

$$
\begin{aligned}
& \sum_{\varphi \in \mathcal{N}[K](X \cup\{y])}\left(\binom{\psi}{\varphi}\right) \stackrel{[1.45}{=} \\
& \sum_{\varphi \in \mathcal{N}[K](X \cup\{y\})} \prod_{x \in X \cup\{y\}}\left(\binom{\psi(x)}{\varphi(x)}\right) \\
&=\sum_{0 \leq k \leq K} \sum_{\varphi \in \mathcal{N}[K-k][X)}\left(\binom{\psi(y)}{k}\right) \cdot \prod_{x \in X}\left(\binom{\psi(x)}{\varphi(x)}\right) \\
&=\sum_{0 \leq k \leq K}\left(\binom{\ell}{k}\right) \cdot \sum_{\varphi \in \mathcal{N}[K-k](X)}\left(\binom{\psi}{\varphi}\right) \\
& \stackrel{(\mathrm{IH})}{=} \sum_{0 \leq k \leq K}\left(\binom{\ell}{k}\right) \cdot\left(\binom{L^{\prime}}{K-k}\right) \\
& \stackrel{I .46}{=}\left(\binom{\ell+L^{\prime}}{K}\right)=\left(\binom{L}{K}\right) .
\end{aligned}
$$

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 $n \geq 1$ elements. The number of natural multisets of size $K$ over $X$ is $\left.\binom{n}{K}\right)$, that is:

$$
|\mathcal{N}[K](X)|=\left(\binom{n}{K}\right)=\binom{n+K-1}{K}
$$

Proof. The statement holds for $K=0$ since there is precisely $\left.1=\binom{n-1}{0}=\binom{n}{0}\right)$ multiset set of 0 , namely the empty multiset $\mathbf{0}$. Hence we may assume $K \geq 1$, so that Lemma 1.8.5 can be used.
We proceed by induction on $n \geq 1$. 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 $X=\left\{x_{1}, \ldots, x_{n}\right\}$ 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 $\varphi\left(x_{1}\right), \ldots, \varphi\left(x_{n}\right)$ is the number of multisets in $\mathcal{N}[K-k](X)$. Thus:

$$
\begin{aligned}
|\mathcal{N}[K](X \cup\{y\})| & =\sum_{0 \leq k \leq K}|\mathcal{N}[K-k](X)| \\
& \stackrel{(\mathrm{IH})}{=} \sum_{0 \leq k \leq K}\left(\binom{n}{K-k}\right) \\
& =\left(\binom{n+1}{K}\right), \quad \text { by Lemma } 1.8 .5
\end{aligned}
$$

There is also a visual proof of this result, described in terms of stars and bars,
see e.g. [52] 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 $\varphi$ over $X$ with size $K$ ?

Lemma 1.8.8. For an element $a \in X$, where $X$ has $n \geq 1$ 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:

$$
\mid\{\varphi \in \mathcal{N}[K](X) \mid \varphi \text { has full support }\} \left\lvert\,=\left(\binom{n}{K-n}\right)\right.
$$

1.8.3 Convince yourself that the following composite

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

is the $L$-fold sum of multisets.
1.8.4 In analogy with the powerset operator, with type $\mathcal{P}: \mathcal{P}(X) \rightarrow \mathcal{P}(\mathcal{P}(X))$, a powerbag operator PB: $\mathcal{N}(X) \rightarrow \mathcal{N}(\mathcal{N}(X))$ is introduced in [124] (see also [39]). It can be defined as:

$$
\mathrm{PB}(\psi):=\sum_{\varphi \leq \psi}\binom{\psi}{\varphi}|\varphi\rangle .
$$

1 Take $X=\{a, b\}$ and show that:

$$
\begin{aligned}
& \mathrm{PB}(1|a\rangle+3|b\rangle) \\
& \begin{array}{l}
=1|\mathbf{0}\rangle+1|1| a\rangle\rangle+3|1| b\rangle\rangle+3|1| a\rangle+1|b\rangle\rangle+3|2| b\rangle\rangle \\
\quad+3|1| a\rangle+2|b\rangle\rangle+1|3| b\rangle\rangle+1|1| a\rangle+3|b\rangle\rangle .
\end{array}
\end{aligned}
$$

2 Check that one can compute the powerbag of $\psi$ as follows. Take a list of elements that accumulate to $\psi$, such as $[a, b, b, b]$ in the previous item. Take the accumulation of all subsequences.
1.8.5 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:

$$
\binom{\psi}{\varphi_{1}, \ldots, \varphi_{N}}:=\frac{\psi \rrbracket}{\varphi_{1} \rrbracket \cdot \ldots \cdot \varphi_{N} \rrbracket} .
$$

1 Check that for $N \geq 3$, in analogy with Exercise 1.2 .3 ,

$$
\binom{\psi}{\varphi_{1}, \ldots, \varphi_{N}}=\binom{\psi}{\varphi_{1}} \cdot\binom{\psi-\varphi_{1}}{\varphi_{2}, \ldots, \varphi_{N}}
$$

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:

$$
\sum_{\substack{\varphi_{1} \leq K_{1}, \not, \ldots, \varphi_{N} \leq K_{N} \psi \\ \sum_{i} \varphi_{i}=\psi}}\binom{\psi}{\varphi_{1}, \ldots, \varphi_{N}}=\binom{K}{K_{1}, \ldots, K_{N}} .
$$

1.8.6 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 $\varphi, \psi$ have disjoint supports, that is, $\operatorname{supp}(\varphi) \cap$ $\operatorname{supp}(\psi)=\emptyset$. Show that now:

$$
(\varphi+\psi)=\binom{K+L}{K} \cdot(\varphi) \cdot(\psi)
$$

1.8.7 Let $\varphi, \psi$ be natural multiset on the same finite set $X$, where $\psi \geq \mathbf{1}$.

1 Show that one has:

$$
\left(\binom{\psi}{\varphi}\right)=\binom{\psi+\varphi-\mathbf{1}}{\varphi}=\frac{(\psi+\varphi-\mathbf{1}) \rrbracket}{\varphi!\cdot(\psi-\mathbf{1}) \rrbracket}
$$

2 Conclude, analogously to Lemma 1.8.3 1, that:

$$
\frac{\mathbf{( \varphi )} \cdot(\psi-\mathbf{1})}{\mathbf{( \psi + \varphi - \mathbf { 1 } )}}=\frac{\left(\begin{array}{l}
\binom{\psi}{\varphi}
\end{array}\right)}{\binom{L}{K}} .
$$

1.8.8 Let $X$ be a non-empty finite set, say with $N=|X|>0$ elements.

1 Let $\psi \in \mathcal{N}[L](X)$ be a fixed multiset of size $L$. Use Proposition 1.8.6 to prove, for arbitrary $K \in \mathbb{N}$,

$$
\sum_{\varphi \in \mathcal{N}[K](X)}\binom{\psi+\varphi}{\varphi}=\left(\binom{L+N}{K}\right) .
$$

2 Now let $L \geq K$ and let $\varphi \in \mathcal{N}[K](X)$ be given. Use the previous point to show that:

$$
\sum_{v \in \mathcal{N}[L](X), \varphi \leq K^{v}}\binom{v}{\varphi}=\binom{L+N-1}{K+N-1}
$$

1.8.9 Let $N \geq 0$ and $M \geq m \geq 1$ and be given. Use the multichoose Vandermonde Equation (1.47) to prove:

$$
\sum_{0 \leq i \leq N}\left(\binom{m}{i}\right) \cdot\binom{M-m+N-i}{N-i}=\binom{N+M}{N}
$$

1.8.10 $\quad$ Let $n \geq 1$ and $m \geq 0$.

1 Show that:

$$
\left(\binom{n+1}{m+1}\right)=\left(\binom{n+1}{m}\right)+\left(\binom{n}{m+1}\right) .
$$

2 Generalise this to:

$$
\left(\binom{n+k}{m+k}\right)=\sum_{0 \leq i \leq k}\binom{k}{i} \cdot\left(\binom{n+i}{m+k-i}\right) .
$$

1.8.11 Prove the following properties.
$1 \quad\left(\binom{n-k}{m-(k+1)}\right)=\left(\binom{m-k}{n-(k+1)}\right)$
$2\left(\binom{n}{m}\right)+\left(\binom{m}{n}\right)=\binom{n+m}{n}$
$3 m \cdot\left(\binom{n}{m}\right)=n \cdot\left(\binom{n+1}{m-1}\right)$.
$4 \quad n \cdot\left(\binom{n+1}{m}\right)=(n+m) \cdot\left(\binom{n}{m}\right)=(m+1) \cdot\left(\binom{n}{m+1}\right)$.
1.8.12 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 $\varphi, \psi$ with $x \in \operatorname{supp}(\psi)$ 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 $x \in \operatorname{supp}(\varphi) \subseteq \operatorname{supp}(\psi)$,

$$
\varphi(x) \cdot\left(\binom{\psi}{\varphi}\right)=\psi(x) \cdot\left(\binom{\psi+1|x\rangle}{\varphi-1|x\rangle}\right) .
$$

$4 \operatorname{For} \operatorname{supp}(\varphi) \subseteq \operatorname{supp}(\psi)$ and $x \in \operatorname{supp}(\psi)$,

$$
\psi(x) \cdot\left(\binom{\psi+1|x\rangle}{\varphi}\right)=(\varphi(x)+\psi(x)) \cdot\left(\binom{\psi}{\varphi}\right)=(\varphi(x)+1) \cdot\left(\binom{\psi}{\varphi+1|x\rangle}\right) .
$$

1.8.13 Let $n \geq 1$ and $m \geq 1$.

1 Show that:

$$
\sum_{j<m}\left(\binom{n}{j}\right)=\left(\binom{m}{n}\right) .
$$

2 Deduce that for $n \geq 1$ and $k \geq 0$,

$$
\sum_{j \leq k}\left(\binom{n}{j}\right)=\binom{n+k}{n}
$$

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 1 Extend Exercise 1.2.2 (1) to: for $k \geq 1$,

$$
\sum_{0 \leq j \leq m}\binom{k+n+j}{n}=\left(\binom{k+m+1}{n+1}\right)-\left(\binom{k}{n+1}\right) .
$$

Hint: One can use induction on $m$.
2 Show that one also has:

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

Hint: Use the multichoose version (1.46) of Vandermonde's formula.
1.8.15 Check that Theorem 1.7.4 (1) can be reformulated as: for a real number $x \in(0,1)$ and $K \geq 1$,

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

1.8.16 Let $s \in[0,1]$ and $n, m \geq 1$.

1 Prove first the following auxiliary result.

$$
\sum_{j<m}\left(\binom{n+1}{j}\right) \cdot s^{j}=\frac{1}{(1-s)} \cdot\left(\sum_{j<m}\left(\binom{n}{j}\right) \cdot s^{j}-\left(\binom{n+1}{m-1}\right) \cdot s^{m}\right) .
$$

2 Take $r=1-s$ so that $r+s=1$ and prove:

$$
r^{n} \cdot \sum_{j<m}\left(\binom{n}{j}\right) \cdot s^{j}+s^{m} \cdot \sum_{i<n}\left(\binom{m}{i}\right) \cdot r^{i}=1 .
$$

3 Show also that:

$$
\sum_{i \geq 0}\left(\binom{n}{m+i}\right) \cdot s^{i}+\left(\binom{m}{n+i}\right) \cdot r^{i}=\frac{1}{r^{n} \cdot s^{m}}
$$

### 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 $M P(K)$. The following triangular prims then gives an overview.


We have already seen the triangle on the left — but not for a fixed size $K$ namely in Diagram 1.34. The front rectangle appeared in Diagram 1.22. The parts of the prism involving the set $M P(K)$ of 'multiset partitions' is new and will be discussed below.

This set $M P(K)$ 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 [84, 92, 83] for further information.

Definition 1.9.1. For a positive number $K \in \mathbb{N}_{>0}$ we write:

$$
\operatorname{MP}(K):=\left\{\alpha \in \mathcal{N}\left(\mathbb{N}_{>0}\right) \mid \operatorname{sum}(\alpha)=K\right\},
$$

where $\operatorname{sum}(\alpha):=\sum_{n} \alpha(n) \cdot n$.
For $\alpha \in \operatorname{MP}(K)$ we write $\operatorname{size}(\alpha)=\|\alpha\|=\sum_{n} \alpha(n)$ for its size, as ordinary
multiset. Since $\operatorname{size}(\alpha) \leq \operatorname{sum}(\alpha)$, this size forms a function $M P(K) \rightarrow(K]$, as in the prism 1.48).

It is not hard to see that the support of a multiset partition $\alpha \in M P(K)$ is contained in $(K]=\{1, \ldots, K\}$. For instance, the set $M P(4)$ contains the five multiset partitions:

$$
4|1\rangle \quad 2|1\rangle+1|2\rangle \quad 1|1\rangle+1|3\rangle \quad 2|2\rangle \quad 1|4\rangle .
$$

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 1.11. 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 $M P(K)$ of multiset partitions, for $K=1,2, \ldots$, are given by the partition numbers:

$$
\begin{array}{llll}
|M P(1)|=1 & |M P(2)|=2 & |M P(3)|=3 & |M P(4)|=5 \\
|M P(5)|=7 & |M P(6)|=11 & |M P(7)|=15 & |M P(8)|=22 \ldots
\end{array}
$$

No closed formula is known for these partition numbers $1,2,3,5,7,11,15$, $22,30,42,56,77,101,135,176,231,297, \ldots$ see [6].

We turn to the maps labeled mc and sc in 1.48 .
Definition 1.9.2. Fix a number $K \geq 1$.
1 For a set $X$ we define the multiplicity count function $\operatorname{mc}[K]: \mathcal{N}[K](X) \rightarrow$ $M P(K)$ on a multiset $\varphi \in \mathcal{N}[K](X)$ as:
$\operatorname{mc}[K]\left(\sum_{i} n_{i}\left|x_{i}\right\rangle\right):=\sum_{i} 1\left|n_{i}\right\rangle \quad$ that is $\quad \operatorname{mc}[K](\varphi):=\sum_{x \in \operatorname{supp}(\varphi)} 1|\varphi(x)\rangle$.
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 $s c[K]: S P(K) \rightarrow M P(K)$ is defined on a set partition $P \in S P(K)$ as:

$$
s c[K](P):=\sum_{B \in P} 1| | B| \rangle .
$$

It thus counts the sizes of blocks $B$ in the set partition $P$.
The parameter $K$ in multiplicity / size count expressions $m c[K]$ and $s c[K]$ is dropped when it is clear from the context.

For instance, for $X=\{a, b, c\}$ and $K=10$,

$$
\begin{aligned}
& m c(2|a\rangle+6|b\rangle+2|c\rangle)=1|2\rangle+1|6\rangle+1|2\rangle=2|2\rangle+1|6\rangle \\
& m c(2|a\rangle+3|b\rangle+5|c\rangle)=1|2\rangle+1|3\rangle+1|5\rangle .
\end{aligned}
$$

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:

$$
\begin{aligned}
& s c(\{\{1,2\},\{3,4,5,6,7,8\},\{9,10\}\})=1|2\rangle+1|6\rangle+1|2\rangle=2|2\rangle+1|6\rangle \\
& s c(\{\{1,10\},\{2,5,9\},\{3,4,6,7,8\})=1|2\rangle+1|3\rangle+1|5\rangle .
\end{aligned}
$$

We collect some basic facts about size and multiplicity count.

## Lemma 1.9.3.

1 The back-rectangle in (1.48) commutes: size $\circ \mathrm{mc}=$ size $\circ$ supp;
2 The right-triangle in (1.48) commutes: size $\circ \mathrm{sc}=$ size;
3 When $|X| \geq K$, the function $\mathrm{mc}: \mathcal{N}[K](X) \rightarrow M P(K)$ 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 permutation $\pi$ : $X \xlongequal{\leftrightharpoons}$ $X$,

$$
\operatorname{mc}(\mathcal{N}(\pi)(\varphi))=\operatorname{mc}(\varphi)
$$

5 For an arbitrary element $x \in X$,

$$
m c(\varphi+1|x\rangle)=m c(\varphi)+1|\varphi(x)+1\rangle-1|\varphi(x)\rangle .
$$

6 Similarly, for $x \in \operatorname{supp}(\varphi)$,

$$
\operatorname{mc}(\varphi-1|x\rangle)= \begin{cases}\operatorname{mc}(\varphi)-1|1\rangle & \text { if } \varphi(x)=1 \\ \operatorname{mc}(\varphi)+1|\varphi(x)-1\rangle-1|\varphi(x)\rangle & \text { otherwise }\end{cases}
$$

$7 \varphi \prod_{0}=\prod_{1 \leq n \leq K}(n!)^{m c(\varphi)(n)}$.
Proof. 1 For $\varphi \in \mathcal{N}[K](X)$,

$$
\begin{aligned}
(\operatorname{size} \circ m c)(\varphi)=\operatorname{size}\left(\sum_{x \in \operatorname{supp}(\varphi)} 1\left|\varphi\left(x_{i}\right)\right\rangle\right) & =\sum_{x \in \operatorname{supp}(\varphi)} 1 \\
& =|\operatorname{supp}(\varphi)| \\
& =(\operatorname{size} \circ \operatorname{supp})(\varphi) .
\end{aligned}
$$

2 For a set partition $P \in S P(K)$,

$$
(\operatorname{size} \circ s c)(P)=\operatorname{size}\left(\sum_{B \in P} 1| | B| \rangle\right)=\sum_{B \in P} 1=|P|=\operatorname{size}(P) .
$$

3 Let $\alpha=\sum_{1 \leq k \leq K} n_{k}|k\rangle \in M P(K)$ be given. We can construct a multiset $\varphi_{K} \in \mathcal{N}[K](X)$ with $\operatorname{mc}\left(\varphi_{K}\right)=\alpha$ 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 $\varphi_{m}=\varphi_{m-1}+m\left|x_{1}\right\rangle+\cdots m\left|x_{n_{m}}\right\rangle$, where $x_{1}, \ldots, x_{n_{m}}$ are freshly chosen elements from $X$, not occurring in $\varphi_{m-1}$. This guarantees that $\varphi_{m}$ has $n_{m}$ many elements occurring $m$ times. By construction $\operatorname{mc}\left(\varphi_{m}\right)=\sum_{1 \leq k \leq m} n_{k}|k\rangle$. In particular, $\operatorname{mc}\left(\varphi_{K}\right)=\sum_{1 \leq k \leq K} n_{k}|k\rangle=$ $\alpha$. This construction uses $\|\alpha\|$ many elements from $X$. Since $K=\operatorname{sum}(\alpha)=$ $\sum_{m} n_{m} \cdot m \geq \sum_{m} n_{m}=\|\alpha\|$, the construction always works for a set $X$ with at least $K$ elements.
4 Obvious, since permuting the elements of a multiset does not change the multiplicities that it has.
5 Let $k=\varphi(x)$. If we add another element $x$ to the multiset $\varphi$, then the number $m c(\varphi)(k)$ of elements in $\varphi$ occuring $k$ times is decreased by one, and the number $\operatorname{mc}(\varphi)(k+1)$ 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),

$$
\begin{aligned}
\prod_{1 \leq m \leq K+1}(k!)^{m c(\varphi+1|x\rangle)(m)} & =((K+1)!)^{m c(\varphi+1|x\rangle)(K+1)} \cdot \prod_{1 \leq m \leq K}(k!)^{m c(\varphi+1|x\rangle)(m)} \\
& = \begin{cases}(K+1)! & \text { if } \varphi=K|x\rangle \\
\prod_{1 \leq m \leq K}(k!)^{m c(\varphi)(m)} \cdot \frac{(\varphi(x)+1)!}{\varphi(x)!} & \text { otherwise }\end{cases} \\
& \stackrel{(\text { IH) }}{=} \begin{cases}(\varphi+1|x\rangle) \rrbracket & \text { if } \varphi=K|x\rangle \\
\varphi!\cdot(\varphi(x)+1) & \text { otherwise }\end{cases} \\
& =(\varphi+1|x\rangle)!.
\end{aligned}
$$

The last item in the above lemma gives rise to a separate 'open' factorial 』 for multiset partitions.

Definition 1.9.4. Let $\alpha \in M P(K)$ be a multiset partition with sum $K$.
1 The factorial $\alpha \rrbracket_{p} \in \mathbb{N}$ is defined as:

$$
\alpha \rrbracket_{p}:=\prod_{1 \leq n \leq K}(n!)^{\alpha(n)}
$$

2 Along the same lines we put:

$$
(\alpha)_{p}:=\frac{\operatorname{sum}(\alpha)!}{\alpha \rrbracket_{p}}=\frac{K!}{\alpha \rrbracket_{p}} .
$$

These definitions make sense in the light of the following result.
Lemma 1.9.5. For $\varphi \in \mathcal{N}[K](X)$,

$$
\begin{equation*}
\varphi \rrbracket=\operatorname{mc}(\varphi) \rrbracket_{\rho} \quad \text { and } \quad(\varphi)=(\operatorname{mc}(\varphi))_{p} . \tag{1.49}
\end{equation*}
$$

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 $\alpha \in M P(K)$ be a multiset partition with $\|\alpha\| \leq N$. Then:

$$
\left|m c[K]^{-1}(\alpha)\right|=\frac{N!}{\alpha \rrbracket \cdot(N-\|\alpha\|)!} \stackrel{\boxed{1.43}}{=}\binom{N}{\alpha} .
$$

Proof. Let $\alpha=\sum_{1 \leq i \leq K} n_{i}|i\rangle \in M P(K)$. If $\varphi \in \mathcal{N}[K](X)$ satisfies $\operatorname{mc}(\varphi)=\alpha$, then $\varphi$ must have $n_{i}$ elements of $X$ occurring $i$ times, for each $i$. The number of choices for these elements is:

$$
\begin{aligned}
& \binom{N}{n_{1}} \cdot\binom{N-n_{1}}{n_{2}} \cdot \ldots \cdot\binom{N-n_{1}-\cdots-n_{K-1}}{n_{K}} \\
& =\frac{N!\cdot\left(N-n_{1}\right)!\cdot \ldots \cdot\left(N-n_{1}-\cdots-n_{K-1}\right)!}{n_{1}!\cdot\left(N-n_{1}\right)!\cdot n_{2}!\cdot\left(N-n_{1}-n_{2}\right)!\cdot \ldots \cdot n_{K}!\cdot\left(N-n_{1}-\cdots-n_{K}\right)!} \\
& =\frac{N!}{n_{1}!\cdot n_{2}!\cdot \ldots \cdot n_{K}!\cdot\left(N-n_{1}-\cdots-n_{K}\right)!} \\
& =\frac{N!}{\alpha \rrbracket \cdot(N-\|\alpha\|)!} .
\end{aligned}
$$

We turn to size count, from set partitions to multiset partitions.
Lemma 1.9.7. For $\alpha \in M P(K)$,

$$
\left|s c[K]^{-1}(\alpha)\right|=\frac{(\alpha)_{p}}{\alpha \rrbracket}=\frac{K!}{\alpha \rrbracket_{\mathrm{o} p} \cdot \alpha \rrbracket} .
$$

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 $\alpha$. We switch to lists $\ell$ of length $K$ over $(K]$ with $\operatorname{supp}(\ell)=(K]$. There are $K$ ! many such lists. Given such a list $\ell$ there are $\prod_{i} n_{i}!=\alpha \rrbracket$ many ways to break up the list into sublists, with $n_{i}$-many subslists of length $i$. This gives $\frac{K!}{\alpha!}$ many set partitions.

We still have to divide by $\alpha \rrbracket_{p}$ 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 $P \in S P(K)$ with,

$$
\begin{aligned}
\operatorname{acc}(\varphi) & =\operatorname{sc}(P) \\
& =\alpha \in M P(K), \text { say. }
\end{aligned}
$$

Then there are $\alpha \rrbracket$ many sequences $\vec{x} \in X^{K}$ with $\operatorname{acc}(\vec{x})=\varphi$ and $\operatorname{mat}(\vec{x})=P$, that is:

$$
\mid\left\{\vec{x} \in X^{K} \mid \operatorname{acc}(\vec{x})=\varphi \text { and } \operatorname{mat}(\vec{x})=P\right\} \mid=\alpha \rrbracket .
$$

Proof. Let $\vec{x} \in X^{K}$ be a candidate sequence with $\operatorname{acc}(\vec{x})=\varphi$ and $\operatorname{mat}(\vec{x})=P$. The multiset $\varphi$ tells us that an element $y \in \operatorname{supp}(\varphi)$ 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 $y \in \operatorname{supp}(\varphi)$ 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 $\alpha=\operatorname{mc}(\varphi)=\operatorname{sc}(P)$, then there are $n$ elements occurring $i$-many times in $\varphi$. 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 $\prod_{i} \alpha(i)!=\alpha \rrbracket$ many sequences $\vec{x}$ with $\operatorname{acc}(\vec{x})=\varphi$ and $\operatorname{mat}(\vec{x})=P$.

## Exercises

1.9.1 1 Double-check that all subdiagrams of Diagram 1.48 commute.

2 Check in detail that starting from the list $[b, a, b, c]$ gives the following outcomes.


### 1.9.2 Fix $K \geq 1$.

1 Let $\ell \in X^{K}$ be a sequence. Show that

$$
\operatorname{supp}(\ell)=\operatorname{supp}(\operatorname{acc}(\ell)) \cong \operatorname{mat}(\ell)
$$

2 Let $P \in S P(K)$ be a set partition. Show that there is a sequence $\ell$ with $\operatorname{mat}(\ell)=P=\operatorname{supp}(\ell)$.
1.9.3 Let $\alpha \in M P(K)$ and $\beta \in M P(L)$ be multiset partitions.

1 Write $\alpha+\beta \in \mathcal{N}\left(\mathbb{N}_{>0}\right)$ for the pointwise sum of multisets, so that $(\alpha+\beta)(i)=\alpha(i)+\beta(i)$.
Check that this $\alpha+\beta$ is an element of $\operatorname{MP}(K+L)$.
2 Calculate: $(2|1\rangle+1|2\rangle+2|3\rangle)+(2|2\rangle+3|4\rangle)$.
3 Write $\alpha \star \beta \in \mathcal{N}\left(\mathbb{N}_{>0}\right)$ for $\sum_{i, j} \alpha(i) \cdot \beta(j)|i \cdot j\rangle$.
Show that $\alpha \star \beta$ is in $\operatorname{MP}(K \cdot L)$.
4 Calculate: $(2|1\rangle+1|2\rangle+2|3\rangle) \star(2|2\rangle+3|4\rangle)$.
5 Check that + and $\star$ on multiset partitions are commutative.
6 Prove that multiplication $\star$ distributes over sum + in:

$$
\left(\alpha_{1}+\alpha_{2}\right) \star \beta=\left(\alpha_{1} \star \beta\right)+\left(\alpha_{2} \star \beta\right) .
$$

1.9.4 Let natural multisets $\varphi, \psi \in \mathcal{N}(X)$ have disjoint supports: $\operatorname{supp}(\varphi) \cap$ $\operatorname{supp}(\psi)=\emptyset$. Show that then:

$$
m c(\varphi+\psi)=m c(\varphi)+m c(\psi)
$$

Give an example that demonstrates that the disjointness requirement is necessary.
1.9.5 In Exercises 1.2 .5 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:

$$
\begin{aligned}
{\left[\begin{array}{l}
K \\
n
\end{array}\right] } & =\sum_{\alpha \in M P(K),\|\alpha\|=n} \frac{K!}{\alpha \rrbracket \cdot \operatorname{prod}(\alpha)} \quad \text { where } \quad \operatorname{prod}(\alpha)=\prod_{i} i^{\alpha(i)} \\
& =\sum_{\alpha \in M P(K),\|\alpha\|=n} \prod_{1 \leq i \leq K} \frac{1}{\alpha(i)!\cdot i^{\alpha(i)-1}},
\end{aligned}
$$

Prove this equation.
1.9.6 Check that the binomial coefficient (1.43) satisfies the following formula.

$$
\sum_{\alpha \in M P(K),\|\alpha\| \leq n}(\alpha)_{p} \cdot\binom{n}{\alpha}=n^{K} \quad \text { for all } K, n \in \mathbb{N}_{>0}
$$

Hint: Combine Exercise 1.7.7 with Lemmas 1.9.5 and 1.9.7

### 1.10 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 $c: X \rightarrow$ $T(Y)$. Such a channel turns an element $x \in X$ into a certain collection $c(x)$ of elements of $Y$. An ordinary function $f: X \rightarrow Y$ 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 $X \rightarrow \mathcal{P}(Y)$ is a non-deterministic computation; for $T=\mathcal{M}$, a channel $X \rightarrow \mathcal{M}(Y)$ 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 $c: X \leadsto Y$, with a special arrow $\rightsquigarrow$ that carries a circle on its shaft.

Definition 1.10.1. 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 $c: X \rightarrow T(Y)$. It will be written as $c: X \leadsto Y$, 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 $c: X \rightarrow T(Y)$ we can form a new state $c \gg=\omega$ in $T(Y)$, on the codomain. It is defined as:

$$
c \gg=\omega:=(\text { flat } \circ T(c))(\omega) \quad \text { where } \quad T(X) \xrightarrow{T(c)} T(T(Y)) \xrightarrow{\text { flat }} T(Y) .
$$

This operation $\omega \mapsto c »=\omega$ is called state tranformation, sometimes with ad-
ditional clarification along the channel c. It may also be called push forward. In functional programming it is commonly called bind ${ }^{1}$.
3 Let $c: X \leadsto Y$ and $d: Y \leadsto Z$ be two channels. Then we can compose them and get a new channel $d \odot c: X \leadsto Z$ via:

$$
(d \odot c)(x):=d \gg=c(x) \quad \text { so that } \quad d \odot c=\text { flat } \circ T(d) \circ c .
$$

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 1.10.2. 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 $f: X \rightarrow \mathcal{L}(Y)$ can for instance be of the form:

$$
f(a)=[u, v] \quad f(b)=[u, u] \quad f(c)=[v, u, v] .
$$

State transformation $f \gg=\omega$ amounts to 'map list' with $f$ and then flattening. It turns a list of lists into a list, as in:

$$
\begin{aligned}
f \gg=\omega=\operatorname{flat}(\mathcal{L}(f)(\omega)) & =\operatorname{flat}([f(c), f(b), f(b), f(a)]) \\
& =\operatorname{flat}([[v, u, v],[u, u],[u, u],[u, v]]) \\
& =[v, u, v, u, u, u, u, u, v] .
\end{aligned}
$$

2 We consider the analogous example for $T=\mathcal{P}$. We thus take as state $\sigma=$ $\{a, b, c\}$ and as channel $g: X \rightarrow \mathcal{P}(Y)$, now given by subsets:

$$
g(a)=\{u, v\} \quad g(b)=\{u\} \quad g(c)=\{u, v\} .
$$

Then:

$$
\begin{aligned}
g \gg=\sigma=\operatorname{flat}(\mathcal{P}(f)(\sigma)) & =\bigcup\{g(a), g(b), g(c)\} \\
& =\bigcup\{\{u, v\},\{u\},\{u, v\}\} \\
& =\{u, v\} .
\end{aligned}
$$

[^2]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 $h: X \rightarrow \mathcal{M}(Y)$ could have:

$$
h(a)=10|u\rangle+5|v\rangle \quad h(b)=1|u\rangle \quad h(c)=4|u\rangle+1|v\rangle .
$$

We then get as state transformation:

$$
\begin{aligned}
h \gg \tau & =\operatorname{flat}(\mathcal{M}(h)(\tau)) \\
& =\operatorname{flat}(3|h(a)\rangle+2|h(b)\rangle+5|h(c)\rangle) \\
& =\operatorname{flat}(3|10| u\rangle+5|v\rangle\rangle+2|1| u\rangle\rangle+5|4| u\rangle+1|v\rangle\rangle) \\
& =30|u\rangle+15|v\rangle+2|u\rangle+20|u\rangle+5|v\rangle \\
& =52|u\rangle+20|v\rangle .
\end{aligned}
$$

We shall mostly be using multiset - and probabilistic channels, as special case - and so we explicitly describe state transformation $>=$ in these cases. So let $c: X \rightarrow \mathcal{M}(Y)$ be an $\mathcal{M}$-channel. Transformation of a state $\omega$ on $X$ can be described as:

$$
\begin{equation*}
(c \gg=\omega)(y)=\sum_{x \in X} \omega(x) \cdot c(x)(y) . \tag{1.50}
\end{equation*}
$$

Equivalently, we can describe the transformed state $c \gg \omega$ as a formal sum:

$$
\begin{equation*}
c \gg=\omega=\sum_{y \in Y}\left(\sum_{x \in X} c(x)(y) \cdot \omega(x)\right)|y\rangle . \tag{1.51}
\end{equation*}
$$

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 1.10 .1 and use the 'monad' properties of flatten and unit.

## Lemma 1.10.3.

1 Channel composition $\odot$ has an identity channel, namely unit: $Y \leadsto Y$, so that:

$$
\text { unit } \odot c=c \quad \text { and } \quad d \odot \text { unit }=d,
$$

for all channels $c: X \rightsquigarrow Y$ and $d: Y \rightsquigarrow Z$. Another way to write the second equation is: $d \gg=u n i t(y)=d(y)$.
2 Channel composition $\odot$ is associative:

$$
e \odot(d \odot c)=(e \odot d) \odot c,
$$

for all channels $c: X \mapsto Y, d: Y \mapsto Z$ and $e: Z \mapsto W$.

3 State tranformation via a composite channel is the same as two consecutive transformations:

$$
(d \odot c) \gg=\omega=d »=(c \gg=\omega) .
$$

4 Each ordinary function $f: Y \rightarrow Z$ gives rise to a 'trivial' or 'deterministic' channel $\langle f\rangle:=$ unit $\circ f: Y \leadsto Z$. This construction 〈-〉 satisfies:

$$
\langle f\rangle \gg=\omega=T(f)(\omega),
$$

where $T$ is the type of channel involved. Moreover:

$$
\langle g\rangle \odot\langle f\rangle=\langle g \circ f\rangle \quad\langle f\rangle \odot c=T(f) \circ c \quad d \odot\langle f\rangle=d \circ f,
$$

for all functions $g: Z \rightarrow W$ and channels $c: X \leadsto Y$ and $d: Y \rightsquigarrow W$.

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 1.10.1 3):

$$
\text { unit } \odot c=\text { flat } \circ T(\text { unit }) \circ c=\text { id } \circ c=c
$$

For the second equation we use naturality of unit in:

$$
d \odot \text { unit }=\text { flat } \circ T(d) \circ \text { unit }=\text { flat } \circ \text { unit } \circ d=i d \circ d=d .
$$

2 The proof of associativity uses naturality and also the commutation of flatten with itself (the 'flat-flat law'), expressed as flat $\circ$ flat $=$ flat $\circ T$ (flat).

$$
\begin{aligned}
e \odot(d \odot c) & =\text { flat } \circ T(e) \circ(d \odot c) & & \\
& =\text { flat } \circ T(e) \circ \text { flat } \circ T(d) \circ c & & \\
& =\text { flat } \circ \text { flat } \circ T(T(e)) \circ T(d) \circ c & & \text { by naturality of flat } \\
& =\text { flat } \circ T(\text { flat }) \circ T(T(e)) \circ T(d) \circ c & & \text { by the flat-flat law } \\
& =\text { flat } \circ T(f l a t \circ T(e) \circ d) \circ c & & \text { by functoriality of } T \\
& =\text { flat } \circ T(e \odot d) \circ c & & \\
& =(e \odot d) \odot c & &
\end{aligned}
$$

3 Along the same lines:

$$
\begin{aligned}
(d \odot c) \gg=\omega & =(\text { flat } \circ T(d \odot c))(\omega) & & \\
& =(\text { flat } \circ T(f f a t \circ T(d) \circ c))(\omega) & & \\
& =(\text { flat } \circ T(\text { flat }) \circ T(T(d)) \circ T(c))(\omega) & & \text { by functoriality of } T \\
& =(\text { flat } \circ \text { flat } \circ T(T(d)) \circ T(c))(\omega) & & \text { by the flat-flat law } \\
& =(\text { flat } \circ T(d) \circ \text { flat } \circ T(c))(\omega) & & \text { by naturality of flat } \\
& =(\text { flat } \circ T(d))((\text { flat } \circ T(c))(\omega)) & & \\
& =(\text { flat } \circ T(d))(c \gg \omega) & & \\
& =d \gg(c \gg=\omega) . & &
\end{aligned}
$$

4 All these properties follow from elementary facts that we have seen before:

$$
\begin{aligned}
\langle f\rangle \gg \omega & =(\text { flat } \circ T(\text { unit } \circ f)(\omega) & & \\
& =(\text { flat } \circ T(\text { unit }) \circ T(f))(\omega) & & \text { by functoriality of } T \\
& =T(f)(\omega) & & \text { by a flat-unit law } \\
\langle g\rangle \odot\langle f\rangle & =\text { flat } \circ T(\text { unit } \circ g) \circ \text { unit } \circ f & & \\
& =\text { flat } \circ \text { unit } \circ(\text { unit } \circ g) \circ f & & \text { by naturality of unit } \\
& =\text { unit } \circ g \circ f & & \text { by a flat-unit law } \\
& =\langle g \circ f\rangle & & \\
\langle f\rangle \circ c & =\text { flat } \circ T(\text { unit } \circ f) \circ c & & \\
& =\text { flat } \circ T(\text { unit }) \circ T(f) \circ c & & \text { by functoriality of } T \\
& =T(f) \circ c & & \text { by a flat-unit law } \\
d \odot\langle f\rangle & =\text { flat } \circ T(d) \circ \text { unit } \circ f & & \\
& =f l a t \circ \text { unit } \circ d \circ f & & \text { by naturality of unit } \\
& =d \circ f & & \text { by a flat-unit law. }
\end{aligned}
$$

In the sequel we often omit writing the brackets $\langle-\rangle$ that turn an ordinary function $f: X \rightarrow Y$ into a channel $\langle f\rangle$. For instance, in a state transformation $f \gg \omega$, it is clear that we use $f$ as a channel, so that the expression should be $\operatorname{read}$ as $\langle f\rangle »>\omega$.

## Exercises

1.10.1 For a function $f: X \rightarrow Y$ define an inverse image (or preimage) $\mathcal{P}_{-}$ channel $f^{-1}: Y \rightsquigarrow X$ by:

$$
f^{-1}(y):=\{x \in X \mid f(x)=y\} .
$$

Prove that:

$$
(g \circ f)^{-1}=f^{-1} \odot g^{-1} \quad \text { and } \quad \quad \text { id }^{-1}=\text { unit }
$$

1.10.2 Notice that a state of type $X$ can be identified with a channel $\mathbf{1} \rightarrow T(Y)$ with singleton set $\mathbf{1}=\{0\}$ as domain. Check that under this identification, state transformation $c \gg=\omega$ corresponds to channel composition $c \odot \omega$.
1.10.3 Let $f: X \mapsto Y$ be a channel.

1 Prove that if $f$ is a $\mathcal{P}_{\text {fin }}$-channel, then the state transformation function $f \gg(-): \mathcal{P}_{\text {fin }}(X) \rightarrow \mathcal{P}_{\text {fin }}(Y)$ can also be defined via freeness, namely as the unique function $\bar{f}$ in Proposition 1.5.4
2 Similarly, show that $f \gg=(-)=\bar{f}$ when $f$ is an $\mathcal{M}$-channel, as in Exercise 1.6 .13
1.10.4 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 7.9.1.)
2 Show that for finite sets $X, Y$ there is a similar correspondence for multiset channels.

### 1.11 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 [129]). Category theory focuses on the structural aspects of mathematics and shows that many mathematical constructions have the same underlying structure. It emphasises similarities between different areas (see e.g. [130]). Category theory has become very useful in (theoretical) computer science too, since it involves a clear distinction between specification and implementation, see books like [7, 11, 123, 149]. 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 [122]. 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 ??, and e.g. [167] and [74] 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. [64, 168]), probabilistic Bayesian reasoning [29, 54, 32, 98], and category theory [57]. Probabilistic methods have received wider attention, for instance, via the current interest in data analytics (see the essay [5]), in quantum probability [139, 28], and in cognition theory [70, 166].

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.

### 1.11.1 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 $f: X \rightarrow Y$, 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 examples.

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 $T \in\{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$ we can form the category $\operatorname{Chan}(T)$. Its objects are arbitrary sets $X$, but its morphisms $X$ to $Y$ are $T$-channels, $X \rightarrow T(Y)$, written as $X \mapsto Y$. We have already seen that channels are closed under composition $\odot$ and have unit as identity, see Lemma 1.10.3. We can now say that 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 $\operatorname{Chan}(\mathcal{P})$ we are in the world of non-deterministic computation, whereas $\operatorname{Chan}(\mathcal{M})$ 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 $\mathbb{C}^{\text {op }}$ can be formed. It has the same objects as $\mathbb{C}$, but its morphisms are reversed. Thus $f: Y \rightarrow X$ in $\mathbb{C}^{\text {op }}$ means $f: X \rightarrow Y$ 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 $(X, A) \rightarrow(Y, B)$ in $\mathbb{C} \times \mathbb{D}$ is given by a pair $(f, g)$ of arrows $f: X \rightarrow Y$ in $\mathbb{C}$ and $g: A \rightarrow B$ in $\mathbb{D}$.

### 1.11.2 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 $F: \mathbb{C} \rightarrow \mathbb{D}$ 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 $f: X \rightarrow Y$ in $\mathbb{C}$ to a morphism $F(f): F(X) \rightarrow F(Y)$ in $\mathbb{D}$. This mapping $F$ should preserve composition and identities, as in: $F(g \circ f)=$ $F(g) \circ F(f)$ and $F\left(i d_{X}\right)=i d_{F(X)}$.
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 $T \in\left\{\mathcal{L}, \mathcal{P}, \mathcal{P}_{\text {fin }}, \mathcal{P}_{*}, \mathcal{P}_{\mathcal{P}}[K], \mathcal{M}, \mathcal{M}_{*}, \mathcal{N}, \mathcal{N}_{*}, \mathcal{N}[K]\right\}$ is a functor $T:$ Sets $\rightarrow$ Sets. This has been described in the beginning of each of the sections $1.4-$ 1.6

2 Taking lists is also a functor $\mathcal{L}:$ Sets $\rightarrow$ Mon. This is in essence the content of Lemman 1.4.2 One can also view $\mathcal{P}, \mathcal{P}_{\text {fin }}$ and $\mathcal{M}$ as functors Sets $\rightarrow$ Mon, see Lemmas 1.5.1 and 1.6.3 Moreover, one can describe $\mathcal{P}, \mathcal{P}_{\text {fin }}$ as a functor Sets $\rightarrow$ PoSets, by considering each set of subsets $\mathcal{P}(X)$ and $\mathcal{P}_{\text {fin }}(X)$ with its subset relation $\subseteq$ as partial order. In order to verify this claim one has to check that $\mathcal{P}(f): \mathcal{P}(X) \rightarrow \mathcal{P}(Y)$ is a morphism of posets, that is, forms a monotone function. But that is easy.

3 There is also a functor $J$ : Sets $\rightarrow \boldsymbol{C h a n}(T)$, for each $T$. It is the identity on sets / objects: $J(X):=X$. But it sends a functon $f: X \rightarrow Y$ to the channel $J(f):=\langle f\rangle=$ unit $\circ f: X \mapsto Y$. We have seen, in Lemma 1.10.3 4, that $J(g \circ f)=J(g) \circ J(f)$ and that $J(i d)=i d$, where the latter identity id is unit in the category $\operatorname{Chan}(T)$. 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 $\times$ : Sets $\times$ Sets $\rightarrow$ Sets. Its action on morphisms was already described at the end of Subsection 1.3.1, see also Exercise 1.3.3

5 If we have functors $F_{i}: \mathbb{C}_{i} \rightarrow \mathbb{D}_{i}$, for $i=1,2$, then we also have a product functor $F_{1} \times F_{2}: \mathbb{C}_{1} \times \mathbb{C}_{2} \rightarrow \mathbb{D}_{1} \times \mathbb{D}_{2}$ between product categories, simply by $\left(F_{1} \times F_{2}\right)\left(X_{1}, X_{2}\right)=\left(F_{1}\left(X_{1}\right), F_{2}\left(X_{2}\right)\right)$, and similarly for morphisms.

### 1.11.3 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 $F, G: \mathbb{C} \rightarrow \mathbb{D}$, a natural transformation $\alpha$ from $F$ to $G$ is a collection of maps $\alpha_{X}: F(X) \rightarrow G(X)$ in $\mathbb{D}$, indexed by objects $X$ in $\mathbb{C}$. Naturality means that $\alpha$ works in the same way on all objects and is expressed as follows: for each morphism $f: X \rightarrow Y$ in $\mathbb{C}$, the rectangle

in $\mathbb{D}$ commutes.
Such a natural transformation is often denoted by a double arrow $\alpha: F \Rightarrow G$. We briefly review some of the examples of natural transformations that we have seen.

1 The support maps supp: $\mathcal{L} \Rightarrow \mathcal{P}_{\text {fin }}$, supp: $\mathcal{M} \Rightarrow \mathcal{P}_{\text {fin }}$ for lists and multisets are natural transformations, see the overview Diagram 1.34. Also accumulation forms a natural transformation acc: $\mathcal{L} \Rightarrow \mathcal{N}$.
2 For each $T \in\{\mathcal{L}, \mathcal{P}, \mathcal{M}\}$ we have described maps unit: $X \rightarrow T(X)$ and flat: $T(T(X)) \rightarrow T(X)$ and have seen naturality results about them. We can now state more precisely that they are natural transformations unit: id $\Rightarrow$ $T$ and flat: $(T \circ T) \Rightarrow T$. Here we have used id as the identity functor Sets $\rightarrow$ Sets, and $T \circ T$ as the composite of $T$ with itself, also as a functor Sets $\rightarrow$ Sets.

### 1.11.4 Monads

A monad on a category $\mathbb{C}$ is a functor $T: \mathbb{C} \rightarrow \mathbb{C}$ that comes with two natural transformations unit : id $\Rightarrow T$ and flat $:(T \circ T) \Rightarrow T$ satisfying:

$$
\begin{gather*}
\text { flat } \circ \text { unit }=\text { id }=\text { flat } \circ T(\text { unit })  \tag{1.52}\\
\text { flat } \circ \text { flat }=\text { flat } \circ T(\text { flat }) .
\end{gather*}
$$

All the collection functors $\mathcal{L}, \mathcal{P}, \mathcal{P}_{*}, \mathcal{P}_{\text {fin }}, \mathcal{M}, \mathcal{M}_{*}, \mathcal{N}, \mathcal{N}_{*}$ 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 Chan $(T)$ of $T$-channels, that capture computations of type $T$, see Subsection 1.11.1 In category theory this is called the Kleisli category of $T$. Composition in this category Chan $(T)$ is called Kleisli composition. In
this book it is written as $\odot$, where the context should make clear what the monad $T$ at hand is.

Monads have become popular in functional programming [136] as mechanisms for including special effects (e.g., for input-output, writing, side-effects, continuations) into a functional programming language ${ }^{2}$ The structure of probabilistic computation is also given by monads, namely by the discrete distribution monads $\mathcal{D}, \mathcal{D}_{\infty}$ and by the continuous distribution monad $\mathcal{G}$.

We thus associate the (Kleisli) category Chan $(T)$ of channels with a monad $T$. A second category is associated with a monad $T$, namely the category $\mathcal{E} \mathcal{M}(T)$ of "Eilenberg-Moore" algebras. The objects of $\mathcal{E} \mathcal{M}(T)$ are algebras $\alpha: T(X) \rightarrow X$, satisfying $\alpha \circ$ unit $=$ id and $\alpha \circ$ flat $=\alpha \circ T(\alpha)$. We have seen algebras for the monads $\mathcal{L}, \mathcal{P}_{\text {fin }}$, 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 $\mathcal{E} \mathcal{M}(T)$ 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 1.11.1. 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 $f: X \rightarrow Y$ we define $T(f): M \times X \rightarrow M \times Y$ by $T(f)(m, x)=(m, f(x))$. There is a unit map unit: $X \rightarrow M \times X$, namely unit $(x)=(0, x)$ and a flattening map flat: $M \times$ $(M \times X) \rightarrow M \times X$ by $\mu\left(m, m^{\prime}, x\right)=\left(m+m^{\prime}, x\right)$. We skip naturality and concentrate on the monad equations $\sqrt{1.52}$. First, for $(m, x) \in T(X)=M \times X$,

$$
\begin{aligned}
(f l a t \circ \operatorname{unit})(m, x) & =\operatorname{flat}(0, m, x)=(0+m, x)=(m, x) \\
(\text { flat } \circ T(\text { unit }))(m, x) & =\operatorname{flat}(m, \text { unit }(x))=\operatorname{flat}(m, 0, x)=(m, x)
\end{aligned}
$$

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 1.11.2. Let $T_{1}=\left(T_{1}\right.$, unit $_{1}$, flat $\left.{ }_{1}\right)$ and $T_{2}=\left(T_{2}\right.$, unit $_{2}$, flat $\left.{ }_{2}\right)$ be two monads (on Sets). A map/homomorphism of monads from $T_{1}$ to $T_{2}$ is a natural

[^3]transformation $\alpha: T_{1} \Rightarrow T_{2}$ that commutes with unit and flatten in the sense that the two diagrams

commute, for each set $X$.
The writer monads from Lemma 1.11 .1 give simple examples of maps of monads: if $f: M_{1} \rightarrow M_{2}$ is a map of monoids, then the maps $\alpha:=f \times i d: M_{1} \times$ $X \rightarrow M_{2} \times X$ form a map of monoids.

For a historical account of monads and their applications we refer to [71].

## Exercises

1.11.1 We have seen the functor $J$ : Sets $\rightarrow \operatorname{Chan}(T)$. Check that there is also a functor $\operatorname{Chan}(T) \rightarrow$ Sets in the opposite direction, which is $X \mapsto T(X)$ on objects, and $c \mapsto c \gg(-)$ on morphisms. Check explicitly that composition is preserved, and find the earlier result that stated that fact implicitly.
1.11.2 Recall from 1.25 the subset $\mathcal{N}[K](X) \subseteq \mathcal{M}(X)$ of natural multisets with $K$ elements. Prove that $\mathcal{N}[K]$ is a functor Sets $\rightarrow$ Sets.
1.11.3 Show that Exercise 1.10 .1 implicitly describes a functor Sets $^{\text {op }} \rightarrow$ $\operatorname{Chan}(\mathcal{P})$, which is the identity on objects.
1.11.4 Show that the zip function from Exercise 1.3.7 is natural: for each pair of functions $f: X \rightarrow U$ and $g: Y \rightarrow V$ the following diagram commutes.

1.11.5 Fill in the remaining details in the proof of Lemma 1.11 .1 that $T$ is a functor, that unit and flat are natural transformation, and that the flatten equation holds.
1.11.6 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 $\kappa_{1}: X \rightarrow X+A$ and $\kappa_{2}: A \rightarrow X+A$ 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 Sets $\rightarrow$ 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.
1.11.7 Check that the support and accumulation functions form maps of monads in the situations:
1 supp: $\mathcal{M}(X) \Rightarrow \mathcal{P}(X)$;
2 acc: $\mathcal{L}(X) \Rightarrow \mathcal{N}(X)$.
1.11.8 Let $T=(T$, unit, flat $)$ be a monad. By definition, it involves $T$ as a functor $T$ : Sets $\rightarrow$ Sets. Show that $T$ can be 'extended' to a functor $\bar{T}: \operatorname{Chan}(T) \rightarrow \boldsymbol{\operatorname { C h a n }}(T)$. It is defined on objects as $\bar{T}(X):=T(X)$ and on a morphism $f: X \leadsto Y$ as:

$$
\bar{T}(f):=(T(X) \xrightarrow{T(f)} T(T(Y)) \xrightarrow{\text { flat }} T(Y) \xrightarrow{\text { unit }} T(T(Y))) .
$$

Prove that $\bar{T}$ is a functor, i.e. that it preserves (channel) identities and composition.

## 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 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 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 that 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, tensors $\otimes$ of distributions are special since they may involve dependencies / correlations which create a dynamic that is essential to the field. This chapter introduces the basics of such 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 computations, 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. 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 wellbehaved. It occurs regularly and is thus included in this first chapter on probability.

One way to compare two distributions (on the same space) is via the socalled 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 distribution over a set $X$ is a finite formal convex sum of the form:

$$
r_{1}\left|x_{1}\right\rangle+\cdots+r_{n}\left|x_{n}\right\rangle \quad \text { where } \quad x_{i} \in X \text { and } r_{i} \in[0,1] \text { with } \sum_{i} r_{i}=1
$$

We can write such an expression as a (formal) sum $\sum_{i} r_{i}\left|x_{i}\right\rangle$. 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 $\omega: X \rightarrow[0,1]$ 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 $\omega \in \mathcal{M}(X) \subseteq \mathbb{R}_{\geq 0}^{X}$ 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 $\omega: X \rightarrow[0,1]$ with the unit 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 $\mathcal{D}_{f s}(X) \subseteq \mathcal{D}(X)$ for the subset of distributions $\omega \in \mathcal{D}(X)$ with full support; this means that $\operatorname{supp}(\omega)=X$, 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. [158], 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 : $X \rightarrow \mathcal{D}(X)$.
This unit map does not exist when one switches to distributions $\mathcal{D}_{f s}$ 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 4.5.9. Therefor 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
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.

$$
\frac{1}{2}|H\rangle+\frac{1}{2}|T\rangle \quad 0.51|H\rangle+0.49|T\rangle .
$$

In general, for a non-empty finite set $X$ there a uniform distribution $u n i f_{X} \in$ $\mathcal{D}(X)$ that assigns the same probability to each element. Thus, it is given by:

$$
\text { unif }_{X}:=\sum_{x \in X} \frac{1}{|X|}|x\rangle \quad \text { where }|X| \in \mathbb{N}_{>0} \text { is the size of } X .
$$

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 en 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:

$$
\begin{aligned}
& 0.082|a\rangle+0.015|b\rangle+0.028|c\rangle+0.043|d\rangle+0.13|e\rangle+0.022|f\rangle \\
& \quad+0.02|g\rangle+0.061|h\rangle+0.07|i\rangle+0.0015|j\rangle+0.0077|k\rangle \\
& \quad+0.04|l\rangle+0.024|m\rangle+0.067|n\rangle+0.075|o\rangle+0.019|p\rangle \\
& \quad+0.00095|q\rangle+0.06|r\rangle+0.063|s\rangle+0.091|t\rangle+0.028|u\rangle \\
& \quad+0.0098|v\rangle+0.024|w\rangle+0.0015|x\rangle+0.02|y\rangle+0.0074|z\rangle .
\end{aligned}
$$

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.

Below we describe several standard examples of distributions that will play an important role in the remainder of the book. We use above the formal convex sum notation. To start, we look at distributions that have a probability $r \in[0,1]$ as parameter.

## Example 2.1.2.

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:

$$
\operatorname{flip}(r):=r|1\rangle+(1-r)|0\rangle .
$$

It uses 1 for 'head' and 0 for 'tail'. We may thus see flip as a function flip: $[0,1] \rightarrow \mathcal{D}(\mathbf{2})$ from probabilities to distributions over the sample space $\mathbf{2}=\{0,1\}$ of Booleans. This flip $(r)$ is often called the Bernoulli distribution, with parameter $r \in[0,1]$.



Figure 2.1 Plots of a slightly biased coin distribution $0.51|H\rangle+0.49|T\rangle$ and a fair (uniform) dice distribution on $\{1,2,3,4,5,6\}$ in the top row, together with the distribution of letter frequencies in English at the bottom. We see that the letter e has the highest probability, as it occurs most frequently.

2 For each number $K \in \mathbb{N}$ and probability $r \in[0,1]$ there is the binomial distribution $b n[K](r) \in \mathcal{D}(\{0,1, \ldots, K\})$. It captures probabilities for iterated coin flips, and is given by the convex sum:

$$
\begin{equation*}
b n[K](r):=\sum_{0 \leq k \leq K}\binom{K}{k} \cdot r^{k} \cdot(1-r)^{K-k}|k\rangle . \tag{2.1}
\end{equation*}
$$

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 (1.39).

Here is an example, for $K=4$ and $r=\frac{1}{3}$.

$$
\begin{aligned}
\operatorname{bn}[4]\left(\frac{1}{3}\right)= & \binom{4}{0} \cdot\left(\frac{1}{3}\right)^{0} \cdot\left(\frac{2}{3}\right)^{4}|0\rangle+\binom{4}{1} \cdot\left(\frac{1}{3}\right)^{1} \cdot\left(\frac{2}{3}\right)^{3}|1\rangle+\binom{4}{2} \cdot\left(\frac{1}{3}\right)^{2} \cdot\left(\frac{2}{3}\right)^{2}|2\rangle \\
& \quad\binom{4}{3} \cdot\left(\frac{1}{3}\right)^{3} \cdot\left(\frac{2}{3}\right)^{1}|3\rangle+\binom{4}{4} \cdot\left(\frac{1}{3}\right)^{4} \cdot\left(\frac{2}{3}\right)^{0}|4\rangle \\
= & \frac{16}{81}|0\rangle+\frac{32}{81}|1\rangle+\frac{8}{27}|2\rangle+\frac{8}{81}|3\rangle+\frac{1}{81}|4\rangle .
\end{aligned}
$$



Figure 2.2 Plots of two binomial distributions, see Examples 2.1.2

We can organise binomial distributions as a function $b n[K]:[0,1] \rightarrow$ $\mathcal{D}(\{0,1, \ldots, K\})$. 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 $\{0,1, \ldots, 10\}$.

In Section 2.6 we shall put these binomial distributions in the more general perspective of 'draw' distributions, associated with drawing coloured balls from an urn.

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 $f: X \rightarrow Y$ we have $\mathcal{D}(f): \mathcal{D}(X) \rightarrow \mathcal{D}(Y)$ defined either as:

$$
\mathcal{D}(f)\left(\sum_{i} r_{i}\left|x_{i}\right\rangle\right):=\sum_{i} r_{i}\left|f\left(x_{i}\right)\right\rangle \quad \text { or as: } \quad \mathcal{D}(f)(\omega)(y):=\sum_{x \in f^{-1}(y)} \omega(x) .
$$

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 $\omega$ 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.

$$
\sum_{y \in Y} \mathcal{D}(f)(\omega)(y)=\sum_{y \in Y} \sum_{x \in f^{-1}(y)} \omega(x)=\sum_{x \in X} \omega(x)=1 .
$$

We present two examples where functoriality of $\mathcal{D}$ is used. It provides a powerful technique to produce new distributions from old, as images.

## Example 2.1.4.

1 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 $\omega \in \mathcal{D}\left(X_{1} \times \cdots \times X_{n}\right)$ is defined as the image distribution $\mathcal{D}\left(\pi_{i}\right)(\omega) \in \mathcal{D}\left(X_{i}\right)$, obtained via the $i$-th projection function $\pi_{i}: X_{1} \times \cdots \times X_{n} \rightarrow X_{i}$.

For instance, the first marginal of the joint distribution,

$$
\omega=\frac{1}{12}|H, 0\rangle+\frac{1}{6}|H, 1\rangle+\frac{1}{3}|H, 2\rangle+\frac{1}{6}|T, 0\rangle+\frac{1}{12}|T, 1\rangle+\frac{1}{6}|T, 2\rangle
$$

on the product space $\{H, T\} \times\{0,1,2\}$ is the distribution on $\{H, T\}$ that is computed explicitly as:

$$
\begin{aligned}
\mathcal{D}\left(\pi_{1}\right)(\omega)= & \frac{1}{12}\left|\pi_{1}(H, 0)\right\rangle+\frac{1}{6}\left|\pi_{1}(H, 1)\right\rangle+\frac{1}{3}\left|\pi_{1}(H, 2)\right\rangle \\
& \quad+\frac{1}{6}\left|\pi_{1}(T, 0)\right\rangle+\frac{1}{12}\left|\pi_{1}(T, 1)\right\rangle+\frac{1}{6}\left|\pi_{1}(T, 2)\right\rangle \\
= & \frac{1}{12}|H\rangle+\frac{1}{6}|H\rangle+\frac{1}{3}|H\rangle+\frac{1}{6}|T\rangle+\frac{1}{12}|T\rangle+\frac{1}{6}|T\rangle \\
= & \frac{7}{12}|H\rangle+\frac{5}{12}|T\rangle .
\end{aligned}
$$

In the same way one obtains as second marginal $\mathcal{D}\left(\pi_{2}\right)(\omega)=\frac{1}{4}|0\rangle+\frac{1}{4}|1\rangle+$ $\frac{1}{2}|3\rangle$.
2 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 : pips $\times$ pips $\rightarrow$ 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 $=\sum_{i, j \in p i p s} \frac{1}{36}|i, j\rangle$. Then, in functional form:

$$
\begin{aligned}
\mathcal{D}(\max )(\text { unif })(k) & =\sum_{i, j \text { with } \max (i, j)=k} \operatorname{unif}(i, j) \\
& =\sum_{i \leq k} \operatorname{unif}(i, k)+\sum_{j<k} \operatorname{unif}(k, j)=\frac{2 k-1}{36} .
\end{aligned}
$$

We can write this image distribution in ket notation as:

$$
\mathcal{D}(\max )(\text { unif })=\frac{1}{36}|1\rangle+\frac{3}{36}|2\rangle+\frac{5}{36}|3\rangle+\frac{7}{36}|4\rangle+\frac{9}{36}|5\rangle+\frac{11}{36}|6\rangle .
$$

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 product $\otimes$ of two dices - which are themselves uniform distributions. This is the same, since products $\otimes$ of uniform distributions are uniform, see Exercise 2.3.3

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. 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 $\omega$ and $\rho$, namely:

$$
\begin{equation*}
\sigma:=s \cdot \omega+(1-s) \cdot \rho \quad \text { that is } \quad \sigma(x)=r \cdot \omega(x)+(1-s) \cdot \rho(x) . \tag{2.2}
\end{equation*}
$$

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
At this stage we shall not axiomatise structures with such convex sums; they are sometimes called 'convex sets' or 'barycentric algebras', see [171] or [72] for details. A brief historical account occurs in [111, 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 2.1.5. For an arbitrary set $X$ we can form a set $\mathcal{D}_{\infty}(X)$ of distributions with (possibly) infinite support as:

$$
\mathcal{D}_{\infty}(X):=\left\{\omega: X \rightarrow[0,1] \mid \sum_{x \in X} \omega(x)=1\right\} .
$$

This operation $\mathcal{D}_{\infty}$ is functorial too, like $\mathcal{D}$ : for a function $f: X \rightarrow Y$ there is a function $\mathcal{D}_{\infty}(f): \mathcal{D}_{\infty}(X) \rightarrow \mathcal{D}_{\infty}(Y)$ given by $\mathcal{D}_{\infty}(f)\left(\sum_{i} r_{i}\left|x_{i}\right\rangle\right)=\sum_{i} r_{i}\left|f\left(x_{i}\right)\right\rangle$.

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:

$$
\sum_{n \in \mathbb{N}}(1-r) \cdot r^{n}|n\rangle \in \mathcal{D}_{\infty}(\mathbb{N}) .
$$

For instance, $r=\frac{1}{4}$ gives an infinite distribution $\sum_{n} \frac{3}{4^{n+1}}|n\rangle=\frac{3}{4}|0\rangle+\frac{3}{16}|0\rangle+$ $\frac{3}{64}|0\rangle+\cdots$.
The sum $\sum$ in the above formulation in Definition 2.1.5 is assumed to exist, as limit of finite sums. We show that it involves at most countably many nonzero probabilities.

Lemma 2.1.6. 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:

$$
\operatorname{supp}_{n}(\omega):=\left\{x \in X \left\lvert\, \omega(x)>\frac{1}{n}\right.\right\} .
$$

Since $\sum_{x \in X} \omega(x)=1$ this set $\operatorname{supp}_{n}(\omega)$ 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:

$$
\operatorname{supp}(\omega)=\{x \in X \mid \omega(x)>0\}=\bigcup_{n \in \mathbb{N}} \operatorname{supp}_{n}(\omega) .
$$

We see that $\operatorname{supp}(\omega)$ 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 2.1.10 contains another example, namely the geometric distribution. There are many more distributions with infinite support, such as the zeta (or zipf) distribution, see e.g. [159].

## Examples 2.1.7.

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:

$$
\begin{equation*}
\operatorname{pois}[\lambda]:=\sum_{k \in \mathbb{N}} e^{-\lambda} \cdot \frac{\lambda^{k}}{k!}|k\rangle \in \mathcal{D}_{\infty}(\mathbb{N}) \tag{2.3}
\end{equation*}
$$

The multiplicities add up to one in the Poisson distribution because of the well known formula describing an exponential via an infinite sum:

$$
\begin{equation*}
e^{\lambda}=\sum_{k \in \mathbb{N}} \frac{\lambda^{k}}{k!} \tag{2.4}
\end{equation*}
$$

The Poisson distribution is typically used for counts of rare events. The rate or intensity parameter $\lambda$ 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 $X$. This happens as follows. Let $X$ have $N$ elements.

$$
\text { mpois }[\lambda]:=\sum_{\varphi \in \mathcal{N}(X)} e^{-\lambda} \cdot \frac{\left(\frac{\lambda}{N}\right)^{\|\varphi\|}}{\varphi \rrbracket}|\varphi\rangle \in \mathcal{D}(\mathcal{N}(X)) .
$$

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 $\operatorname{Pmn}[\lambda]\left(u n i f_{X}\right)$, see Definition 3.9.1, for the uniform distribution on $X$.

The probabilities in mpois $[\lambda]$ add up to one since:

$$
\begin{aligned}
\sum_{\varphi \in \mathcal{N}(X)} e^{-\lambda} \cdot \frac{\left(\frac{\lambda}{N}\right)^{\|\varphi\|}}{\varphi!} & =e^{-\lambda} \cdot \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\left(\frac{\lambda}{N}\right)^{\|\varphi\|}}{\varphi!} \\
& =e^{-\lambda} \cdot \sum_{K \in \mathbb{N}} \frac{\left(\frac{\lambda}{N}\right)^{K}}{K!} \sum_{\varphi \in \mathcal{N}[K](X)} \frac{K!}{\varphi \rrbracket} \\
& =e^{-\lambda} \cdot \sum_{K \in \mathbb{N}} \frac{\lambda^{K}}{N^{K} \cdot K!} \cdot N^{K} \quad \text { by Exercise } 1.7 .7 \\
& \frac{[2.4]}{-} 1 .
\end{aligned}
$$

This is a countable sum since $\mathcal{N}(X) \cong \mathbb{N}^{N}$, see Exercise 1.6.5. The multiset Poisson and ordinary Poisson distributions are connected via the size function, see Exercise 2.1.11, in a so-called sufficient statistic situation, see Exercise 4.3.3
3 Our next example of an infinite distribution is the negative binomial distribution, of the form $n b n[K](s) \in \mathcal{D}_{\infty}(\mathbb{N})$, 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:

$$
\begin{aligned}
\operatorname{nbn}[K](s) & :=\sum_{n \in \mathbb{N}}\left(\binom{K}{n}\right) \cdot s^{K} \cdot(1-s)^{n}|n\rangle \\
& =\sum_{n \in \mathbb{N}}\binom{K+n-1}{K-1} \cdot s^{K} \cdot(1-s)^{n}|n\rangle \\
& =\sum_{m \geq K}\binom{m-1}{K-1} \cdot s^{K} \cdot(1-s)^{m-K}|m-K\rangle .
\end{aligned}
$$

One can use Theorem 1.7.4 (or Exercise 1.7.14) to show that this forms a distribution.

$$
\begin{aligned}
\sum_{n \in \mathbb{N}}\left(\binom{K}{n}\right) \cdot s^{K} \cdot(1-s)^{n} & =s^{K} \cdot \sum_{n \in \mathbb{N}}\binom{n+K-1}{K-1} \cdot(1-s)^{n} \\
& =s^{K} \cdot \frac{1}{(1-(1-s))^{K}} \quad \text { by Theorem 1.7.4 (1) } \\
& =1 .
\end{aligned}
$$

We shall encounter multivariate negative distributions in Section ??.

## Exercises

2.1.1 Check that a marginal of a uniform distribution is again a uniform distribution; more precisely, $\mathcal{D}\left(\pi_{1}\right)\left(u n i f_{X \times Y}\right)=u n i f_{X}$, for non-empty finite sets $X, Y$.
2.1.2 1 Prove that flip: $[0,1] \rightarrow \mathcal{D}(\mathbf{2})$ is an isomorphism.

2 Check that flip $(r)$ is the same as $b n[1](r)$.
3 Describe the distribution bn[3]((%5Cfrac%7B1%7D%7B4%7D)) concretely and interpret this distribution in terms of coin flips.
2.1.3 Recall that $\boldsymbol{n}:=\{0, \ldots, n-1\}$ and check that:

$$
\mathcal{D}(\mathbf{0}) \cong \mathbf{0} \quad \mathcal{D}(\mathbf{1}) \cong \mathbf{1} \quad \mathcal{D}(\mathbf{2}) \cong[0,1]
$$

The set $\mathcal{D}(\boldsymbol{n}+\mathbf{1})$ 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 $\mathbb{R}^{n}$.
2.1.4 Consider the binomial distribution $b n[K]$ from Example 2.1.2 (2). Show that for $r \in[0,1]$ and $k \in\{0,1, \ldots, K\}$,

$$
b n[K](r)(K-k)=b n[K](1-r)(k) .
$$

Check that this means that the rectangle

commutes, where minus $(x, y):=x-y$.
2.1.5 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 $b n[K](r)(k)$, that is, what is $\operatorname{argmax}_{r \in[0,1]} b n[K](r)(k)$ ?
1 What do you expect this probability to be?
2 Write $f(r)=\ln (b n[K](r)(k))$, where $\ln$ is the natural logarithm. Check for yourself that we may as well determine the maximum of $f$.

3 Check that the derivative of $f$ is:

$$
f^{\prime}(r)=\frac{k}{r}-\frac{K-k}{1-r}
$$

4 Show that solving $f^{\prime}(r)=0$ gives $r=\frac{k}{K}$.
Proposition 3.3.10 will generalise this result to multinomial distributions.
2.1.6 Let a number $r \in[0,1]$ and a finite set $X$ be given. Show that:

$$
\sum_{U \in \mathcal{P}_{(X)}} r^{|U|} \cdot(1-r)^{|X \backslash U|}=1 .
$$

Hint: Recall the binomial distribution from Example 2.1.2 (2) and recall also Exercise 1.5.6
2.1.7 Let $\omega \in \mathcal{D}(X)$ be a distribution, considered as a function $\omega: X \rightarrow$ [ 0,1 ] with finite support. Use functoriality of $\mathcal{D}$ to show that:

$$
\mathcal{D}(\omega)(\omega)=\sum_{x \in \operatorname{supp}(\omega)} \omega(x)|\omega(x)\rangle \in \mathcal{D}([0,1]) .
$$

2.1.8 Check that the powerbag operation from Exercise 1.8 .4 can be turned into a probabilistic powerbag PPB via:

$$
\operatorname{PPB}(\psi):=\sum_{\varphi \leq \psi} \frac{\binom{\psi}{\varphi}}{2^{\|\psi\|}}|\varphi\rangle .
$$

2.1.9 1 Recall Theorem 1.7.4 (1) and conclude that $\lim _{n \rightarrow \infty}\binom{n+K}{K} \cdot r^{n}=0$, 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:

$$
\lim _{n \rightarrow \infty} b n[n+m](r)(m)=0 .
$$

Explain yourself what this means.
2.1.10 For a (non-zero, non-one) probability $r \in(0,1)$ one defines the geometric distribution geo $[r] \in \mathcal{D}_{\infty}\left(\mathbb{N}_{>0}\right)$ as:

$$
\operatorname{geo}[r]:=\sum_{k \in \mathbb{N}_{>_{0}}} r \cdot(1-r)^{k-1}|k\rangle .
$$

It captures the probability of being successful for the first time after $k-1$ unsuccesful tries. Prove that this is a distribution indeed: its multiplicities add up to one.
2.1.11 Recall the ordinary and multiset Poisson distributions (channels) from Example 2.1.7(1) and (2). Show that they are connected by size, as in the following diagram.

2.1.12 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 $\omega^{\star}: \mathcal{L}(X) \rightarrow[0,1]$ by:

$$
\omega^{\star}\left(\left[x_{1}, \ldots, x_{n}\right]\right):=\frac{\omega\left(x_{1}\right) \cdot \ldots \cdot \omega\left(x_{n}\right)}{2^{n+1}}
$$

1 Prove that $\omega^{\star} \in \mathcal{D}_{\infty}(\mathcal{L}(X))$.
2 Consider the function $f:\{a, b, c\} \rightarrow\{1,2\}$ 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.

$$
\mathcal{D}(f)(\omega)^{\star}(\ell)=\frac{245}{27648}=\mathcal{D}_{\infty}(\mathcal{L}(f))\left(\omega^{\star}\right)(\ell)
$$

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:


3 Prove in general that $(-)^{\star}$ is a natural transformation from $\mathcal{D}$ to $\mathcal{D}_{\infty} \circ \mathcal{L}$.

### 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. When a probrability 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 part of this section focuses on 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 1.34. We now add the mapping Flrn: $\mathcal{M}_{*}(X) \rightarrow \mathcal{D}(X)$, 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 4.5 .9 we show that these outcomes of frequentist learning ly dense in the set of 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 $\sum_{i} r_{i}\left|x_{i}\right\rangle$, with $r_{i} \neq 0$ for at least one index $i$. Equivalently one can require that the size $s:=$ $\sum_{i} r_{i}=\| \sum_{i} r_{i}\left|x_{i}\right\rangle \|$ is non-zero.
The Flrn maps turns a (non-empty) multiset into a distribution, essentially by normalisation. It is defined as follows.

$$
\begin{align*}
\operatorname{Flrn}\left(r_{1}\left|x_{1}\right\rangle+\cdots+r_{k}\left|x_{k}\right\rangle\right):= & \frac{r_{1}}{s}\left|x_{1}\right\rangle+\cdots+\frac{r_{k}}{s}\left|x_{k}\right\rangle  \tag{2.5}\\
& \text { where } s:=\sum_{i} r_{i} .
\end{align*}
$$

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:

$$
\begin{equation*}
\operatorname{Flrn}(\varphi):=\frac{1}{\|\varphi\|} \cdot \varphi \quad \text { where } \quad\|\varphi\|:=\sum_{x} \varphi(x) \tag{2.6}
\end{equation*}
$$

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 $v \in \mathcal{N}(X)$ over a set of colours $X$, the probability of drawing a ball of colour $x \in X$ from $v$ is given by the fraction of balls of colour $X$ and the total number of balls in the urn:

$$
\operatorname{Flrn}(v)(x)=\frac{v(x)}{\|v\|} \in[0,1] .
$$

This formalises the idea that was presented pictorially 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, 20 times come up head $(H)$ and 30 yield tail ( $T$ ). We can represent these 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:

$$
F \operatorname{lrn}(\varphi)=\frac{20}{20+30}|H\rangle+\frac{30}{20+30}|T\rangle=\frac{2}{5}|H\rangle+\frac{3}{5}|T\rangle
$$

Thus, the bias (twowards 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 (2.7) 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 (1.28) captured by the multiset $\tau \in \mathcal{N}(B \times M)$. Learning from $\tau$ yields the following joint distribution:

$$
\begin{aligned}
& \operatorname{Flrn}(\tau)=\frac{10}{100}|H, 0\rangle+\frac{35}{100}|H, 1\rangle+\frac{25}{100}|H, 2\rangle \\
&+\frac{5}{100}|L, 0\rangle+\frac{10}{100}|L, 1\rangle+\frac{15}{100}|L, 2\rangle .
\end{aligned}
$$

Such a distribution, directly derived from a table, is sometimes called an empirical distribution [37].

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:

$$
\begin{equation*}
\operatorname{Flrn}(s \cdot \varphi)=\operatorname{Flrn}(\varphi) \quad \text { for } s \in \mathbb{R}_{>0} . \tag{2.7}
\end{equation*}
$$

In [62] 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.

The following elementary observation captures the essence of frequentist learning.

Lemma 2.2.3. For each set $X$ the tuple of size and frequentist learning forms an isomorphism in:

$$
\mathcal{M}_{*}(X) \xrightarrow[\cong]{\lfloor\text { size, Flrn }\rangle} \mathbb{R}_{>0} \times \mathcal{D}(X)
$$

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 $s \cdot \omega=\sum_{x} s \cdot \omega(x)|x\rangle$.

It turns out that the learning map Flrn is 'natural', in the sense that it works uniformly for each set.

Lemma 2.2.4. The frequentist learning maps Flrn: $\mathcal{M}_{*}(X) \rightarrow \mathcal{D}(X)$ from 2.5 are natural in $X$. This means that for each function $f: X \rightarrow Y$ the following diagram commutes.


As a special case, frequentist learning commutes with marginalisation, via projection functions.

Proof. Pick an arbitrary non-empty multiset $\varphi=\sum_{i} r_{i}\left|x_{i}\right\rangle$ in $\mathcal{M}_{*}(X)$ and write
$s:=\|\varphi\|=\sum_{i} r_{i}$. By non-emptyness of $\varphi$ we have $s \neq 0$. Then:

$$
\begin{aligned}
\left(F l r n \circ \mathcal{M}_{*}(f)\right)(\varphi) & =F \operatorname{lrn}\left(\sum_{i} r_{i}\left|f\left(x_{i}\right)\right\rangle\right) \\
& \left.=\sum_{i} \frac{r_{i}}{s}\left|f\left(x_{i}\right)\right\rangle\right) \\
& =\mathcal{D}(f)\left(\sum_{i} \frac{r_{i}}{s}\left|x_{i}\right\rangle\right) \\
& =(\mathcal{D}(f) \circ F \operatorname{lrn})(\varphi) .
\end{aligned}
$$

We can apply this basic result to the medical data in Table 1.28, , via the multiset $\tau \in \mathcal{N}(B \times M)$. We have already seen in Section 1.6 that the multisetmarginals $\mathcal{N}\left(\pi_{i}\right)(\tau)$ produce the marginal columns and rows, with their totals. We can learn the distributions from the colums as:

$$
\operatorname{Flrn}\left(\mathcal{M}\left(\pi_{1}\right)(\tau)\right)=\operatorname{Flrn}(70|H\rangle+30|L\rangle)=\frac{7}{10}|H\rangle+\frac{3}{10}|L\rangle
$$

We can also take the distribution-marginal of the 'learned' distribution from the table, as described in Example 2.2.2 (2):

$$
\begin{aligned}
\mathcal{M}\left(\pi_{1}\right)(\operatorname{Flrn}(\tau)) & =\left(\frac{10}{100}+\frac{35}{100}+\frac{25}{100}\right)|H\rangle+\left(\frac{5}{100}+\frac{10}{100}+\frac{15}{100}\right)|L\rangle \\
& =\frac{7}{10}|H\rangle+\frac{3}{10}|L\rangle
\end{aligned}
$$

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 2.2.4

Remark 2.2.5. In Lemma 2.2.4 we have seen that Flrn: $\mathcal{M}_{*} \Rightarrow \mathcal{D}$ 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 1.11.2. This is not the case.

It is easy to see that Flrn commutes with the unit maps, simply because $\operatorname{Flrn}(1|x\rangle)=1|x\rangle$. But commutation with flatten's fails. Here is a simple counterexample. Consider the multiset of multsets $\Phi \in \mathcal{M}(\mathcal{M}(\{a, b, c\}))$ given by:

$$
\Phi:=1|2| a\rangle+4|c\rangle\rangle+2|1| a\rangle+1|b\rangle+1|c\rangle\rangle .
$$

First flattening the multiset, and then doing frequentist learning gives:

$$
\operatorname{Flrn}(\operatorname{flat}(\Phi))=\operatorname{Flrn}(4|a\rangle+2|b\rangle+6|c\rangle)=\frac{1}{3}|a\rangle+\frac{1}{6}|b\rangle+\frac{1}{2}|c\rangle
$$

However, first (outer en inner) learning and then flattening the resulting distribution of distributions yields:

$$
\begin{aligned}
\operatorname{flat}(\text { Flrn }(\mathcal{M}(\text { Flrn })(\Phi))) & \left.\left.\left.\left.=\operatorname{flat}\left(\frac{1}{3}\left|\frac{1}{3}\right| a\right\rangle+\frac{2}{3}|c\rangle\right\rangle+\frac{2}{3}\left|\frac{1}{3}\right| a\right\rangle+\frac{1}{3}|b\rangle+\frac{1}{3}|c\rangle\right\rangle\right) \\
& =\frac{1}{3}|a\rangle+\frac{2}{9}|b\rangle+\frac{4}{9}|c\rangle .
\end{aligned}
$$

### 2.2.1 Sampling

Sampling is a technique for choosing individual elements from the support $\operatorname{supp}(\omega) \subseteq X$ of a distribution $\omega \in \mathcal{D}(X)$, in accordance with the probabilities in $\omega$. 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: $\operatorname{Flrn}(5|0\rangle+4|1\rangle+1|2\rangle)=\frac{1}{2}|0\rangle+\frac{2}{5}|1\rangle+\frac{1}{5}|2\rangle$.

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: 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 $\omega$ increases, the resulting multiset $\varphi_{N}$ of samples, of size $N$, approaches $\omega$ via frequentist learning: $\operatorname{Flrn}\left(\varphi_{N}\right) \rightarrow \omega$ as $N \rightarrow \infty$.

Thus, informally, sampling may be seen as an approximate inverse to fre-
quentist learning. It exists in many probabilistic programming languages, see e.g. [64] for an overview. Sampling a single element $x$ from a distribution $\omega \in \mathcal{D}(X)$ may be written as:

$$
x \leftarrow \omega \quad \text { or as } \quad x \sim \omega
$$

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 pseudorandom 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 distribution $\sum_{i} r_{i}\left|x_{i}\right\rangle$, 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 $f_{100}$ be the uniform distribution on $\{1,2, \ldots, 100\}$.

$$
\begin{align*}
& \text { i } \leftarrow \text { unif }_{100} \\
& j \leftarrow \text { unif }_{100} \\
& \mathrm{k} \leftarrow \text { unif }_{100} \\
& \text { if } i<=i \text { and } j<=\mathrm{k}:  \tag{2.8}\\
& \quad \text { return yes } \\
& \text { else }: \\
& \quad \text { return no }
\end{align*}
$$

This program yields as distribution, approximately, $0.17 \mid$ yes $\rangle+0.83 \mid$ no $\rangle$.

This book is not about such probabilistic programming: it has a more mathematical ('exact') focus. From this mathematical perspective, the multinomial distribution $m n[K](\omega) \in \mathcal{D}(\mathcal{N}[K](X))$, that we introduce later on, can be understood as a 'correct' distribution over multiset samples, see Theorem 3.3.3 and also Theorem 5.5.4, 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. [165] or [106, Defn. 1]) that will be introduced below and will be used occasionally in the sequel. We follow the description of [86] and start with some basic observations.

We write $\mathcal{D}_{f s}(\{R, G, B, Y\})$ for the set of all distributions with full support on the set $\{R, G, B, Y\}$ of four colours (Red, Green, Blue, Yellow). Explicitly:

$$
\begin{aligned}
\mathcal{D}_{f s}(\{R, G, B, Y\})=\left\{r_{0}|R\rangle+r_{1}|G\rangle+r_{2}|B\rangle+\right. & r_{3}|Y\rangle \mid r_{0}, r_{1}, r_{2}, r_{3} \in(0,1) \\
& \text { with } \left.r_{0}+r_{1}+r_{2}+r_{3}=1\right\} .
\end{aligned}
$$

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:

$$
\mathcal{D}_{f s}(\{R, G, B, Y\}) \cong\left\{\left(r_{0}, r_{1}, r_{2}\right) \in(0,1)^{3} \mid r_{0}+r_{1}+r_{2}<1\right\} .
$$

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 yields an isomorphism:

$$
\begin{equation*}
\mathcal{D}_{f s}(\{R, G, B, Y\}) \cong(0,1)^{3} . \tag{2.9}
\end{equation*}
$$

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 $s_{1}\left(1-s_{0}\right)$. The unpainted part is now $\left(1-s_{2}\right)(1-$ $s_{0}$ ). We paint the $s_{2}$-proportion of this remainder
 blue. The final remainder is then of length ( $1-$ $\left.s_{2}\right)\left(1-s_{2}\right)\left(1-s_{0}\right)$. 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 $\operatorname{map}(0,1)^{3} \rightarrow \mathcal{D}_{f s}(\{R, G, B, Y\})$.

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 $N=\{0,1, \ldots, N-1\}$, there is a "stick breaking" isomorphism:

$$
\begin{align*}
& (0,1)^{N-1} \xrightarrow{s t b r} \\
& \left(r_{0}, \ldots, r_{N-2}\right) \longmapsto \mathcal{D}_{f s}(\boldsymbol{N})  \tag{2.10}\\
& \\
& \quad r_{0}|0\rangle+r_{0}^{\perp} r_{1}|1\rangle+r_{0}^{\perp} r_{1}^{\perp} r_{2}|2\rangle+\cdots \\
& \quad+r_{0}^{\perp} \cdots r_{N-3}^{\perp} r_{N-2}|N-2\rangle+r_{0}^{\perp} \cdots r_{N-2}^{\perp}|N-1\rangle .
\end{align*}
$$

2 This isomorphism extends to countably infinite sequences, as:

$$
\begin{gather*}
(0,1)^{\mathbb{N}} \xrightarrow{\stackrel{s t b r}{\cong}} \mathcal{D}_{\infty, f s}(\mathbb{N}) \\
\vec{r} \longmapsto \tag{2.11}
\end{gather*}
$$

where $\mathcal{D}_{\infty, \text { fs }}$ is used for distributions with full support, without finiteness restriction, see Definition 2.1.5.

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 (2.10) satisfies for $1 \leq i<N-1$,

$$
\begin{equation*}
1-\sum_{j \leq i} \operatorname{stbr}(\vec{r})(j)=\prod_{j \leq i} r_{j}^{\perp} \tag{2.12}
\end{equation*}
$$

This gives $\sum_{i<N} \operatorname{stbr}(\vec{r})(i)=1$, so that stickbreaking produces a probability distribution, on $N$.

In the reverse direction we start from a distribution $\omega \in \mathcal{D}(N)$ and define a sequence of $N-1$ numbers in $(0,1)$ by:

$$
\operatorname{stbr}^{-1}(\omega)_{0}:=\omega(0) \quad \text { and } \quad \operatorname{stbr}^{-1}(\omega)_{i+1}:=\frac{\omega(i+1)}{1-\sum_{j \leq i} \omega(j)}
$$

This inverse operation satisfies, for $0 \leq i<N-2$,

$$
\begin{equation*}
\prod_{j \leq i}\left(s t b r^{-1}(\omega)_{j}\right)^{\perp}=1-\sum_{j \leq i} \omega(j) \tag{2.13}
\end{equation*}
$$

These operations are each other's inverses, since:

$$
\begin{aligned}
\operatorname{stbr}^{-1}(\operatorname{stbr}(\vec{r}))_{0} & =\operatorname{stbr}(\vec{r})(0)=r_{0} \\
\operatorname{stbr}^{-1}(\operatorname{stbr}(\vec{r}))_{i+1} & =\frac{\operatorname{stbr}(\vec{r})(i+1)}{1-\prod_{j \leq i} \operatorname{stbr}(\vec{r})(j)} \frac{[2.12]}{=} \frac{\left(\prod_{j \leq i} r_{j}^{\perp}\right) r_{i+1}}{\prod_{j \leq i} r_{j}^{\perp}}=r_{i+1} .
\end{aligned}
$$

And, the other way around, the case $i=0$ is simple again. And for $0<i<$ $N-1$,

$$
\begin{aligned}
\operatorname{stbr}\left(\operatorname{stbr}^{-1}(\omega)\right)(i+1) & =\left(\prod_{j \leq i}\left(s t b r^{-1}(\omega)_{j}\right)^{\perp}\right) \operatorname{stbr}^{-1}(\omega)_{i+1} \\
& \stackrel{\boxed{2.13}}{=}\left(1-\sum_{j \leq i} \omega(j)\right) \frac{\omega(i+1)}{1-\sum_{j \leq i} \omega(j)}=\omega(i+1)
\end{aligned}
$$

2 For an infinite sequence $\vec{r} \in(0,1)^{\mathbb{N}}$ definition 2.11 yields a probability distribution since:

$$
\begin{aligned}
1-\sum_{n \in \mathbb{N}} \operatorname{stbr}(\vec{r})(n) & =1-\lim _{N \rightarrow \infty} \sum_{n \leq N}\left(\prod_{i<n} r_{i}^{\perp}\right) r_{n} \\
& =\lim _{N \rightarrow \infty} 1-\sum_{n \leq N} \operatorname{stbr}\left(r_{0}, \ldots, r_{n}\right)(n) \\
& \lim _{N \rightarrow \infty} \prod_{n \leq N} r_{n}^{\perp}=0 .
\end{aligned}
$$

The latter holds since $r_{n}^{\perp}=1-r_{n}<1$ for each $n$.

In the other direction we can define, for $\omega \in \mathcal{D}_{\infty, f s}(X)$,

$$
\operatorname{stbr}^{-1}(\omega)_{i}:=\frac{\omega(i)}{1-\sum_{j<i} \omega(j)}
$$

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 $N=4$.

$$
\begin{aligned}
\operatorname{stbr}\left(\frac{3}{4}, \frac{2}{3}, \frac{1}{5}\right) & =\frac{3}{4}|0\rangle+\frac{1}{4} \cdot \frac{2}{3}|1\rangle+\frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{5}|2\rangle+\frac{1}{4} \cdot \frac{1}{3} \cdot \frac{4}{5}|3\rangle \\
& =\frac{3}{4}|0\rangle+\frac{1}{6}|1\rangle+\frac{1}{60}|2\rangle+\frac{1}{15}|3\rangle .
\end{aligned}
$$

In the other direction we get, as expected, the original sequence:

$$
\begin{aligned}
\operatorname{stbr}^{-1}\left(\frac{3}{4}|0\rangle+\frac{1}{6}|1\rangle+\frac{1}{60}|2\rangle+\frac{1}{15}|3\rangle\right) & =\left(\frac{3}{4}, \frac{1 / 6}{1-3 / 4}, \frac{1 / 60}{1-3 / 4-1 / 6}\right) \\
& =\left(\frac{3}{4}, \frac{1 / 6}{1 / 4}, \frac{1 / 60}{1 / 12}\right) \\
& =\left(\frac{3}{4}, \frac{2}{3}, \frac{1}{5}\right) .
\end{aligned}
$$

2 Consider the infinite distribution:

$$
\omega=\sum_{n \in \mathbb{N}} \frac{2}{5} \cdot\left(\frac{3}{5}\right)^{n}|n\rangle=\frac{2}{5}|0\rangle+\frac{6}{25}|1\rangle+\frac{18}{125}|2\rangle+\frac{54}{625}|3\rangle+\cdots
$$

We can see that it is a distribution via Theorem 1.7.4(2):

$$
\sum_{n \in \mathbb{N}} \omega(n)=\frac{2}{5} \cdot \sum_{n \in \mathbb{N}}\left(\frac{3}{5}\right)^{n}=\frac{2}{5} \cdot \frac{1}{1-3 / 5}=\frac{2}{5-3}=1
$$

The sequence of numbers in $(0,1)$ corresponding to $\omega$ is constant:

$$
\operatorname{stbr}^{-1}(\omega)=\left(\frac{2}{5}, \frac{2}{5}, \frac{2}{5}, \ldots\right)
$$

In general, for $r \in(0,1)$, we have $\operatorname{stbr}(r, r, r, \ldots)=\sum_{n \in \mathbb{N}} r(1-r)^{n}|n\rangle$.

## 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 (1.26) and 1.27?
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 (2.7).
2.2.3 1 Prove that for multisets $\varphi, \psi \in \mathcal{M}_{*}(X)$ one has:

$$
\operatorname{Flrn}(\varphi+\psi)=\frac{\|\varphi\|}{\|\varphi\|+\|\psi\|} \cdot \operatorname{Flrn}(\varphi)+\frac{\|\psi\|}{\|\varphi\|+\|\psi\|} \cdot \operatorname{Flrn}(\psi)
$$

This means that when one has already learned $\operatorname{Flrn}(\varphi)$ and new data $\psi$ arrives, all probabilities have to be adjusted, as in the above convex sum of distributions.
2 Check that the following formulation for natural multisets of fixed sizes $K>0, L>0$ is a special case of the previous item.

2.2.4 Show that Diagram (1.34) can be refined to:

where $\mathcal{L}_{*}(X) \subseteq \mathcal{L}(X)$ is the subset of non-empty lists.
2.2.5 Consider the program in (2.8) and show analytically that the outcome distribution is $\left.\frac{1717}{10000} \right\rvert\,$ yes $\rangle \left.+\frac{8283}{10000} \right\rvert\,$ no $\rangle$.
2.2.6 Consider $\omega=\frac{1}{6}|0\rangle+\frac{1}{5}|1\rangle+\frac{1}{4}|2\rangle+\frac{1}{3}|3\rangle+\frac{1}{20}|4\rangle \in \mathcal{D}(5)$.

1 Calculate $\vec{x}:=\operatorname{stbr}^{-1}(\omega) \in(0,1)^{4}$.
2 Check what $\operatorname{stbr}(\vec{x})$ is.
2.2.7 $\quad$ Fix $N>1$.

1 Let $v=\operatorname{unif}_{N}=\sum_{i \in N} \frac{1}{N}|i\rangle$. Show that:

$$
\operatorname{stbr}^{-1}(v)=\left(\frac{1}{N}, \frac{1}{N-1}, \frac{1}{N-2}, \ldots, \frac{1}{2}\right) \in(0,1)^{N-1} .
$$

2 Show that for an arbitrary $r \in(0,1)$ one has:

$$
\operatorname{stbr}(\underbrace{r, \ldots, r}_{N-1 \text { times }})=\sum_{0 \leq i<N-1} r(1-r)^{i}|i\rangle+(1-r)^{N-1}|N-1\rangle .
$$

2.2.8 Prove Equations (2.12) and (2.13).

### 2.3 Parallel products

Early on this book, 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.
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 $\otimes: \mathcal{P}(X) \times \mathcal{P}(Y) \rightarrow \mathcal{P}(X \times Y)$ and $\otimes: \mathcal{M}(X) \times \mathcal{M}(Y) \rightarrow$ $\mathcal{M}(X \times Y)$ and $\otimes: \mathcal{D}(X) \times \mathcal{D}(Y) \rightarrow \mathcal{D}(X \times Y)$.

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:

$$
U \otimes V:=\{(x, y) \in X \times Y \mid x \in U \text { and } y \in V\} .
$$

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:

$$
\varphi \otimes \psi:=\sum_{x \in X, y \in Y} \varphi(x) \cdot \psi(y)|x, y\rangle \text { that is }(\varphi \otimes \psi)(x, y)=\varphi(x) \cdot \psi(y) .
$$

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:

$$
\begin{aligned}
& \left(\frac{1}{6}|u\rangle+\frac{1}{3}|v\rangle+\frac{1}{2}|w\rangle\right) \otimes\left(\frac{3}{4}|0\rangle+\frac{1}{4}|1\rangle\right) \\
& \quad=\frac{1}{8}|u, 0\rangle+\frac{1}{24}|u, 1\rangle+\frac{1}{4}|v, 0\rangle+\frac{1}{12}|v, 1\rangle+\frac{3}{8}|w, 0\rangle+\frac{1}{8}|w, 1\rangle .
\end{aligned}
$$

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:

$$
|U \otimes V|=|U| \cdot|V| \quad\|\varphi \otimes \psi\|=\|\varphi\| \cdot\|\psi\| .
$$

2 Marginalisation separates tensors into the original components - assuming that the subsets $U, V$ are non-empty.

$$
\begin{array}{lll}
\mathcal{P}\left(\pi_{1}\right)(U \otimes V)=U & \mathcal{M}\left(\pi_{1}\right)(\varphi \otimes \psi)=\|\psi\| \cdot \varphi & \mathcal{D}\left(\pi_{1}\right)(\omega \otimes \rho)=\omega \\
\mathcal{P}\left(\pi_{2}\right)(U \otimes V)=V & \mathcal{M}\left(\pi_{2}\right)(\varphi \otimes \psi)=\|\varphi\| \cdot \psi & \mathcal{D}\left(\pi_{2}\right)(\omega \otimes \rho)=\rho .
\end{array}
$$

3 Tensors are natural: for functions $f: X \rightarrow A$ and $g: Y \rightarrow B$,

$$
\begin{aligned}
\mathcal{P}(f \times g)(U \otimes V) & =\mathcal{P}(f)(U) \otimes \mathcal{P}(g)(V) \\
\mathcal{M}(f \times g)(\varphi \otimes \psi) & =\mathcal{M}(f)(\varphi) \otimes \mathcal{M}(g)(\psi) \\
\mathcal{D}(f \times g)(\omega \otimes \rho) & =\mathcal{D}(f)(\omega) \otimes \mathcal{D}(g)(\rho) .
\end{aligned}
$$

4 The support maps supp: $\mathcal{M} \Rightarrow \mathcal{P}$ and supp: $\mathcal{D} \Rightarrow \mathcal{P}$ commute with tensors, in the sense that:

$$
\operatorname{supp}(\varphi \otimes \psi)=\operatorname{supp}(\varphi) \otimes \operatorname{supp}(\psi) \quad \operatorname{supp}(\omega \otimes \rho)=\operatorname{supp}(\omega) \otimes \operatorname{supp}(\rho) .
$$

5 The frequentist learning map Flrn: $\mathcal{M}_{*} \Rightarrow \mathcal{D}$ also commutes with tensors:

$$
\operatorname{Flrn}(\varphi \otimes \psi)=\operatorname{Flrn}(\varphi) \otimes \operatorname{Flrn}(\psi) .
$$

The last two items express that support and frequentist learning are 'monoidal' natural transformations.

Proof. 1 The case of subsets is trivial. For multisets:

$$
\|\varphi \otimes \psi\|=\sum_{x \in X, y \in Y} \varphi(x) \cdot \psi(y)=\left(\sum_{x \in X} \varphi(x)\right) \cdot\left(\sum_{y \in Y} \psi(y)\right)=\|\varphi\| \cdot\|\psi\| .
$$

2 For subsets we have

$$
\mathcal{P}\left(\pi_{1}\right)(U \otimes V)=\left\{\pi_{1}(x, y) \mid x \in U, y \in V\right\}=U, \quad \text { using that } V \neq \emptyset .
$$

For multisets the size shows up:

$$
\begin{aligned}
\mathcal{M}\left(\pi_{1}\right)(\varphi \otimes \psi) & =\sum_{x \in X, y \in Y} \varphi(x) \cdot \psi(y)\left|\pi_{1}(x, y)\right\rangle \\
& =\sum_{x \in X}\left(\sum_{y \in Y} \varphi(x) \cdot \psi(y)\right)|x\rangle \\
& =\sum_{x \in X} \varphi(x) \cdot\left(\sum_{y \in Y} \psi(y)\right)|x\rangle=\sum_{x \in X} \varphi(x) \cdot\|\psi\||x\rangle=\|\psi\| \cdot \varphi .
\end{aligned}
$$

For distributions $\omega, \rho$, the computation $\mathcal{D}\left(\pi_{1}\right)(\omega \otimes \rho)$ 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.

$$
\begin{aligned}
\mathcal{D}(f \times g)(\omega \otimes \rho) & =\sum_{x \in X, y \in Y} \omega(x) \cdot \rho(y)|(f \times g)(x, y)\rangle \\
& =\sum_{x \in X, y \in Y} \omega(x) \cdot \rho(y)|f(x), g(y)\rangle \\
& =\left(\sum_{x \in X} \omega(x)|f(x)\rangle\right) \otimes\left(\sum_{y \in y} \rho(y)|g(y)\rangle\right) \\
& =\mathcal{D}(f)(\omega) \otimes \mathcal{D}(g)(\rho) .
\end{aligned}
$$

4 We prove the required result for multisets; the proof for distributions works in the same way.

$$
\begin{aligned}
\operatorname{supp}(\varphi \otimes \psi) & =\{(x, y) \mid(\varphi \otimes \psi)(x, y) \neq 0\} \\
& =\{(x, y) \mid \varphi(x) \cdot \psi(y) \neq 0\} \\
& =\{(x, y) \mid \varphi(x) \neq 0 \text { and } \psi(y) \neq 0\} \\
& =\{(x, y) \mid x \in \operatorname{supp}(\varphi), y \in \operatorname{supp}(\psi)\} \\
& =\operatorname{supp}(\varphi) \otimes \operatorname{supp}(\psi) .
\end{aligned}
$$

5 We use commutation of size with tensor from item (1) in:

$$
\begin{aligned}
\operatorname{Flrn}(\varphi \otimes \psi)(x, y)=\frac{(\varphi \otimes \psi)(x, y)}{\|\varphi \otimes \psi\|} & =\frac{\varphi(x)}{\|\varphi\|} \cdot \frac{\psi(y)}{\|\psi\|} \\
& =\operatorname{Flrn}(\varphi)(x) \cdot \operatorname{Flrn}(\psi)(y) \\
& =(\operatorname{Flrn}(\varphi) \otimes \operatorname{Flrn}(\psi))(x, y)
\end{aligned}
$$

As promised we look into why parallel products do not work for lists.
Remark 2.3.3. 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:

$$
\begin{array}{r}
{[\langle a, u\rangle,\langle a, v\rangle,\langle b, u\rangle,\langle b, v\rangle,\langle c, u\rangle,\langle c, v\rangle]} \\
{[\langle a, u\rangle,\langle b, u\rangle,\langle c, u\rangle,\langle a, v\rangle,\langle b, v\rangle,\langle c, v\rangle] .}
\end{array}
$$

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 [112].

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 $\omega^{K}=\omega \otimes \cdots \otimes \omega \in \mathcal{D}\left(X^{K}\right)$. 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:

$$
\begin{equation*}
\mathcal{D}(X) \xrightarrow{\text { iid }[K]} \mathcal{D}\left(X^{K}\right) \quad \text { with } \quad \text { iid }[K](\omega):=\underbrace{\omega \otimes \cdots \otimes \omega}_{K \text { times }} . \tag{2.15}
\end{equation*}
$$

2 The $K$-ary tensors form a function:

$$
\begin{equation*}
\mathcal{D}(X)^{K} \xrightarrow{\otimes[K]} \mathcal{D}\left(X^{K}\right) \quad \text { with } \quad \bigotimes[K](\vec{\omega}):=\omega_{1} \otimes \cdots \otimes \omega_{K} . \tag{2.16}
\end{equation*}
$$

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 $\mathcal{D}\left(X^{K}\right)$ is also a singleton.

We can describe iid $[K]$ diagrammatically as composite, making the copying of distributions explicit:


We collect some basic properties, involving the zip isomorphism zip: $X^{K} \times$ $Y^{K} \xrightarrow{\cong}(X \times Y)^{K}$ from Exercises 1.3.7 and 1.11.4

Lemma 2.3.5. Fix a number $K \in \mathbb{N}$.

1 The iid and $\bigotimes$ maps are natural: for each function $f: X \rightarrow Y$ the following two diagrams commutes.



2 The zip function commutes with multiple tensors in the following way.


3 Zip also commutes with copied tensors:


Proof. 1 This is a direct consequence of Lemma 2.3.2 (3).
2 For distributions $\omega_{i} \in \mathcal{D}(X), \rho_{i} \in \mathcal{D}(Y)$ and elements $x_{i} \in X, y_{i} \in Y$ we elaborate:

$$
\begin{aligned}
& (\mathcal{D}(z i p) \circ \otimes \circ(\otimes \times \otimes))(\vec{\omega}, \vec{\rho}) \\
& =\sum_{\vec{x} \in X^{K}, \vec{y} \in Y^{K}}(\otimes(\vec{\omega}) \otimes \otimes(\vec{\rho}))(\vec{x}, \vec{y})|z i p(\vec{x}, \vec{y})\rangle \\
& =\sum_{\vec{x} \in X^{K}, \vec{y} \in Y^{K}} \otimes(\vec{\omega})(\vec{x}) \cdot \otimes(\vec{\rho})(\vec{y})|z i p(\vec{x}, \vec{y})\rangle \\
& =\sum_{\vec{x} \in X^{K}, \vec{y} \in Y^{K}} \omega_{1}\left(x_{1}\right) \cdot \ldots \cdot \omega_{K}\left(x_{K}\right) \cdot \rho_{1}\left(y_{1}\right) \cdot \ldots \cdot \rho_{K}\left(y_{K}\right)|z i p(\vec{x}, \vec{y})\rangle \\
& =\sum_{\vec{x} \in X^{K}, \vec{y} \in Y^{K}} \omega_{1}\left(x_{1}\right) \cdot \rho_{1}\left(y_{1}\right) \cdot \ldots \cdot \omega_{K}\left(x_{K}\right) \cdot \rho_{K}\left(y_{K}\right)|z i p(\vec{x}, \vec{y})\rangle \\
& =\sum_{\vec{x} \in X^{K}, \vec{y} \in Y^{K}}\left(\omega_{1} \otimes \rho_{1}\right)\left(x_{1}, y_{1}\right) \cdot \ldots \cdot\left(\omega_{K} \otimes \rho_{K}\right)\left(x_{K}, y_{K}\right)|z i p(\vec{x}, \vec{y})\rangle \\
& =\sum_{\vec{x} \in X^{K}, \vec{y} \in Y^{K}} \otimes\left(\omega_{1} \otimes \rho_{1}, \ldots, \omega_{K} \otimes \rho_{K}\right)\left(\left(x_{1}, y_{1}\right), \ldots,\left(x_{K}, y_{K}\right)\right)|z i p(\vec{x}, \vec{y})\rangle \\
& =\sum_{\vec{z} \in\left(X \times Y Y^{K}\right.} \otimes\left(\omega_{1} \otimes \rho_{1}, \ldots, \omega_{K} \otimes \rho_{K}\right)(\vec{z})|\vec{z}\rangle \\
& =\left(\otimes \circ \otimes^{K}\right)\left(\left(\omega_{1}, \rho_{1}\right), \ldots,\left(\omega_{K}, \rho_{K}\right)\right) \\
& =\left(\otimes \circ \otimes^{K} \circ z i p\right)(\vec{\omega}, \vec{\rho}) .
\end{aligned}
$$

3 This follows from the previous item and (2.17), 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 $\omega$ 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 2.2:

$$
\begin{equation*}
\operatorname{mix}(\omega, \vec{\rho}):=\omega(0) \cdot \rho_{0}+\cdots+\omega(n-1) \cdot \rho_{n-1}=\sum_{i \in \boldsymbol{n}} \omega(i) \cdot \rho_{i} . \tag{2.18}
\end{equation*}
$$

Using tensors we can describe this function mix as a composite of the following form:

where $\operatorname{proj}(i, \vec{x})=x_{i}$. Indeed:

$$
\begin{aligned}
(\text { proj } \circ \otimes \circ(\text { id } \times \otimes))(\omega, \vec{\rho}) & =\sum_{i \in \boldsymbol{n}, \vec{x} \in X^{n}}(\omega \otimes(\otimes \vec{\rho}))(i, \vec{x})|\operatorname{proj}(i, \vec{x})\rangle \\
& =\sum_{i \in \boldsymbol{n}, \vec{x} \in X^{n}} \omega(i) \cdot(\otimes \vec{\rho})(\vec{x})\left|x_{i}\right\rangle \\
& =\sum_{i \in \boldsymbol{n}, \vec{x} \in X^{n}} \omega(i) \cdot \rho_{0}\left(x_{0}\right) \cdot \ldots \cdot \rho_{n-1}\left(x_{n-1}\right)\left|x_{i}\right\rangle \\
& =\sum_{i \in \boldsymbol{n}, y \in X} \omega(i) \cdot \rho_{i}(y)|y\rangle \\
& =\sum_{i \in \boldsymbol{n}} \omega(i) \cdot \rho_{i} .
\end{aligned}
$$

Exercise 2.4.5 will describe this mixture in terms of state transformation $\gg=$.

### 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 $\operatorname{iid}[K](\omega)=\omega^{K}=\omega \otimes \cdots \otimes \omega \in \mathcal{D}\left(X^{K}\right)$ and then take image / pushforward distributions of the form $\mathcal{D}(f)\left(\omega^{K}\right)$, for two different functions $f: X^{K} \rightarrow Y$. The two functions $f$ that we use come from the triangular prism (1.48), 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 [83] and use the match
function mat: $X^{K} \rightarrow S P(K)$ from lists to set partitions (1.21) and the support fuction supp: $X^{K} \rightarrow \mathcal{P}_{s}[K](X)$ from lists to subsets (1.13). Explanations and illustrations will be provided below.

Definition 2.3.7. Let $\omega \in \mathcal{D}(X)$ be distribution and let $K \in \mathbb{N}$.
1 The coincidence distribution $c d$ is defined via pushforward along the match function as:

$$
\begin{equation*}
c d[K](\omega):=\mathcal{D}(\operatorname{mat})(\operatorname{iid}[K](\omega)) \in \mathcal{D}(S P(K)) \tag{2.19}
\end{equation*}
$$

2 The coupon distribution cpn arises via pushforward along the support function:

$$
\begin{equation*}
\operatorname{cpn}[K](\omega):=\mathcal{D}(\operatorname{supp})(\operatorname{iid}[K](\omega)) \in \mathcal{D}\left(\mathcal{P}_{\leq}[K](X)\right) \tag{2.20}
\end{equation*}
$$

We explain the latter coupon distribution first. We now think of the elements of the set $X$ as different 'coupons' 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: $X^{K} \rightarrow \mathcal{P}_{\leq}[K](X)$ to a product distribution $\omega^{K}$, see $(2.20)$. The recipe $(2.20)$ tells that for a subset $U$ we have to sum the probabilities of all list $\ell$ with $\operatorname{supp}(\ell)=U$.

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.

$$
\begin{array}{r}
\operatorname{cpn}[4](\rho)=\frac{1}{16}|\{a\}\rangle+\frac{1}{81}|\{b\}\rangle+\frac{11}{27}|\{a, b\}\rangle+\frac{1}{1296}|\{c\}\rangle \\
+\frac{29}{216}|\{a, c\}\rangle+\frac{4}{81}|\{b, c\}\rangle+\frac{1}{3}|\{a, b, c\}\rangle .
\end{array}
$$

We elaborate the probability $\frac{11}{27}$ of the subset $\{a, b\}$. Lemma 1.4 .9 tells that the number of lists $\ell$ of length 4 with support $\{a, b\}$ is $2 \cdot\left\{\begin{array}{l}4 \\ 2\end{array}\right\}=2 \cdot 7=14$. Below we describe the probabilities $\omega^{4}(\ell)$ for all these lists $\ell$ with $\operatorname{supp}(\ell)=\{a, b\}$.

$$
\begin{aligned}
& \omega^{4}(a, a, a, b)=\frac{1}{24} \\
& \omega^{4}(a, b, a, a)=\frac{1}{24} \\
& \omega^{4}(a, b, b, b)=\frac{1}{54} \\
& \omega^{4}(b, a, b, a)=\frac{1}{36} \\
& \omega^{4}(b, b, a, b)=\frac{1}{54}
\end{aligned}
$$

$$
\omega^{4}(a, a, b, a)=\frac{1}{24}
$$

$$
\omega^{4}(a, a, b, b)=\frac{1}{36}
$$

$$
\omega^{4}(a, b, a, b)=\frac{1}{36}
$$

$$
\omega^{4}(a, b, b, a)=\frac{1}{36}
$$

$$
\omega^{4}(b, a, a, a)=\frac{1}{24}
$$

$$
\omega^{4}(b, a, a, b)=\frac{1}{36}
$$

$$
\omega^{4}(b, a, b, b)=\frac{1}{54}
$$

$$
\omega^{4}(b, b, a, a)=\frac{1}{36}
$$

$$
\omega^{4}(b, b, b, a)=\frac{1}{54}
$$

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 $\operatorname{cpn}[4](\rho)$.

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 $v=\sum_{x \in X} \frac{1}{N}|x\rangle$. The resulting coupon probabilities are determined by the size of the subsets:

$$
\operatorname{cpn}[K](v)=\sum_{U \in \mathcal{P}[K](X)} \frac{|U|!}{N^{K}} \cdot\left\{\begin{array}{c}
K \\
|U|
\end{array}\right\}|U\rangle .
$$

This follows from Equation 1.14.
We next illustrate the coincidence distribution (2.19), obtained via a pushforward along the match function mat: $X^{K} \rightarrow S P(K)$ from (1.21). It assigns to each set partition $P \in S P(K)$ 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 $S P(3)$ of set partitions over $\{1,2,3\}$.

$$
\begin{gathered}
\operatorname{cd}[3](\rho)=\frac{1}{6}|\{\{1,2,3\}\}\rangle+\frac{2}{9}|\{\{1,2\},\{3\}\}\rangle+\frac{2}{9}|\{\{1,3\},\{2\}\}\rangle \\
+\frac{2}{9}|\{\{1\},\{2,3\}\}\rangle+\frac{1}{6}|\{\{1\},\{2\},\{3\}\}\rangle .
\end{gathered}
$$

We elaborate the probability $\frac{2}{9}$ of the set partition $P=\{\{1,2\},\{3\}\} \in S P(3)$. According to Lemma 1.5 .9 (2) there are $\frac{3!}{(3-2)!}=6$ lists $\ell$ of length 3 with $\operatorname{mat}(\ell)=P$. 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 $v$ is the uniform distribution on $X$, then via Equation (1.23) we get as coincidence distribution:

$$
c d[K](v)=\sum_{P \in S P(K)} \frac{N!}{(N-|P|)!\cdot N^{K}}|P\rangle \stackrel{1.23}{-} \sum_{P \in S P(K)} \frac{\left|m a t^{-1}(P)\right|}{N^{K}}|P\rangle .
$$

This is a proper probability distribution by Exercise 1.5 .7 For instance, for $N=4$ and $K=3$,

$$
\begin{gathered}
c d[3](v)=\frac{1}{16}|\{\{1,2,3\}\}\rangle+\frac{3}{16}|\{\{1,2\},\{3\}\}\rangle+\frac{3}{16}|\{\{1,3\},\{2\}\}\rangle \\
+\frac{3}{16}|\{\{1\},\{2,3\}\}\rangle+\frac{3}{8}|\{\{1\},\{2\},\{3\}\}\rangle .
\end{gathered}
$$

### 2.3.2 Marginalisation and copying

We introduce special, post-fix notation for marginalisation via 'masks'. It corresponds to the idea of listing only the relevant variables, where a distribution on a product set is often written as $\omega(x, y)$ and its first marginal as $\omega(x)$.

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 $\omega \in T\left(X_{1} \times \cdots \times X_{n}\right)$ 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 $\omega$ 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:

$$
\omega[1,0,1,0,1]=T\left(\left\langle\pi_{1}, \pi_{3}, \pi_{5}\right\rangle\right)(\omega) \in T\left(X_{1} \times X_{3} \times X_{5}\right)
$$

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 re-arrangement of components.
3 Let $\sigma \in T(X)$ and $\tau \in T(Y)$ be given. A coupling of $\sigma, \tau$ is a joint state $\omega \in T(X \times Y)$ that marginalises to both $\sigma$ and $\tau$, as in:

$$
\omega[1,0]=\sigma \quad \omega[0,1]=\tau
$$

In the probabilistic case there typically are infinitely many couplings, see Exercise 2.3.5

A product distribution of the form $\omega_{1} \otimes \omega_{2} \in \mathcal{D}\left(X_{1} \times X_{2}\right)$ is always nonentwined, see Lemma 2.3.2 (2). But in general, joint distributions are entwined, as will be illustrated below.

This 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 [28]. 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:

$$
W[1,0]=\mathcal{P}\left(\pi_{1}\right)(W)=\{u, v\}=X \quad W[0,1]=\mathcal{P}\left(\pi_{2}\right)(W)=\{a, b\}=A .
$$

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:

$$
\omega=\frac{1}{8}|u, a\rangle+\frac{1}{4}|u, b\rangle+\frac{3}{8}|v, a\rangle+\frac{1}{4}|v, b\rangle .
$$

We claim that $\omega$ is entwined. The first and second marginals $\omega[1,0]=$ $\mathcal{D}\left(\pi_{1}\right)(\omega) \in \mathcal{D}(X)$ and $\omega[0,1]=\mathcal{D}\left(\pi_{2}\right)(\omega) \in \mathcal{D}(A)$ are:

$$
\omega[1,0]=\frac{3}{8}|u\rangle+\frac{5}{8}|v\rangle \quad \text { and } \quad \omega[0,1]=\frac{1}{2}|a\rangle+\frac{1}{2}|b\rangle .
$$

The original state $\omega$ differs from the product of its marginals:

$$
\omega[1,0] \otimes \omega[0,1]=\frac{3}{16}|u, a\rangle+\frac{3}{16}|u, b\rangle+\frac{5}{16}|v, a\rangle+\frac{5}{16}|v, b\rangle .
$$

This entwinedness follows from a general characterisation, see Exercise 2.3.7 below.

We move from discarding (via marginalisation) to copying. Recall the copy function $\Delta: X \rightarrow X^{K}$ from Subsection 1.3.2 It is used in the following observation.

Fact 2.3.10. A pushforward along a copy differes from an iterated tensor. This will be illustrated for subsets and distributions. We consider the binary case, with copier $\Delta: X \rightarrow X \times X$, 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 $\omega \in \mathcal{D}(X)$,

$$
\mathcal{D}(\Delta)(\omega) \neq \omega \otimes \omega
$$

The following simple example illustrates this fact. First, for $\omega=\frac{1}{3}|0\rangle+$ $\frac{2}{3}|1\rangle$,

$$
\begin{equation*}
\mathcal{D}(\Delta)(\omega)=\frac{1}{3}|\Delta(0)\rangle+\frac{2}{3}|\Delta(1)\rangle=\frac{1}{3}|0,0\rangle+\frac{2}{3}|1,1\rangle . \tag{2.21}
\end{equation*}
$$

In contrast:

$$
\begin{equation*}
\omega \otimes \omega=\frac{1}{9}|0,0\rangle+\frac{2}{9}|0,1\rangle+\frac{2}{9}|1,0\rangle+\frac{4}{9}|1,1\rangle . \tag{2.22}
\end{equation*}
$$

In Exercise 2.3.9 we see that only point distributions are 'copyable'.

## 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 $\omega=\frac{1}{2}|a, b\rangle+\frac{1}{2}\left|a^{\perp}, b^{\perp}\right\rangle \in \mathcal{D}(A \times B)$ for $A=\left\{a, a^{\perp}\right\}$ and $B=\left\{b, b^{\perp}\right\}$. Check that both marginals $\omega[1,0] \in$ $\mathcal{D}(A)$ and $\omega[0,1] \in \mathcal{D}(B)$ have full support, but $\omega$ 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:

$$
\begin{aligned}
& \omega=\frac{1}{24}|a a a\rangle+\frac{1}{12}|a a b\rangle+\frac{1}{12}|a b a\rangle+\frac{1}{6}|a b b\rangle \\
&+\frac{1}{6}|b a a\rangle+\frac{1}{3}|b a b\rangle+\frac{1}{24}|b b a\rangle+\frac{1}{12}|b b b\rangle
\end{aligned}
$$

satisfies:

$$
\omega=\omega[1,1,0] \otimes \omega[0,0,1] .
$$

2.3.3 Let $X, Y$ be two non-empty finite sets. Show that:

$$
u n i f_{X} \otimes u n i f_{Y}=u n i f_{X \times Y}
$$

Conclude the uniform distribution on a product is non-entwined.
2.3.4 Find different joint states $\sigma, \tau$ 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).
2.3.5 1 For numbers $0 \leq s \leq r \leq 1$ form the joint distribution:

$$
\omega:=s|1,1\rangle+(r-s)|1,0\rangle+(1-r)|0,0\rangle \in \mathcal{D}(\mathbf{2} \times \mathbf{2}) .
$$

Show that $\omega$ is a coupling of $\operatorname{flip}(r)$ and flip(s), i.e. that: $\omega[1,0]=$ flip $(r)$ and $\omega[0,1]=$ flip $(s)$, see Definition 2.3.8, (3).

2 Let $s, t \in(0,1)$ be given. Show that there infinitely many couplings of flip( $s$ ) and flip $(t)$.
2.3.6 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?
2.3.7 Let $X=\{u, v\}$ and $A=\{a, b\}$ as in Example 2.3.9. Prove that a state $\omega=r_{1}|u, a\rangle+r_{2}|u, b\rangle+r_{3}|v, a\rangle+r_{4}|v, b\rangle \in \mathcal{D}(X \times A)$, where $r_{1}+$ $r_{2}+r_{3}+r_{4}=1$, is non-entwined if and only if $r_{1} \cdot r_{4}=r_{2} \cdot r_{3}$.
2.3.8 Show that tensoring of multisets is linear, in the sense that for $\varphi \in$ $\mathcal{M}(X)$ the 'tensor with $\varphi$ ' operation $\varphi \otimes(-): \mathcal{M}(Y) \rightarrow \mathcal{M}(X \times Y)$ 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:

$$
\varphi \otimes\left(r_{1} \cdot \psi_{1}+\cdots+r_{n} \cdot \psi_{n}\right)=r_{1} \cdot\left(\varphi \otimes \psi_{1}\right)+\cdots+r_{n} \cdot\left(\varphi \otimes \psi_{n}\right) .
$$

The same hold in the other coordinate, for $(-) \otimes \psi$. As a special case we obtain that when $\varphi$ is a probability distribution, then $\varphi \otimes(-)$ preserves convex sums of distributions.
2.3.9 For a state $\omega \in \mathcal{D}(X)$, show that the following statements are equivalent.

- $\omega$ is copyable, that is, $\omega \otimes \omega=\mathcal{D}(\Delta)(\omega)$;
- $\omega$ is a point distribution;
- $\omega$ is $\{0,1\}$-valued, i.e. restricts to a function $X \rightarrow\{0,1\}$.
2.3.10 Show that the following diagram commutes, in which flat is the flatten operation for lists, that removes inner brackets.

2.3.11 Show that the big tensor $\otimes: \mathcal{D}(X)^{K} \rightarrow \mathcal{D}\left(X^{K}\right)$ from 2.17) commutes with unit and flatten, as described below.



Abstractly this shows that the $K$-fold tensor functor $(-)^{K}$ distributes over the monad $\mathcal{D}$.

### 2.4 Probabilistic channels

In the previous chapter we have seen channels of the form $X \rightarrow \mathcal{L}(Y)$, or $X \rightarrow \mathcal{P}(Y)$, or $X \rightarrow \mathcal{M}(Y)$. We will now introduce probabilistic channels of the form $X \rightarrow \mathcal{D}(Y)$. Recall that we often use a special arrow $\rightarrow$ for such channels and then simply write them as $X \leadsto Y$.

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 Chan $(\mathcal{L}), \operatorname{Chan}(\mathcal{P})$ and $\operatorname{Chan}(\mathcal{M})$, for associated forms of computation. In this section we will 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 also organise probabilistic channels in the form of a category $\operatorname{Chan}(\mathcal{D})$, with sets as objects and probabilistic computations $X \rightarrow \mathcal{D}(Y)$ as morphisms $X \leadsto Y$. 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 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 : $X \rightarrow \mathcal{D}(X)$ is simply unit $(x):=$ $1|x\rangle$. Flattening is the function flat: $\mathcal{D}(\mathcal{D}(X)) \rightarrow \mathcal{D}(X)$ with:

$$
\operatorname{flat}\left(\sum_{i} r_{i}\left|\omega_{i}\right\rangle\right):=\sum_{i} r_{i} \cdot \omega_{i}=\sum_{x \in X}\left(\sum_{i} r_{i} \cdot \omega_{i}(x)\right)|x\rangle
$$

The formulation in the middle uses a convex sum of distributions 2.2.
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:

$$
\sum_{x}\left(\sum_{i} r_{i} \cdot \omega_{i}(x)\right)=\sum_{i} r_{i} \cdot \sum_{x} \omega_{i}(x)=\sum_{i} r_{i} \cdot 1=1
$$

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

In Lemma 2.1.3 we have seen pushforward of a distribution $\omega$ along a function $f$, producing an image distribution $\mathcal{D}(f)(\omega)$. There is also pushforward along a channel, which we describe next, with a special notation $\gg=$. It may be formulated in terms of multiplication of stochastic matrices, see Exercise 2.4.2.

## Definition 2.4.2.

1 A probabilistic channel $c: X \leadsto Y$ is a function $c: X \rightarrow \mathcal{D}(Y)$. 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 $c: X \leadsto Y$ we can do a pushforward, also known as state transformation or prediction, and produce a new distribution / state $c \gg=\omega \in \mathcal{D}(Y)$ on the channel's codomain $Y$. This happens via the standard definition for $>=$ from Section 1.10 , see especially $(1.50$ and (1.51):

$$
\begin{align*}
c \gg=\omega:=\operatorname{flat}(\mathcal{D}(c)(\omega)) & =\sum_{x \in X} \omega(x) \cdot c(x) \\
& =\sum_{y \in Y}\left(\sum_{x \in X} c(x)(y) \cdot \omega(x)\right)|y\rangle . \tag{2.23}
\end{align*}
$$

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.

We have already seen that unit and flatten restrict to distribution, so as a result the description 2.23) produces a new distribuition $c \gg=\omega$. But we can double-check that probabilities add up to one:

$$
\begin{aligned}
\sum_{y \in Y}(c \gg=\omega)(y) & =\sum_{y \in Y}\left(\sum_{x \in X} c(x)(y) \cdot \omega(x)\right) \\
& =\sum_{x \in X}\left(\sum_{y \in Y} c(x)(y)\right) \cdot \omega(x)=\sum_{x \in X} 1 \cdot \omega(x)=1 .
\end{aligned}
$$

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:

$$
\begin{align*}
& \mathrm{x} \leftarrow \omega \\
& \mathrm{y} \leftarrow c(\mathrm{x})  \tag{2.24}\\
& \text { return } \mathrm{y}
\end{align*}
$$

Later, in Subsection 3.7.1, we provide a mathematical justification for this notation.

Example 2.4.3 (Copied from [81]). Let's assume we wish to capture the mood of a teacher, as a probabilistic mixture of three possible options namely: pessimistic $(p)$, neutral ( $n$ ) or optimistic ( $o$ ). We thus have a three-element proba-
bility space $X=\{p, n, o\}$. We assume a mood distribution:

$$
\omega=\frac{1}{8}|p\rangle+\frac{3}{8}|n\rangle+\frac{1}{2}|o\rangle \quad \text { with plot }
$$



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 $c: X \leadsto Y$. It is defined via the following three mark distributions, one for each element in the set of moods $X=\{p, n, o\}$.

$$
\begin{aligned}
& c(p) \\
& =\frac{1}{50}|1\rangle+\frac{2}{50}|2\rangle+\frac{10}{50}|3\rangle+\frac{15}{50}|4\rangle+\frac{10}{50}|5\rangle \\
& \quad+\frac{6}{50}|6\rangle+\frac{3}{50}|7\rangle+\frac{1}{50}|8\rangle+\frac{1}{50}|9\rangle+\frac{1}{50}|10\rangle
\end{aligned}
$$



$$
\begin{aligned}
& c(n) \\
& \quad=\frac{1}{50}|1\rangle+\frac{2}{50}|2\rangle+\frac{4}{50}|3\rangle+\frac{10}{50}|4\rangle+\frac{15}{50}|5\rangle \\
& \quad+\frac{10}{50}|6\rangle+\frac{5}{50}|7\rangle+\frac{1}{50}|8\rangle+\frac{1}{50}|9\rangle+\frac{1}{50}|10\rangle
\end{aligned}
$$



$$
\begin{aligned}
& c(o) \\
& =\frac{1}{50}|1\rangle+\frac{1}{50}|2\rangle+\frac{1}{50}|3\rangle+\frac{2}{50}|4\rangle+\frac{4}{50}|5\rangle \\
& \quad+\frac{10}{50}|6\rangle+\frac{15}{50}|7\rangle+\frac{10}{50}|8\rangle+\frac{4}{50}|9\rangle+\frac{2}{50}|10\rangle .
\end{aligned}
$$



Now that the state $\omega \in \mathcal{D}(X)$ and the channel $c: X \leadsto Y$ have been described, we can form the transformed state $c \gg \omega \in \mathcal{D}(Y)$. Following the formulation (2.23) we get for each mark $i \in Y$ the predicted probability:

$$
\begin{aligned}
(c \gg=\omega)(i) & =\sum_{x} \omega(x) \cdot c(x)(i) \\
& =\omega(p) \cdot c(p)(i)+\omega(n) \cdot c(n)(i)+\omega(o) \cdot c(o)(i) \\
& =\frac{1}{8} \cdot c(p)(i)+\frac{3}{8} \cdot c(n)(i)+\frac{1}{2} \cdot c(o)(i) .
\end{aligned}
$$

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 $\omega$. This plot contains the 'predicted' marks, corresponding to the state of mind of the teacher.


We view channels $X \leadsto Y$ as probabilistic computations. Such computations can be composed, both sequentially, via $\odot$, and in parallel, via $\otimes$.

## Definition 2.4.4

1 Let $X \stackrel{c}{\rightarrow} Y$ and $Y \stackrel{d}{\rightarrow} Z$ be two channels. We can form the composite channel $d \odot c: X \leadsto Z$ via the monad structure.

$$
d \odot c:=(X \xrightarrow{c} \mathcal{D}(Y) \xrightarrow{\mathcal{D}(d)} \mathcal{D}(\mathcal{D}(Z)) \xrightarrow{\text { flat }} \mathcal{D}(Z)) .
$$

More explicitly, there is a formula, for $x \in X$,

$$
(d \odot c)(x)=d \gg c(x)=\sum_{z \in Z}\left(\sum_{y \in Y} c(x)(y) \cdot d(y)(z)\right)|z\rangle .
$$

This composition $\odot$ is associative and has unit channels unit: $X \leadsto X$ as identity maps, see Lemma 1.10 .3 . It gives us a category of probabilistic channels, written as Chan $(\mathcal{D})$, or simply as Chan.
2 Now let's assume that we have channels $X \stackrel{c}{\rightarrow} Y$ and $A \stackrel{d}{\hookrightarrow} B$. Then we can form the parallel product (or tensor) channel $c \otimes d: X \times A \nrightarrow Y \times B$ via:

$$
c \otimes d:=(X \times A \xrightarrow{c \times d} \mathcal{D}(Y) \times \mathcal{D}(B) \xrightarrow{\otimes} \mathcal{D}(Y \times B)) .
$$

This means for elements $x \in X$ and $a \in A$,

$$
(c \otimes d)(x, a)=c(x) \otimes d(a)=\sum_{y \in Y, b \in B} c(x)(y) \cdot d(a)(b)|y, b\rangle .
$$

3 When two channels $c: Z \leadsto X$ and $d: Z \rightsquigarrow Y$ have the same domain we can form a tuple channel $\langle c, d\rangle: Z \mapsto X \times Y$ as:

$$
\langle c, d\rangle:=(c \otimes d) \circ \Delta \quad \text { so that } \quad\langle c, d\rangle(z)=c(z) \otimes d(z) .
$$

4 In particular, for a distribution $\omega$ on $Z$ and a channel $d: Z \rightarrow Y$ we define the graph as the joint distribution on $Z \times Y$ defined by:

$$
\langle\mathrm{id}, d\rangle \gg=\omega \quad \text { so that } \quad(\langle\mathrm{id}, d\rangle \gg=\omega)(y, z)=\omega(y) \cdot d(y)(z) .
$$

5 Finally, for a channel $c: X \leadsto Y_{1} \times \cdots \times Y_{n}$ and a mask $M$ of length $n$ we write $c M$ 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 $=\operatorname{Chan}(\mathcal{D})$ of probabilistic channels has sets $X$ as objects and channels $X \leadsto Y$ as morphisms / maps. These morphisms can be composed sequentially via $\odot$ 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 $f: X \rightarrow Y$ can be turned into a (probabilistic) channel $\langle f\rangle: X \leadsto Y$. Explicitly, $\langle f\rangle(x)=1|f(x)\rangle$. This can be formalised in terms of a functor Sets $\rightarrow$ 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 channels - see Lemma 1.10.3, 4 .

$$
\begin{align*}
& \langle g\rangle \circ\langle f\rangle=\langle g \circ f\rangle \quad c \circ\langle f\rangle=c \circ f \quad\langle f\rangle \otimes\langle g\rangle=\langle f \times g\rangle  \tag{2.25}\\
& \langle f\rangle \gg \omega=\mathcal{D}(f)(\omega) \quad g \odot c=\mathcal{D}(g) \circ c \quad\langle\langle f\rangle,\langle g\rangle\rangle=\langle\langle f, g\rangle\rangle .
\end{align*}
$$

As a special case we get:

$$
\begin{equation*}
\text { unit } \otimes \text { unit }=\langle i d\rangle \otimes\langle\mathrm{id}\rangle=\langle\mathrm{id} \times \mathrm{id}\rangle=\langle\mathrm{id}\rangle=\text { unit. } \tag{2.26}
\end{equation*}
$$

We have already seen several examples of probabilistic channels. For instance, the flip and binomial distributions from Example 2.1.2 form channels:

$$
[0,1] \xrightarrow{\text { fip }}\{0,1\} \quad \text { and } \quad[0,1] \xrightarrow{\text { bn }[K]}\{0,1, \ldots, K\}
$$

Frequentist learning has been described as a function Flrn: $\mathcal{M}_{*}(X) \rightarrow \mathcal{D}(X)$ and thus forms a channel, as on the left below. A tensor also forms a channel, as in the middle. Even the identity function $\mathcal{D}(X) \rightarrow \mathcal{D}(X)$ can be described as a channel, which we call sample, abbreviated to sam.

$$
\mathcal{M}_{*}(X) \xrightarrow{\text { Flrn }} X \quad \mathcal{D}(X) \times \mathcal{D}(Y) \xrightarrow{\otimes \rightarrow \rightarrow} X \times Y \quad \mathcal{D}(X) \xrightarrow{\text { sam }} X
$$

The coincidence and coupon functions from Subsection 2.3.1 form channels of the form:



The formalism of channels, including their sequential and parallel composition
$\odot$ and $\otimes$, allows us to express some of the basic properties in probability theory. This is an underlying theme in this book. It forms the basis for the area of categorical probability theory.
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 2.4.5. Let $X$ be an arbitrary set and $K$ be a fixed positive natural number. We recall the accumulation function acc: $X^{K} \rightarrow \mathcal{N}[K](X)$ mapping lists to natural multisets, via $\operatorname{acc}\left(x_{1}, \ldots, x_{K}\right)=1\left|x_{1}\right\rangle+\cdots+1\left|x_{K}\right\rangle$, see (1.36).
In Proposition 1.7.2 we have seen that for a natural multiset $\varphi \in \mathcal{N}[K](X)$ there are $(\varphi)$ many lists $\vec{x} \in X^{K}$ that accumulate to $\varphi$, that is, with $\operatorname{acc}(\vec{x})=\varphi$. We are interested in an inverse to accumulation. Which of these $(\varphi)$ many lists in $\operatorname{acc}^{-1}(\varphi)$ do we pick? The trick is not to make a choice, but to take them all, in a uniform distribution. This gives a channel $\mathcal{N}[K](X) \rightarrow \mathcal{D}\left(X^{K}\right)$ in the other direction. We call it arrangement, abbreviated as arr.

$$
\begin{equation*}
\operatorname{arr}(\varphi):=\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{(\varphi)}|\vec{x}\rangle=\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{\prod_{x} \varphi(x)!}{K!}|\vec{x}\rangle \tag{2.27}
\end{equation*}
$$

For instance, for $X=\{a, b\}$ with multiset $\varphi=3|a\rangle+1|b\rangle$ there are $(\varphi)=$ $\binom{4}{3,1}=\frac{4!}{3!\cdot 1!}=4$ arrangements of $\varphi$, namely $[a, a, a, b],[a, a, b, a],[a, b, a, a]$, and $[b, a, a, a]$, so that:

$$
\operatorname{arr}(3|a\rangle+1|b\rangle)=\frac{1}{4}|a, a, a, b\rangle+\frac{1}{4}|a, a, b, a\rangle+\frac{1}{4}|a, b, a, a\rangle+\frac{1}{4}|b, a, a, a\rangle .
$$

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 acc: $X^{K} \rightarrow \mathcal{N}[K](X)$ into a (deterministic) channel $\langle\mathrm{acc}\rangle: X^{K} \rightarrow$ $\mathcal{D}(\mathcal{N}[K](X))$, where $\langle\operatorname{acc}\rangle(\vec{x})=1|\operatorname{acc}(\vec{x})\rangle$, 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.


We elaborate the details, for an arbitrary multiset $\varphi \in \mathcal{N}[K](X)$.

$$
\begin{aligned}
&(\langle\operatorname{acc}\rangle \odot \operatorname{arr})(\varphi) \stackrel{\sqrt{2.25}}{-} \mathcal{D}(\operatorname{acc})(\operatorname{arr}(\varphi)) \stackrel{\sqrt{2.27}}{-} \mathcal{D}(\operatorname{acc})\left(\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{(\varphi)}|\vec{x}\rangle\right) \\
&=\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{(\varphi)}|\operatorname{acc}(\vec{x})\rangle \\
&=\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{(\varphi)}|\varphi\rangle \\
&=1|\varphi\rangle \quad \text { by Proposition 1.7.2 } \\
&=\text { unit }(\varphi) .
\end{aligned}
$$

In the other direction, $\operatorname{arr}(\operatorname{acc}(\vec{x}))$ does not return $\vec{x}$. It yields a uniform distribution over all permutations of the sequence $\vec{x} \in X^{K}$, see Section 2.9 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 2.4.6. Let $h: X \rightarrow Y$ 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 $h^{\sim 1}: Y \rightarrow \mathcal{D}(X)$ via uniform distributions:

$$
\begin{equation*}
h^{\sim 1}(y):=\sum_{x \in h^{-1}(y)} \frac{1}{\left|h^{-1}(y)\right|}|x\rangle . \tag{2.29}
\end{equation*}
$$

Then one has: $\langle h\rangle \odot h^{\sim 1}=$ unit, precisely as in the previous example.
Thus, the arrangement map is defined in 2.27) as arr $=\mathrm{acc}^{\sim 1}$, 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 $c: X \rightarrow \mathcal{D}(Y)$ given by:

$$
c(x)=\frac{1}{2}|u\rangle+\frac{1}{2}|v\rangle \quad c(y)=1|u\rangle \quad c(z)=\frac{3}{4}|u\rangle+\frac{1}{4}|v\rangle .
$$

1 Compute the state transformation $c »=\omega \in \mathcal{D}(Y)$.

2 Consider a new channel $d:\{u, v\} \nrightarrow\{1,2,3,4\}$ given by:

$$
d(u)=\frac{1}{4}|1\rangle+\frac{1}{8}|2\rangle+\frac{1}{2}|3\rangle+\frac{1}{8}|4\rangle \quad d(v)=\frac{1}{4}|1\rangle+\frac{1}{8}|3\rangle+\frac{5}{8}|4\rangle .
$$

Describe the composite channel $d \odot c:\{a, b, c\} \leadsto\{1,2,3,4\}$ concretely.
2.4.2 Consider for the channel $c$ and the distribution $\omega$ in the previous exercise the associated matrices:

$$
M_{c}=\left(\begin{array}{ccc}
\frac{1}{2} & 1 & \frac{3}{4} \\
\frac{1}{2} & 0 & \frac{1}{4}
\end{array}\right) \quad M_{\omega}=\left(\begin{array}{c}
\frac{1}{6} \\
\frac{1}{2} \\
\frac{1}{3}
\end{array}\right)
$$

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 $c \gg=\omega$ 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 4.3.5])
2.4.3 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 unif $\mathcal{N}_{\mathcal{N}[K](X)} \in \mathcal{D}(\mathcal{N}[K](X))$ for the uniform distribution over such multisets.

1 Check that:

$$
\text { unif }_{\mathcal{N}[K](X)}=\sum_{\varphi \in \mathcal{N}[K](n)} \frac{1}{\left.\binom{n}{K}\right)}|\varphi\rangle .
$$

2 Use Lemma 1.8.8 to prove that Flrn 》= unif $\mathcal{N}_{\mathcal{N}[K](X)}=$ unif $_{X} \in \mathcal{D}(X)$.
2.4.4 Recall that the identity function $\mathcal{D}(X) \rightarrow \mathcal{D}(X)$ is used as sample channel sam: $\mathcal{D}(X) \mapsto X$. Check that for $\omega \in \mathcal{D}(X), c: X \leadsto Y$, and $\Omega \in \mathcal{D}(\mathcal{D}(X))$ one has

$$
(c \odot \operatorname{sam})(\omega)=c \gg=\omega \quad \text { and } \quad \operatorname{sam} \gg \Omega=\operatorname{flat}(\Omega) \text {. }
$$

2.4.5 Recall the mixture map mix: $\mathcal{D}(\boldsymbol{n}) \times \mathcal{D}(X)^{n} \rightarrow \mathcal{D}(X)$ from Example 2.3.6. Show that it is essentially state transformation:

$$
\operatorname{mix}(\omega, c)=c \gg=\omega,
$$

where a channel $c: n \mapsto X$ is identified with an $n$-tuple of distributions $c(0), \ldots, c(n-1) \in \mathcal{D}(X)$.
2.4.6 Let $f: X \rightarrow Y$ be a function and $c: Y \leadsto Z$ be a channel. Show that for a distribution $\omega \in \mathcal{D}(X)$ the following expressions are all the same.

$$
c \gg=\mathcal{D}(f)(\omega)=c \gg=(f \gg=\omega)=(c \odot f) \gg=\omega .
$$

We have left the brackets $\langle-\rangle$ that promote $f$ to a channel implicit. Fill them in yourself where appropriate.
2.4.7 For an arbitrary set $X$ and a number $K \in \mathbb{N}_{>0}$ define a 'uniform projection' channel unpr $[K]: X^{K} \leadsto X$ by:

$$
\operatorname{unpr}[K](\vec{x}):=\sum_{1 \leq i \leq K} \frac{1}{K}\left|x_{i}\right\rangle .
$$

This channel is written as $\varepsilon[K]$ in [82]. It can be understood as frequentist learning for sequences, see the triangles below.

1 Prove that unpr $[K]$ is natural in $X$.
2 Check that the following diagram commutes, for $K, L \geq 1$.


3 Show that the following two triangles (of channels) commute.


4 Let $\omega_{1}, \ldots, \omega_{K} \in \mathcal{D}(X)$ be given. Show that:

$$
\begin{aligned}
\text { unpr }[K] »=\left(\omega_{1} \otimes \cdots \otimes \omega_{K}\right) & =\frac{1}{K} \cdot \omega_{1}+\cdots+\frac{1}{K} \cdot \omega_{K} \\
& =\text { flat }\left(\operatorname{unpr}[K]\left(\omega_{1}, \ldots, \omega_{K}\right)\right) .
\end{aligned}
$$

This means that the following rectangle commute, via Exercise 2.4.4.


Deduce that unpr $[K] \gg=\omega^{K}=\omega$.
2.4.8 1 Check that the following two diagrams commute.



Express these equations also in string diagrammatic form.
2 Show that:

$$
(c \otimes d) \gg=(\omega \otimes \rho)=(c \gg=\omega) \otimes(d \gg=\rho) .
$$

3 Use the previous equation to show:

$$
(e \otimes f) \odot(c \otimes d)=(e \odot c) \otimes(f \odot d) .
$$

Check that this equation trivialises when written in diagrammatic form.
4 Prove that for a channel $c: X \leadsto Y$ the following diagram commutes.


5 Show that the following diagram commutes.

2.4.9 Recall that we write sam: $\mathcal{D}(X) \leadsto X$ for the identity function, as channel. Prove that one can write the iid and $\otimes$ channels from Definition 2.3.4 as:

$$
\text { iid }=\langle\operatorname{sam}, \ldots, \text { sam }\rangle \quad \text { and } \quad \otimes=\operatorname{sam} \otimes \cdots \otimes \text { sam. }
$$

2.4.10 Prove that state transformation along $\mathcal{D}$-channels preserves convex combinations (2.2), that is, for $r \in[0,1]$,

$$
c \gg=(r \cdot \omega+(1-r) \cdot \rho)=r \cdot(c \gg=\omega)+(1-r) \cdot(c \gg=\rho) .
$$

2.4.11 1 Prove that the arrangement map $\operatorname{arr}=\operatorname{acc}^{\sim 1}$ from Example 2.4.5 is
natural: for each function $f: X \rightarrow Y$ the following diagram commutes.

(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: $\mathcal{N}(X) \mapsto$ $\mathcal{L}(X)$ and that it is also natural in this form, without explicit size $K$.
2.4.12 Recall the match function mat: $X^{K} \rightarrow S P(K)$ from 1.21. We assume $K \leq|X|$. Use $(1.23)$ to check that the probabilistic inverse channel mat ${ }^{\sim 1}: S P(K) \rightsquigarrow X^{K}$ can be described as:

$$
\operatorname{mat}^{\sim 1}(P)=\sum_{\vec{x} \in \operatorname{mat}^{-1}(P)} \frac{(|X|-|P|)!}{|X|!}|\vec{x}\rangle
$$

Describe this channel for $K=3$ and $X=\{a, b, c, d\}$.
2.4.13 Use Lemma 1.4.9 to define a probabilistic inverse for the support function $\operatorname{supp}[K]: X^{K} \rightarrow \mathcal{P}_{\leq}[K](X)$, where $K \geq 1$.
2.4.14 This exercise is meant to show that probabilistic inverses do not compose. Consider the following commutative triangle of surjective functions with finite inverse images:

$$
\left\{\begin{array}{l}
\text { \{1, } 2,3,4\} \longrightarrow\{u, v, w\} \\
\text { where } \quad \begin{array}{l}
h(1)=h(2)=u \\
h(3)=v \\
h(4)=w \\
g(u)=g(v)=a \\
g(v)=b .
\end{array} \\
\{a, b\}
\end{array}\right.
$$

Check the inequality $\neq$ in:

$$
h^{\sim 1} \odot g^{\sim 1} \neq f^{\sim 1}=(g \circ h)^{\sim 1} .
$$

2.4.15 Recall the bijective correspondences from Exercise 1.10.4.

1 Let $X, Y$ be finite sets and $c: X \rightarrow \mathcal{D}(Y)$ be a $\mathcal{D}$-channel. We can then define an $\mathcal{M}$-channel $c^{\dagger}: Y \rightarrow \mathcal{M}(X)$ 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 $A=\left\{a_{0}, a_{1}\right\}$ and $B=\left\{b_{0}, b_{1}\right\}$ and define a channel bell: $A \times$ $2 \rightarrow \mathcal{D}(B \times 2)$ as:

$$
\begin{array}{r}
\operatorname{bell}\left(a_{0}, 0\right)=\frac{1}{2}\left|b_{0}, 0\right\rangle+\quad \frac{3}{8}\left|b_{1}, 0\right\rangle+\frac{1}{8}\left|b_{1}, 1\right\rangle \\
\operatorname{bell}\left(a_{0}, 1\right)=r \\
\operatorname{bell}\left(a_{1}, 0\right)=\frac{3}{8}\left|b_{0}, 0\right\rangle+\frac{1}{8}\left|b_{0}, 1\right\rangle+\frac{1}{8}\left|b_{0}, 1\right\rangle+\frac{1}{8}\left|b_{1}, 0\right\rangle+\frac{3}{8}\left|b_{1}, 1\right\rangle \\
\operatorname{bell}\left(a_{1}, 1\right)=\frac{3}{8}\left|b_{1}, 1\right\rangle
\end{array}
$$

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 $\operatorname{bell}{ }^{\dagger}(y, z)[1,0]$ are uniform.
2.4.16 Check that the inclusions $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ form a map of monads, as described in Definition 1.11.2
2.4.17 Let $c: X \rightarrow \mathcal{D}(Y)$ be a $\mathcal{D}$-channel and $\varphi \in \mathcal{M}(X)$ be a multiset. Because $\mathcal{D}(Y) \subseteq \mathcal{M}(Y)$ we can also consider $c$ as an $\mathcal{M}$-channel, and use $c \gg=\varphi$. Prove that:

$$
\operatorname{Flrn}(c \gg=\varphi)=c »>=\operatorname{lrn}(\varphi)=F \operatorname{lrn}(c \gg=\operatorname{Flrn}(\varphi)) .
$$

2.4.18 Consider a state $\sigma \in \mathcal{D}(X)$ and an 'endo' channel $c: X \rightarrow X$.

1 Check that the following two statements are equivalent.

- $\langle\mathrm{id}, c\rangle \gg=\sigma=\langle c$, id $\rangle \gg=\sigma$ 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 $\sigma$ is a fixed point for state transformation, that is:

$$
c \gg=\sigma=\sigma .
$$

Such a $\sigma$ is also called a stationary state.

### 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 2.28 For channels we will use a second graphical formalism, called string diagrams. 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. We introduce them via an example from the literature and show that they are also formalisations of channels. However, Bayesian networks do not reflect sequential and parallel composition of channels. Hence we prefer to work with string diagrams.

### 2.5.1 A graphical language for channels

At this stage we introduce these string diagrams informally. How they are used will become clear along the way. For more information we refer to the literature, see e.g. [163]. For usage of string diagram in probability theory, see [54, 57], and in quantum theory, see [28].

In this language of string diagrams a channel $c: X \leadsto Y$ 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 $d \odot c$ 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 $\bar{\mp}$. 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 finite set $X$ there is a uniform distribution $u n i f_{X} \in \mathcal{D}(X)$, for which
we use special notation, on the left below, as upside-down of discarding.

$$
\text { unif } \left._{X}=X|=\quad 1| y\right\rangle=\begin{array}{r}
Y  \tag{2.30}\\
y
\end{array}
$$

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:


The dashed box on the right indicates 'nothing'.
A copy is written as a splitting of wires: $\zeta$. It satisfies a number of diagrammatic equations, expressing associativity, commutativity and a co-unit property:


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:



$$
\begin{equation*}
X \times Y \overline{\overline{\bar{T}}}=\overline{\overline{X^{X} \mid Y}}=\overline{\overline{\eta_{X}}} \overline{\overline{\rceil_{Y}}} \tag{2.32}
\end{equation*}
$$

We illustrate the use of string diagrams in the following result 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 equation below expresses that the bit-sum function is a sufficient statistic, see Section 7.6

Theorem 2.5.1. Consider the $K$-fold bit-sum function sum : $\mathbf{2}^{K} \rightarrow\{0, \ldots, K\}$ given by $\operatorname{sum}\left(b_{1}, \ldots, b_{K}\right)=\sum_{i} b_{i}$.

1 The probabilistic inverse sum $^{\sim 1}:\{0, \ldots, K\} \rightsquigarrow \mathbf{2}^{K}$ is given by:

$$
\operatorname{sum}^{\sim 1}(n)=\sum_{\vec{b} \in \operatorname{sum}^{-1}(n)} \frac{1}{\binom{K}{n}}|\vec{b}\rangle .
$$

2 There is the following equality of channels $[0,1] \leadsto\{0, \ldots, K\} \times \mathbf{2}^{K}$.


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 $\operatorname{sum}^{\sim 1}(n)$ each of these ways gets probability ${ }^{1 /( }\left({ }_{n}^{K}\right)$.
2 We have for $r \in[0,1]$,

$$
\begin{aligned}
& \langle\operatorname{sum}, \text { id }\rangle \gg=\langle\text { flip }, \ldots, \text { flip }\rangle(r) \\
& =\sum_{\vec{b} \in 2^{K}} \sum_{n \in\{0, \ldots, K\}}\langle\operatorname{sum}\rangle(\vec{b})(n) \cdot \operatorname{flip}(r)\left(b_{1}\right) \cdot \ldots \cdot \operatorname{flip}(r)\left(b_{K}\right)|n, \vec{b}\rangle \\
& =\sum_{n \in\{0, \ldots, K\}} \sum_{\vec{b} \in s u u^{-1}(n)} r^{b_{1}} \cdot(1-r)^{1-b_{1}} \cdot \ldots \cdot r^{b_{K}} \cdot(1-r)^{1-b_{K}}|n, \vec{b}\rangle \\
& =\sum_{n \in\{0, \ldots, K\}} \sum_{\vec{b} \in \operatorname{sum}^{-1}(n)} r^{\sum_{i} b_{i}} \cdot(1-r)^{K-\sum_{i} b_{i}}|n, \vec{b}\rangle \\
& =\sum_{n \in\{0, \ldots, K\}} \sum_{\vec{b} \in \operatorname{sum}^{-1}(n)} \frac{1}{\binom{K}{n}} \cdot\binom{K}{n} \cdot r^{n} \cdot(1-r)^{K-n}|n, \vec{b}\rangle \\
& =\sum_{\vec{b} \in 2^{K}} \sum_{n \in\{0, \ldots, K\}} \operatorname{sum}^{\sim}(n)(\vec{b}) \cdot \operatorname{bn}[K](r)(n)|n, \vec{b}\rangle \\
& =\left\langle i d, \operatorname{sum}^{\sim}\right\rangle \gg \operatorname{bn}[K](r) .
\end{aligned}
$$

Concentrating only on the left, or only on the right, side of Diagram 2.33 gives interesting equations.

Corollary 2.5.2. There are the following two equations of channels $[0,1] \rightarrow$
$\{0, \ldots, K\}$ and $[0,1] \rightarrow \mathbf{2}^{K}$.


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 (2.31), and the rule for discarding of boxes:


Alternatively, the corresponding equational proof uses Exercise 2.4.8

$$
\begin{aligned}
\text { sum } \odot\langle\text { flip, }, \ldots, \text { flip }\rangle & =\pi_{1} \odot\langle\text { sum, id }\rangle \odot\langle\text { flip, } \ldots, \text { flip }\rangle \\
& \frac{2.33]}{-} \pi_{1} \odot\left\langle i d, \text { sum }^{\sim}\right\rangle \odot \text { bn }[K]=\text { bn }[K] .
\end{aligned}
$$

Remark 2.5.3. 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 as characteristic of such deterministic channels, see Exercise 2.5.2.

### 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


Figure 2.3 The wetness Bayesian network (from [37 Chap. 6]), with only the nodes and edges between them; the conditional probability tables associated with the nodes are given separately in the text.
observations, the network may give a doctor (updated) information about the likelihoods of various scenarios, see e.g. [51]. 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.

The particular Bayesian network that we use as illustration is a standard one, copied from the literature, namely from [37]. Bayesian networks were introduced in [145], see also [120, 9, 14, 15, 102, 115]. They form a popular technique for displaying probabilistic connections and for efficiently presenting joint states, without state explosion.

Consider the diagram / network in Figure 2.3. 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's 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 [37]. Here they are not used as sets of Booleans, with inhabitants true and false, but instead we use these sets with elements:

$$
A=\left\{a, a^{\perp}\right\} \quad B=\left\{b, b^{\perp}\right\} \quad C=\left\{c, c^{\perp}\right\} \quad D=\left\{d, d^{\perp}\right\} \quad E=\left\{e, e^{\perp}\right\}
$$

The notation $a^{\perp}$ is read as 'not $a^{\prime}$. In this way the name of an elements suggests to which set the element belongs.

The diagram in Figure 2.3 becomes a Bayesian network when we provide it with conditional probability tables. For the lower three nodes they look as follows.

馬䔍 | $a$ | $a^{\perp}$ |
| :---: | :---: |
| $3 / 5$ | $2 / 5$ |

$$
\begin{array}{l|ccc}
\dot{t} & A & b & b^{\perp} \\
\hline \hline a & 1 / 5 & 4 / 5 \\
\hline a_{0}^{2} & 3 / 4 & 1 / 4
\end{array}
$$

$$
\sqrt{A} \begin{array}{ccc}
A & c & c^{\perp} \\
\hline \hline a & 4 / 5 & 1 / 5 \\
\hline a^{\perp} & 1 / 10 & 9 / 10
\end{array}
$$

And for the upper two nodes we have:

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 $A=\left\{a, a^{\perp}\right\}$, which, in ket notation is:

$$
w i=\frac{3}{5}|a\rangle+\frac{2}{5}\left|a^{\perp}\right\rangle \in \mathcal{D}(A) .
$$

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, namely a channel $A \rightarrow \mathcal{D}(B)$. This is a crucial insight! We abbreviate this channel as $s p$, and define it explicitly as:

$$
A \xrightarrow{s p} B \quad \text { with } \quad\left\{\begin{aligned}
s p(a) & =\frac{1}{5}|b\rangle+\frac{4}{5}\left|b^{\perp}\right\rangle \\
s p\left(a^{\perp}\right) & =\frac{3}{4}|b\rangle+\frac{1}{4}\left|b^{\perp}\right\rangle .
\end{aligned}\right.
$$

We read this channel as: if it's winter, there is a $20 \%$ chance that the sprinkler is on, but if it's not winter, there is $75 \%$ chance that the sprinkler is on.

Similarly, the 'rain' table corresponds to a channel:

$$
A \xrightarrow{\mathrm{ra}} C \quad \text { with } \quad\left\{\begin{aligned}
\mathrm{ra}(a) & =\frac{4}{5}|c\rangle+\frac{1}{5}\left|c^{\perp}\right\rangle \\
\mathrm{ra}\left(a^{\perp}\right) & =\frac{1}{10}|c\rangle+\frac{9}{10}\left|c^{\perp}\right\rangle .
\end{aligned}\right.
$$

Before continuing we can see that the formalisation (partial, so far) of the wetness Bayesian network in Figure 2.3 in terms of states and channels, already allows us to do something meaningful, namely state transformation $\gg$. Indeed, we can form distributions:

$$
s p \gg=w i \text { on } B \quad \text { and } \quad r a \gg=w i \text { on } C .
$$

These (transformed) distributions capture the derived, predicted probabilities that the sprinkler is on, and that it rains. Using the definition of state transformation, see 2.23, we get:

$$
\begin{aligned}
(s p \gg=w i)(b) & =\sum_{x} s p(x)(b) \cdot w i(x) \\
& =s p(a)(b) \cdot w i(a)+\operatorname{sp}\left(a^{\perp}\right)(b) \cdot w i\left(a^{\perp}\right) \\
& =\frac{1}{5} \cdot \frac{3}{5}+\frac{3}{4} \cdot \frac{2}{5}=\frac{21}{50} \\
(s p \gg w i)\left(b^{\perp}\right) & =\sum_{x} s p(x)\left(b^{\perp}\right) \cdot w i(x) \\
& =s p(a)\left(b^{\perp}\right) \cdot w i(a)+\operatorname{sp}\left(a^{\perp}\right)\left(b^{\perp}\right) \cdot w i\left(a^{\perp}\right) \\
& =\frac{4}{5} \cdot \frac{3}{5}+\frac{1}{4} \cdot \frac{2}{5}=\frac{29}{50} .
\end{aligned}
$$

Thus the overall distribution for the sprinkler (being on or not) is:

$$
s p \gg=w i=\frac{21}{50}|b\rangle+\frac{29}{50}\left|b^{\perp}\right\rangle .
$$

In a similar way one can compute the probability distribution for rain as:

$$
r a \gg=w i=\frac{13}{25}|c\rangle+\frac{12}{25}\left|c^{\perp}\right\rangle
$$

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 2.3 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:

$$
B \times C \longrightarrow \stackrel{\text { wg }}{\longrightarrow} D \quad \text { with } \quad\left\{\begin{array}{r}
w g(b, c)=\frac{19}{20}|d\rangle+\frac{1}{20}\left|d^{\perp}\right\rangle \\
w g\left(b, c^{\perp}\right)=\frac{9}{10}|d\rangle+\frac{1}{10}\left|d^{\perp}\right\rangle \\
w g\left(b^{\perp}, c\right)=\frac{4}{5}|d\rangle+\frac{1}{5}\left|d^{\perp}\right\rangle \\
w g\left(b^{\perp}, c^{\perp}\right)=1\left|d^{\perp}\right\rangle .
\end{array}\right.
$$

Finally, the table for the 'slippery road' table gives:

$$
C \xrightarrow[0]{\text { sr }} E \quad \text { with } \quad\left\{\begin{aligned}
s r(c) & =\frac{7}{10}|e\rangle+\frac{3}{10}\left|e^{\perp}\right\rangle \\
\operatorname{sr}\left(c^{\perp}\right) & =1\left|e^{\perp}\right\rangle .
\end{aligned}\right.
$$

We illustrate how to obtain predictions for 'rain' and for 'slippery road'. We start from the latter. Looking at the network in Figure 2.3 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:

$$
\begin{aligned}
(s r \odot r a) \gg=w i & =s r \gg=(r a \gg w i) \\
& =s r \gg=\left(\frac{13}{25}|c\rangle+\frac{12}{25}\left|c^{\perp}\right\rangle\right)=\frac{91}{250}|e\rangle+\frac{159}{250}\left|e^{\perp}\right\rangle .
\end{aligned}
$$

The first equation follows from Lemma 1.10.3 (3). The second one involves elementary calculations, where we can use the distribution ra $\gg=$ wi that we calculated earlier.

Getting the predicted wet grass probability requires some care. Inspection of the network in Figure 2.3 is of some help, but leads to some 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 $w g$, as in:

$$
w g \gg=((s p \gg=w i) \otimes(r a \gg=w i)) .
$$

But this is wrong, since the winter probabilities are now not used consistently, see the different outcomes in the calculations 2.21) and 2.22. The correct way to obtain the wet grass prediction involves copying the winter state, via the copy channel $\Delta$, see:

$$
\begin{aligned}
(w g \odot(s p \otimes r a) \odot \Delta) \gg=w i & =w g \gg=((s p \otimes r a) \gg=(\Delta \gg=w i)) \\
& =w g \gg=(\langle s p, r a\rangle \gg=w i) \\
& =\frac{1399}{2000}|d\rangle+\frac{601}{2000}\left|d^{\perp}\right\rangle .
\end{aligned}
$$

Such calcutions are laborious, but essentially straightforward. We shall do this in one 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.

$$
\begin{aligned}
(\langle s p, r a\rangle \gg=w i)(b, c) & =\sum_{x} s p(x)(b) \cdot r a(x)(c) \cdot w i(x) \\
& =\operatorname{sp}(a)(b) \cdot r a(a)(c) \cdot \frac{3}{5}+\operatorname{sp}\left(a^{\perp}\right)(b) \cdot r a\left(a^{\perp}\right)(c) \cdot \frac{2}{5} \\
& =\frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5}+\frac{3}{4} \cdot \frac{1}{10} \cdot \frac{2}{5}=\frac{63}{500} \\
(\langle s p, r a\rangle \gg w i)\left(b, c^{\perp}\right) & =\cdots=\frac{147}{500} \\
(\langle s p, r a\rangle \gg w i)\left(b^{\perp}, c\right) & =\frac{197}{500} \\
(\langle s p, r a\rangle \gg w i)\left(b^{\perp}, c^{\perp}\right) & =\frac{93}{500} .
\end{aligned}
$$

We conclude that:

$$
\begin{aligned}
(s p \otimes r a) \gg=(\Delta \gg=w i) & =\langle s p, r a\rangle \gg=w i \\
& =\frac{63}{500}|b, c\rangle+\frac{147}{500}\left|b, c^{\perp}\right\rangle+\frac{197}{500}\left|b^{\perp}, c\right\rangle+\frac{93}{500}\left|b^{\perp}, c^{\perp}\right\rangle
\end{aligned}
$$

This distribution is used in the next step:

$$
\begin{aligned}
(w g \gg(\langle s p, r a\rangle \gg=w i))(d)= & \sum_{x, y} w g(x, y)(d) \cdot((s p \otimes r a) \gg=(\Delta \gg=w i))(x, y) \\
= & w g(b, c)(d) \cdot \frac{63}{500}+w g\left(b, c^{\perp}\right)(d) \cdot \frac{147}{500} \\
& \quad+w g\left(b^{\perp}, c\right)(d) \cdot \frac{197}{500}+w g\left(b^{\perp}, c^{\perp}\right)(d) \cdot \frac{93}{500} \\
= & \frac{19}{20} \cdot \frac{63}{500}+\frac{9}{10} \cdot \frac{147}{500}+\frac{4}{5} \cdot \frac{197}{500}+0 \cdot \frac{93}{500} \\
= & \frac{1399}{2000} \\
(w g \gg(\langle s p, r a\rangle \gg=w i))\left(d^{\perp}\right)= & \frac{601}{2000} .
\end{aligned}
$$

We have thus shown that the predicted wet grass probability is around $70 \%$ :

$$
w g \gg=(\langle s p, r a\rangle \gg w i)=\frac{1399}{2000}|d\rangle+\frac{601}{2000}\left|d^{\perp}\right\rangle \approx 0.7|d\rangle+0.3\left|d^{\perp}\right\rangle
$$

### 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 ( $\odot$ 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 2.3, but via string diagrams. The switch involves several changes.

- Copying is written explicity in string diagrams, as $\gamma$ in the binary case; in general one can have $n$-ary copying, for $n \geq 2$;
- Wires between nodes in Figure 2.3 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 2.3 according to these points yields its representation as string diagram in Figure 2.4. In this way the nodes (with their conditional probability tables) are clearly recognisable as channels,


Figure 2.4 The wetness Bayesian network from Figure 2.3 redrawn as a string diagram, reflecting the underlying channel-based semantics, via explicit copiers and typed wires.
in general of type $A_{1} \times \cdots \times A_{n} \leadsto B$, 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 $\mathbf{1} \leadsto B$, where $\mathbf{1}=\{0\}$ is the empty product. As we have seen, such channels $\mathbf{1} \leadsto B$ 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 $(w g \odot(s p \otimes r a) \odot \Delta) \gg w i$. We may even replace the symbol $\gg$ in the latter expression by $\odot$, when we see the winter distribution also as a channel, of type wi: $\mathbf{1} \leadsto A$. This tightens the match between the string diagram and the computation.

## Exercises

2.5.1 Consider the probabilistic channel $f: X \leadsto Y$ from Exercise 2.4.1 and show that on the one hand $\Delta \odot f: X \leadsto Y \times Y$ is given by:

$$
\begin{aligned}
x & \longmapsto \frac{1}{2}|u, u\rangle+\frac{1}{2}|v, v\rangle \\
y & \longmapsto 1|u, u\rangle \\
z & \longmapsto \frac{3}{4}|u, u\rangle+\frac{1}{4}|v, v\rangle .
\end{aligned}
$$

On the other hand, $(f \otimes f) \odot \Delta: X \rightarrow Y \times Y$ is described by:

$$
\begin{aligned}
x & \longmapsto \frac{1}{4}|u, u\rangle+\frac{1}{4}|u, v\rangle+\frac{1}{4}|v, u\rangle+\frac{1}{4}|v, v\rangle \\
y & \longmapsto 1|u, u\rangle \\
z & \longmapsto \frac{9}{16}|u, u\rangle+\frac{3}{16}|u, v\rangle+\frac{3}{16}|v, u\rangle+\frac{1}{16}|v, v\rangle .
\end{aligned}
$$

This demonstrates that channels do not commute with copiers, as claimed in Remark 2.5.3.
2.5.2 Deterministic channels are in fact the only ones that commute with copy (see for an abstract setting [20]). Let $c: X \leadsto Y$ commute with diagonals, in the sense that $\Delta \odot c=(c \otimes c) \odot \Delta$. Use Exercise 2.3.9 to prove that $c$ commutes with diagonals if and only if $c$ is deterministic, i.e. of the form $c=\langle f\rangle=u n i t \circ f$ for a function $f: X \leadsto Y$.
2.5.3 Show that the tuple operation $\langle-,-\rangle$ for channels, described in Definition 2.4.4 (4), satisfies both:

$$
\pi_{i} \odot\left\langle c_{1}, c_{2}\right\rangle=c_{i} \quad \text { and } \quad\left\langle\pi_{1}, \pi_{2}\right\rangle=\text { unit }
$$

but, essentially as in Exercise 2.5.1.

$$
\left\langle c_{1}, c_{2}\right\rangle \odot d \neq\left\langle c_{1} \odot d, c_{2} \odot d\right\rangle .
$$

2.5.4 Prove that, for finite sets $X, Y$, and for elements $u, v$,
$1 \quad X \times Y \mid=\underline{\underline{|X|}}=\underline{\underline{|X| Y}}=\underline{\underline{=}}$

$3 \begin{aligned} & \overline{\text { ㄱ}} \\ & 1\end{aligned}{ }^{-\cdots}=\underline{\underline{1}}$
2.5.5 1 Describe the channel in Theorem 2.5.1 (1) in detail for $K=3$.

2 Add types to all wires in the string diagrams in Theorem 2.5.1 (2].
2.5.6 In [37, §6.2] the (predicted) joint distribution on $D \times E$ that arises from the Bayesian network example Subsection 2.5.2 is reprented as a table. It translates into a joint distribution:

$$
\frac{30,443}{100,000}|d, e\rangle+\frac{39,507}{100,000}\left|d, e^{\perp}\right\rangle+\frac{5,957}{100,000}\left|d^{\perp}, e\right\rangle+\frac{24,093}{100,000}\left|d^{\perp}, e^{\perp}\right\rangle .
$$

Following the structure of the diagram in Figure 2.4, it is obtained in the present setting as:

$$
\begin{aligned}
& ((w g \otimes s r) \odot(i d \otimes \Delta) \odot(s p \otimes r a) \odot \Delta) \gg=w i \\
& \quad=(w g \otimes s r) \gg=((i d \otimes \Delta) \gg(\langle s p, r a\rangle \gg=w i)) .
\end{aligned}
$$

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 [37].)

### 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 these 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 $\mathcal{D}(\mathcal{N}[5](\{R, G, B\}))$.
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 returend to the urn together with an additional ball of the same colour.

These three modes 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 $v$ - the Greek letter upsilon - from the set $\mathcal{N}(X)$. The probability of drawing a ball with colour $x$ from an urn $v \in \mathcal{N}(X)$ is given by the fraction $\frac{v(x)}{\|v\|}=\operatorname{Flrn}(v)(x) \in[0,1]$ of $x$-coloured balls in the urn $v$. When a ball with colour $x$ is removed or added to an urn $v \in \mathcal{N}(X)$, we describe the resulting urn as $v-1|x\rangle$ or as $v+1|x\rangle$.

### 2.6.1 Hypergeometric draw-and-delete draws

Let us take, for instance, the multiset $v=4|R\rangle+6|B\rangle+2|G\rangle$ 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 $v$ is $\operatorname{Flrn}(v)(R)=\frac{v(R)}{\|v\|}=$ $\frac{4}{12}=\frac{1}{3}$. The resulting urn is $v_{1}=v-1|R\rangle=3|R\rangle+6|B\rangle+2|G\rangle$.
2 The probability of drawing the second red ball, now from $v_{1}=v-1|R\rangle$ is given by the same formula: $\operatorname{Flrn}\left(v_{1}\right)(R)=\frac{v_{1}(R)}{\left\|v_{1}\right\|}=\frac{3}{11}$. This results in an urn $v_{2}=v_{1}-1|R\rangle=2|R\rangle+6|B\rangle+2|G\rangle$.
3 The probability of the third red ball is now $\operatorname{Flrn}\left(v_{2}\right)(R)=\frac{v_{2}(R)}{\left\|v_{2}\right\|}=\frac{2}{10}=\frac{1}{5}$.
The probability that we thus assign to the draw $3|R\rangle$ from $v$ 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. The resulting probability for the draw $2|R\rangle+$ $1|B\rangle$ 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 $v \in \mathcal{N}(X)$ be an urn of size $L=\|v\|$, filled with $X$-coloured balls. Let $K \in \mathbb{N}$ describe the size of the draws from $v$, where we assume $K \leq L=\|v\|$. This guarantees that the urn contains sufficiently many balls and that we do not 'overdraw'.

The hypergeometric distribution $h g[K](v)$ over $\mathcal{N}[K](X)$ assigns probabilities to $K$-sized draws $\varphi \leq_{K} v$ via the formula:

$$
\begin{equation*}
\operatorname{hg}[K](v):=\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{L}{K}}|\varphi\rangle=\sum_{\varphi \leq K^{v}} \frac{\prod_{x}\binom{v(x)}{\varphi(x)}}{\binom{L}{K}}|\varphi\rangle . \tag{2.34}
\end{equation*}
$$

The formula (2.34) involves binomial coefficients for numbers and for multisets (see Definition 1.8.1). The probabilities add up to one by Vandermonde's


Figure 2.5 Plots of hypergeometric and Pólya distributions; see Definitions 2.6.1 and 2.6.3 for details.
result, see Lemma 1.8.2 The demonstration that this formula is the right one will be provided soon, in Theorem 2.6.2.

First we provide an illustration. The hypergeometric distribution for draws of size 3 , that we used above, is:

$$
\begin{aligned}
& h g[3](4|R\rangle+6|B\rangle+2|G\rangle) \\
& \begin{array}{l}
\left.\left.\left.\left.\left.\left.\left.\left.=\frac{1}{55}|3| R\right\rangle\right\rangle+\frac{9}{55}|2| R\right\rangle+1|B\rangle\right\rangle+\frac{3}{11}|1| R\right\rangle+2|B\rangle\right\rangle+\frac{1}{11}|3| B\right\rangle\right\rangle \\
\left.\left.\left.\left.\left.\left.\quad+\frac{3}{55}|2| R\right\rangle+1|G\rangle\right\rangle+\frac{12}{55}|1| R\right\rangle+1|B\rangle+1|G\rangle\right\rangle+\frac{3}{22}|2| B\right\rangle+1|G\rangle\right\rangle \\
\left.\left.\left.\left.\quad+\frac{1}{55}|1| R\right\rangle+2|G\rangle\right\rangle+\frac{3}{110}|1| B\right\rangle+2|G\rangle\right\rangle .
\end{array}
\end{aligned}
$$

The hypergeometric distribution can be described as a channel of the form:

$$
\begin{equation*}
\mathcal{N}[L](X) \xrightarrow{h g[K]} \mathcal{N}[K](X) . \tag{2.35}
\end{equation*}
$$

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 the role in the formalisation of drawing, see Remark 2.2.1.

Theorem 2.6.2. Let $X$ be a set of colours and $v \in \mathcal{N}[L](X)$ be a multiset / urn of size $L$. For each $K \leq L$ and draw $\varphi \leq_{K} v$,

$$
\begin{aligned}
& \operatorname{hg}[K](v)(\varphi) \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \operatorname{Flrn}(v)\left(x_{1}\right) \cdot F \operatorname{lrn}\left(v-1\left|x_{1}\right\rangle\right)\left(x_{2}\right) \cdot \ldots \cdot \operatorname{Flrn}\left(v-\sum_{i<K} 1\left|x_{i}\right\rangle\right)\left(x_{K}\right) .
\end{aligned}
$$

Proof. Let's write the draw multiset as $\varphi=\sum_{j} n_{j}\left|y_{j}\right\rangle$. Then, for each list $\vec{x}=\left(x_{1}, \ldots, x_{K}\right) \in X^{K}$ with $\operatorname{acc}(\vec{x})=\varphi$, each element $y_{j}$ occurs $n_{j}$ times in $\vec{x}$. We first prove the following claim, for $\vec{x} \in \operatorname{acc}^{-1}(\varphi)$.

$$
\begin{equation*}
\prod_{1 \leq i \leq K}\left(v-\sum_{j<i} 1\left|x_{j}\right\rangle\right)\left(x_{i}\right)=\frac{v \rrbracket}{(v-\varphi) \rrbracket} \tag{*}
\end{equation*}
$$

The product on the left-hand-side in (*) does not depend on the order of the elements in $\vec{x}$ : each element $y_{j}$ occurs $n_{j}$ times in this product, with multiplicities $v\left(y_{j}\right), v\left(y_{j}\right)-1, \ldots, v\left(y_{j}\right)-n_{j}+1$, independently of the posititions of the $y_{j}$ in $\vec{x}$. Thus:

$$
\begin{aligned}
\prod_{1 \leq i \leq K}\left(v-\sum_{j<i} 1\left|x_{j}\right\rangle\right)\left(x_{i}\right) & =\prod_{j} v\left(y_{j}\right) \cdot\left(v\left(y_{j}\right)-1\right) \cdot \ldots \cdot\left(v\left(y_{j}\right)-n_{j}+1\right) \\
& =\prod_{j} v\left(y_{j}\right) \cdot \ldots \cdot\left(v\left(y_{j}\right)-\varphi\left(y_{j}\right)+1\right) \\
& =\prod_{j} \frac{v\left(y_{j}\right)!}{\left(v\left(y_{j}\right)-\varphi\left(y_{j}\right)\right)!}=\frac{v \rrbracket}{(v-\varphi) \rrbracket}
\end{aligned}
$$

Now we can finish the proof:

$$
\begin{aligned}
& \sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \prod_{1 \leq i \leq K} \operatorname{Flrn}\left(v-\sum_{j<i} 1\left|x_{j}\right\rangle\right)\left(x_{i}\right) \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \prod_{1 \leq i \leq K} \frac{\left(v-\sum_{j<i} 1\left|x_{j}\right\rangle\right)\left(x_{i}\right)}{L-i+1} \\
& \stackrel{(*)}{=} \sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{v \rrbracket}{(v-\varphi) \rrbracket} \cdot \frac{1}{\prod_{1 \leq i \leq K} L-i+1} \\
& =(\varphi) \cdot \frac{v \rrbracket}{(v-\varphi) \rrbracket} \cdot \frac{(L-K)!}{L!} \quad \text { by Proposition 1.7.2 } \\
& =\frac{v \rrbracket}{\varphi \rrbracket \cdot(v-\varphi) \rrbracket} \cdot \frac{K!\cdot(L-K)!}{L!} \quad \text { by Definition 1.7.1 (5) } \\
& \text { [1.42] } \frac{\binom{v}{\varphi}}{\binom{L}{K}} \\
& \stackrel{[2.34}{-} \operatorname{hg}[K](v)(\varphi) \text {. }
\end{aligned}
$$

### 2.6.2 Pólya draw-and-duplicate draws

In the hypergeometric distributions 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 grows - instead of decreases, 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 in the spread of contagious diseases [67], the flow of tourists [119], or topic classification [16].

Let's elaborate this dynamics for the urn $v=4|R\rangle+6|B\rangle+2|G\rangle \in \mathcal{N}[12](X)$, 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 $\operatorname{Flrn}(v)(R)=\frac{v(R)}{\|v\|}=\frac{1}{3}$. The resulting urn now has an additional red ball: $v_{1}=v+1|R\rangle=5|R\rangle+6|B\rangle+$ $2|G\rangle$.
2 The draw of the next red ball happens with probability $\operatorname{Flrn}\left(v_{1}\right)(R)=\frac{v_{1}(R)}{\left\|v_{1}\right\|}=$ $\frac{5}{13}$, giving an urn $v_{2}=v_{1}+1|R\rangle=6|R\rangle+6|B\rangle+2|G\rangle$.
3 The probability of the third red ball is now $\operatorname{Flrn}\left(v_{2}\right)(R)=\frac{v_{2}(R)}{\left\|v_{2}\right\|}=\frac{3}{7}$.
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 $v \in \mathcal{N}[L](X)$ be a non-empty urn over a set $X$, of size $L=\|v\|>0$, with a number $K \in \mathbb{N}$ describing the size of the draws from $v$. Since Pólya urns grow in size, we do not have need to put any restrictions on $K$.

The Pólya distribution $p l[K](v)$ uses the multichoose coefficients from Definitions 1.2 .3 and 1.8 .4 both for numbers and for multisets.

$$
\begin{equation*}
p l[K](v)=\sum_{\varphi \in \mathcal{N}[K](\operatorname{supp}(v))} \frac{\left.\binom{v}{\varphi}\right)}{\left(\binom{\|v\|}{K}\right)}|\varphi\rangle . \tag{2.36}
\end{equation*}
$$

Notice that the draws $\varphi$ are required to be multisets over $\operatorname{supp}(v)$. This guarantees that $\operatorname{supp}(\varphi) \subseteq \operatorname{supp}(v)$, so that we can only draw balls that are in the urn. In this situation it is most natural to use urns $v$ with full support, that is with $v(x)>0$ for each $x \in X$. In that case the urn contains at least one ball of each colour.
The above Pólya formula 2.36 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 .

$$
\begin{aligned}
& p 1[3](4|R\rangle+6|B\rangle+2|G\rangle) \\
& \left.\left.\left.\left.\left.\left.\left.\left.=\frac{5}{91}|3| R\right\rangle\right\rangle+\frac{15}{91}|2| R\right\rangle+1|B\rangle\right\rangle+\frac{3}{13}|1| R\right\rangle+2|B\rangle\right\rangle+\frac{2}{13}|3| B\right\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.\quad+\frac{5}{91}|2| R\right\rangle+1|G\rangle\right\rangle+\frac{12}{91}|1| R\right\rangle+1|B\rangle+1|G\rangle\right\rangle+\frac{3}{26}|2| B\right\rangle+1|G\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.\quad+\frac{3}{91}|1| R\right\rangle+2|G\rangle\right\rangle+\frac{9}{182}|1| B\right\rangle+2|G\rangle\right\rangle+\frac{1}{91}|3| G\right\rangle\right\rangle .
\end{aligned}
$$

In general, we can describe Pólya distributions as a channel of the form:

$$
\mathcal{N}_{f s}(X) \xrightarrow{p l[K]} \mathcal{N}[K](X)
$$

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 2.5 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 $v \in \mathcal{N}[L](X)$ be a multiset / urn of size $L$. For each draw $\varphi \in \mathcal{N}(X)$ with $\operatorname{supp}(\varphi) \subseteq \operatorname{supp}(v)$,

$$
\begin{aligned}
& p l[K](v)(\varphi) \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \operatorname{Flrn}(v)\left(x_{1}\right) \cdot \operatorname{Flrn}\left(v+1\left|x_{1}\right\rangle\right)\left(x_{2}\right) \cdot \ldots \cdot \operatorname{Flrn}\left(v+\sum_{i<K} 1\left|x_{i}\right\rangle\right)\left(x_{K}\right) .
\end{aligned}
$$

Proof. We write the draw multiset as $\varphi=\sum_{j} n_{j}\left|y_{j}\right\rangle$. We now use, for $\vec{x} \in$
$\operatorname{acc}^{-1}(\varphi)$, the equation:

$$
\begin{equation*}
\prod_{1 \leq i \leq K}\left(v+\sum_{j<i} 1\left|x_{j}\right\rangle\right)\left(x_{i}\right)=\frac{(v+\varphi-\mathbf{1}) \rrbracket}{(v-\mathbf{1}) \rrbracket} \tag{*}
\end{equation*}
$$

where $\mathbf{1}=\sum_{x \in \operatorname{supp}(v)} 1|x\rangle$. It holds by the same reasoning as in the proof of Theorem 2.6.2.

$$
\begin{aligned}
\prod_{0 \leq i<K}\left(v+\sum_{j<i} 1\left|x_{j}\right\rangle\right)\left(x_{i}\right) & =\prod_{j} v\left(y_{j}\right) \cdot\left(v\left(y_{j}\right)+1\right) \cdot \ldots \cdot\left(v\left(y_{j}\right)+\varphi\left(y_{j}\right)-1\right) \\
& =\prod_{j} \frac{\left(v\left(y_{j}\right)+\varphi\left(y_{j}\right)-1\right)!}{\left(v\left(y_{j}\right)-1\right)!} \\
& =\frac{(v+\varphi-\mathbf{1}) \rrbracket}{(v-\mathbf{1}) \rrbracket}
\end{aligned}
$$

Now we can derive the Pólya probability:

$$
\begin{aligned}
& \sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \prod_{1 \leq i \leq K} F \operatorname{lrn}\left(v+\sum_{j<i} 1\left|x_{j}\right\rangle\right)\left(x_{i}\right) \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \prod_{1 \leq i \leq K} \frac{\left(v+\sum_{j<i} 1\left|x_{j}\right\rangle\right)\left(x_{i}\right)}{L+i-1} \\
& \stackrel{(*)}{=} \sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{(v+\varphi-\mathbf{1}) \rrbracket}{(v-\mathbf{1}) \rrbracket} \cdot \frac{1}{\prod_{1 \leq i \leq K} L+i-1} \\
& =(\varphi) \cdot \frac{v \rrbracket}{(v-\varphi) \rrbracket} \cdot \frac{(L-1)!}{(L+K-1)!} \\
& =\frac{(v+\varphi-\mathbf{1}) \rrbracket}{\varphi \rrbracket \cdot(v-\mathbf{1}) \rrbracket} \cdot \frac{K!\cdot(L-1)!}{(L+K-1)!} \\
& =\frac{\left(\binom{v}{\varphi}\right)}{\left(\binom{L}{K}\right)} \\
& p l[K](v)(\varphi) \text {. }
\end{aligned}
$$

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 $\Gamma$, defined as:

$$
\begin{equation*}
\Gamma(z):=\int_{0}^{\infty} x^{z-1} \cdot e^{x} \mathrm{~d} x . \tag{2.37}
\end{equation*}
$$

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:

$$
\begin{equation*}
\Gamma(n)=(n-1)! \tag{2.38}
\end{equation*}
$$

This property guarantees that the formulation below really is a generalisation of the earlier formulation 2.36) of the Polya distribution.

Let $v \in \mathcal{M}_{f s}(X)$ be an urn, with full support, whose multiplicities $v(x)$ are thus positive real numbers. We (re)define:

$$
\begin{equation*}
p l[K](v):=\sum_{\varphi \in \mathcal{N}[K](X)}(\varphi) \cdot \frac{\Gamma(\|v\|)}{\Gamma(\|v\|+K)} \cdot \prod_{x \in X} \frac{\Gamma(v(x)+\varphi(x))}{\Gamma(v(x))}|\varphi\rangle . \tag{2.39}
\end{equation*}
$$

Notice that the draws $\varphi$ 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-anddelete 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 $m n[K](\omega)$ on $\mathcal{N}[K](X)$ is defined as:

$$
\begin{align*}
m n[K](\omega) & :=\sum_{\varphi \in \mathcal{N}[K](X)}(\varphi) \cdot \omega^{\varphi}|\varphi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)}(\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)}|\varphi\rangle . \tag{2.40}
\end{align*}
$$

For a channel $c: X \rightarrow \mathcal{D}(Y)$ we sometimes write more generally $m n[K](c):=$ $m n[K] \circ c: X \rightarrow \mathcal{D}(\mathcal{N}[K](Y))$. This general pointwise form of multinomial, for a channel $c$, arises naturally in an axiomatic setting [82].

We recall that the number $(\varphi)$ is the multinomial coefficient $\frac{K!}{\Pi_{x} \varphi(x)!}$, see Definition 1.7.1 (5]. It takes care of the different lists of single draws that accumulate to $\varphi$, see Proposition 1.7.2

The Multinomial Theorem (1.40) ensures that the probabilities in 2.40 add up to one:

$$
\sum_{\varphi \in \mathcal{N}[K](X)}(\varphi) \cdot \prod_{x} \omega(x)^{\varphi(x)} \stackrel{\boxed{1.40}}{=}\left(\sum_{x} \omega(x)\right)^{K}=1^{K}=1
$$

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:

$$
\begin{aligned}
m n[3](\omega)= & \left.\left.\left.\left.\left.\left.\left.\left.\frac{1}{27}|3| R\right\rangle\right\rangle+\frac{1}{6}|2| R\right\rangle+1|B\rangle\right\rangle+\frac{1}{4}|1| R\right\rangle+2|B\rangle\right\rangle+\frac{1}{8}|3| B\right\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.+\frac{1}{18}|2| R\right\rangle+1|G\rangle\right\rangle+\frac{1}{6}|1| R\right\rangle+1|B\rangle+1|G\rangle\right\rangle+\frac{1}{216}|3| G\right\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.+\frac{1}{36}|1| R\right\rangle+2|G\rangle\right\rangle+\frac{1}{24}|1| B\right\rangle+2|G\rangle\right\rangle+\frac{1}{8}|2| B\right\rangle+1|G\rangle\right\rangle .
\end{aligned}
$$

In general, we can describe multinomial distributions as a channel:

$$
\mathcal{D}(X) \xrightarrow{m_{n}[K]} \mathcal{N}[K](X) .
$$

In the end, we note that the draw multisets $\varphi$ in $\sqrt{2.40}$ can be restricted to those with $\operatorname{supp}(\varphi) \subseteq \operatorname{supp}(\omega)$. 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 $\Pi$ becomes zero, and so that $\varphi$ does not contribute to the above multinomial distribution.
There is an analogue of Theorem 2.6.2 and 2.6.2 for multinomial distributions. For a distribution $\omega \in \mathcal{D}(X)$, considered as urn, the product distribution $\operatorname{iid}[K](\omega)=\omega^{K} \in \mathcal{D}\left(X^{K}\right)$ 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: $X^{K} \rightarrow \mathcal{M}[K](X)$ from (1.35). In this way we recover the multinomial distribution, in analogy with coupon an 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:

$$
\begin{equation*}
m n[K](\omega)=\mathcal{D}(\operatorname{acc})(i i d[K](\omega)) \tag{2.41}
\end{equation*}
$$

Proof. By unfolding the relevant definitions:

$$
\begin{aligned}
\mathcal{D}(\operatorname{acc})(\operatorname{iid}[K](\omega)) & =\sum_{\vec{x} \in X^{K}} \omega^{K}(\vec{x})|\operatorname{acc}(\vec{x})\rangle \\
& \stackrel{(*)}{=} \sum_{\vec{x} \in X^{K}} \prod_{y \in X} \omega(y)^{\operatorname{acc}(\vec{x})(y)}|\operatorname{acc}(\vec{x})\rangle \\
& =\sum_{\varphi \in \mathcal{M}[K](X)} \sum_{\vec{x} \in \operatorname{acc}-1} \prod_{y \in X} \omega(y)^{\operatorname{acc}(\vec{x})(y)}|\operatorname{acc}(\vec{x})\rangle \\
& =\sum_{\varphi \in \mathcal{M}[K](X)} \sum_{\vec{\in} \in \operatorname{acc}^{-1}(\varphi)} \prod_{y \in X} \omega(y)^{\varphi(y)}|\varphi\rangle \\
& =\sum_{\varphi \in \mathcal{M}[K](X)}(\varphi) \cdot \prod_{y \in X} \omega(x)^{\varphi(y)}|\varphi\rangle \quad \text { by Proposition 1.7.2 } \\
& =m n[K](\omega) .
\end{aligned}
$$

The marked equation $\stackrel{(*)}{=}$ is the crucial step:

$$
\begin{aligned}
\omega^{K}(\vec{x}) & =\omega\left(x_{1}\right) \cdot \ldots \cdot \omega\left(x_{K}\right) \\
& =\prod_{y \in X} \omega(y)^{n_{y}} \quad \text { where } n_{y} \text { is the number of occurrences of } y \text { in } \vec{x} \\
& =\prod_{y \in X} \omega(y)^{\varphi(y)} \quad \text { where } \varphi=\operatorname{acc}(\vec{x}) .
\end{aligned}
$$

The literature sometimes distinguishes 'bivariate' and 'multivariate' draw distributions. In the bivarite 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) balls. 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 $[K]:[0,1] \rightarrow \mathcal{D}(\{0, \ldots, K\})$ is a bivariate version of the multinomial distribution $m n[K]: \mathcal{D}(X) \rightarrow \mathcal{D}(\mathcal{N}[K](X))$, 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 2, via the flip function from Example 2.1.2 (1), given by flip $(r)=r|1\rangle+(1-r)|0\rangle$.
- The set $\{0,1, \ldots, K\}$ of the first $\left.K+1=\binom{2}{K}\right)$ 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 ones $(k)=k|1\rangle+(K-k)|0\rangle$.

Via these two isomorphisms we get a commuting diagram that connects binomial and multinomial distributions:


Proof. For $r \in[0,1]$,

$$
\begin{aligned}
& (m n[K] \circ \text { flip })(r) \\
& =\sum_{\varphi \in \mathcal{N}[K](2)}(\varphi) \cdot \prod_{i \in 2} f \operatorname{fip}(r)(i)^{\varphi(i)}|\varphi\rangle \\
& \left.\left.=\sum_{k \in\{0, \ldots, K\}} \frac{K!}{k!\cdot(K-k)!} \cdot r^{k} \cdot(1-r)^{K-k}|k| 1\right\rangle+(K-k)|0\rangle\right\rangle \\
& \left.=\sum_{k \in\{0, \ldots, K\}} b n[K](r)(k) \mid \text { ones }(k)\right\rangle \\
& =(\mathcal{D}(\text { ones }) \circ b n[K])(r) .
\end{aligned}
$$

Via stick breaking one can express multinomial distributions in terms of multiple binomial distributions. This is a folklore result that occurs for instance in [86].

Proposition 2.6.9. For a sequence of probabilities $\vec{r}=\left(r_{0}, \ldots, r_{n-2}\right) \in(0,1)^{n-1}$ of length $n>1$, and for a draw multiset $\varphi \in \mathcal{N}[K](\boldsymbol{n})$ over $\boldsymbol{n}=\{0, \ldots, n-1\}$,

$$
\begin{aligned}
m n[K](\operatorname{stbr}(\vec{r}))(\varphi)=\operatorname{bn}[ & K]\left(r_{0}\right)(\varphi(0)) \\
& \cdot b n[K-\varphi(0)]\left(r_{1}\right)(\varphi(1)) \\
& \cdot b n[K-\varphi(0)-\varphi(1)]\left(r_{2}\right)(\varphi(2)) \\
& \cdot \ldots \cdot \operatorname{bn}\left[K-\sum_{i<n-2} \varphi(i)\right](\varphi(n-2)) .
\end{aligned}
$$

Proof. We use induction on $n>1$. When $n=2$ one has $\operatorname{stbr}\left(r_{0}\right)=r_{0}|0\rangle+(1-$ $\left.r_{0}\right)|1\rangle$, so that for $\varphi \in \mathcal{N}[K](\mathbf{2})$,

$$
m n[K]\left(\operatorname{stbr}\left(r_{0}\right)\right)(\varphi)=\binom{K}{\varphi(0)} \cdot r_{0}^{\varphi(0)} \cdot\left(1-r_{0}\right)^{\varphi(1)}=b n[K]\left(r_{0}\right)(\varphi(0))
$$

Next, let $\varphi=\sum_{i \leq n} k_{i}|i\rangle \in \mathcal{M}[K](\boldsymbol{n}+\mathbf{1})$ and $\vec{r}=r_{0}, \ldots, r_{n-1} \in(0,1)^{n}$ be given.

We use a shifted multiset $\varphi^{\prime}=\sum_{i<n-1} k_{i+1}|i\rangle$ of size $K-k_{0}$. Then:

$$
\begin{aligned}
& \operatorname{bn}[K]\left(r_{0}\right)\left(k_{0}\right) \cdot \operatorname{bn}\left[K-k_{0}\right]\left(r_{1}\right)\left(k_{1}\right) \cdot \ldots \cdot \operatorname{bn}\left[K-\sum_{i<n-1} k_{i}\right]\left(r_{n-1}\right)\left(k_{n-1}\right) \\
& \stackrel{(\mathrm{HH})}{=} \operatorname{bn}[K]\left(r_{0}\right)\left(k_{0}\right) \cdot \operatorname{mn}\left[K-k_{0}\right]\left(\operatorname{stbr}\left(r_{1}, \ldots, r_{n-1}\right)\right)\left(\varphi^{\prime}\right) \\
& \quad=\binom{K}{k_{0}} \cdot r_{0}^{k_{0}} \cdot\left(1-r_{0}\right)^{K-k_{0}} \cdot\left(\varphi^{\prime}\right) \cdot \prod_{i>0} \operatorname{stbr}\left(r_{1}, \ldots, r_{n-1}\right)(i)^{k_{i}} \\
& \quad=\frac{K!}{k_{0}!\cdot\left(K-k_{0}\right)!} \cdot \frac{\left(K-k_{0}\right)!}{k_{1}!\cdots k_{n-1}!} \cdot r_{0}^{k_{0}} \cdot \prod_{i>0}\left(\operatorname{stbr}\left(r_{1}, \ldots, r_{n-1}\right)(i) \cdot\left(1-r_{0}\right)\right)^{k_{i}} \\
& =(\varphi) \cdot \prod_{i \geq 0} \operatorname{stbr}\left(r_{0}, \ldots, r_{n-1}\right)(i)^{k_{i}} \\
& =\operatorname{mn}[K](\operatorname{stbr}(\vec{r}))(\varphi) .
\end{aligned}
$$

## 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:

$$
m n[7](\omega)(\varphi)=\frac{35}{432}
$$

Explain this outcome in terms of iterated single draws.
2 Consider urn $v=4|R\rangle+6|B\rangle+2|G\rangle \in \mathcal{N}[12](X)$ and compute:

$$
\operatorname{hg}[7](v)(\varphi)=\frac{5}{33} .
$$

3 Check that the Pólya probability of the same draw from the same urn is given by:

$$
p l[7](v)(\varphi)=\frac{35}{663} .
$$

2.6.2 Let $X$ be a non-empty finite set with $n$ elements, carrying the uniform distribution unif $X_{X}=\sum_{x \in X} \frac{1}{n}|x\rangle$. Show that:

$$
m n[K]\left(\text { unif }_{X}\right)=\sum_{\varphi \in \mathcal{N}[K](X)} \frac{(\varphi)}{n^{K}}|\varphi\rangle .
$$

Describe this distribution for $X=\{R, B, G\}$ and $K=4$. Relate it to Exercise 1.7.7.
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:

$$
p l[K](\mathbf{1})=\text { unif }_{\mathcal{N}[K](X)}=\sum_{\varphi \in \mathcal{N}[K](X)} \frac{1}{\binom{n}{K}}|\varphi\rangle .
$$

2.6.4 Let $v$ be a natural multiset. Use (2.38) to show that the two Pólya formulations 2.36 and 2.39) 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

$$
\left.\left.m n[1](\omega)=\sum_{x \in \operatorname{supp}(\omega)} \omega(x)|1| x\right\rangle\right\rangle .
$$

2.6.7 Use the formulation of the multinomial distribution in Theorem 2.6.7 together with the naturality of accumulation, in Exercise 1.7.12, and of iid, in Lemma 2.3.5 1 , to obtain the naturality of the multinomial channel: for each function $f: X \rightarrow Y$ the following diagram commutes.

2.6.8 Show that one can obtain the coupon distribution from Definition 2.3.7 via pushforward of the multionial distribution, i.e. show that:

$$
\mathcal{D}(\operatorname{supp})(m n[K](\omega))=\operatorname{cpn}[K](\omega) .
$$

2.6.9 Can you generalise Exercise 2.1.6to numbers $r_{1}, \ldots, r_{n}$ and set partitions of $X$ of size $n$ ?
2.6.10 Define yourself 'sequence' versions of the hypergeometric and Pólya distributions, forming channels:

where $X$ is finite and $K \leq L$. Do this in such a way that:

$$
\operatorname{acc} \odot \operatorname{seqhg}[K]=h g[K] \quad \text { and } \quad \text { acc } \odot \operatorname{seqpl}[K]=p l[K] .
$$

Hint: Use the formulations of Theorem 2.6.2 and 2.6.4, or more abstractly, the arrangement channel from (2.27).

### 2.7 Convolution

A convolution is a particular kind of (binary) operation in mathematics. In general, a convolution of parallel maps $f, g: X \rightarrow Y$ is a composite of the form:

$$
\begin{equation*}
X \xrightarrow{\text { split }} X \times X \xrightarrow{f \otimes g} Y \times Y \xrightarrow{\text { join }} Y \tag{2.43}
\end{equation*}
$$

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 (2.43).

Definition 2.7.1. Let $M=(M,+, 0)$ be a commutative monoid. One defines a sum and zero element on $\mathcal{D}(M)$ via:

$$
\begin{align*}
\omega+\rho & :=\mathcal{D}(+)(\omega \otimes \rho) \quad \text { using } \quad\left\{\begin{array}{l}
+: M \times M \rightarrow M \\
0 \in M
\end{array} \quad:=1|0\rangle\right. \tag{2.44}
\end{align*}
$$

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.

Alternative, equivalent descriptions of this sum of distributions are:

$$
\begin{aligned}
\omega+\rho & =\sum_{a, b \in M} \omega(a) \cdot \rho(b)|a+b\rangle \\
\left(\sum_{i} r_{i}\left|a_{i}\right\rangle\right)+\left(\sum_{j} s_{j}\left|b_{j}\right\rangle\right) & =\sum_{i, j} r_{i} \cdot s_{j}\left|a_{i}+b_{j}\right\rangle
\end{aligned}
$$

We may also 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.


$$
\begin{align*}
& \mathrm{a} \leftarrow \omega \\
& \mathrm{~b} \leftarrow \rho  \tag{2.45}\\
& \text { return } \quad \mathrm{a}+\mathrm{b}
\end{align*}
$$

The sum of discrete distributions $\omega+\rho$ occurs in [111, p.82], but does not seem to be widely used and/or familiar. It is an instance of an abstract form of convolution in [114, §10], like in Diagram (2.43).

As announced, convolution gives a commutative monoid structure. But there is more to say.

## Proposition 2.7.2.

1 Via sums,+ 0 of distributions (2.44), the set $\mathcal{D}(M)$ forms a commutative monoid.
2 If $f: M \rightarrow N$ is a homomorphism of commutative monoids, then so is $\mathcal{D}(f): \mathcal{D}(M) \rightarrow \mathcal{D}(N)$.

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 $\mathcal{D}$ : Sets $\rightarrow$ Sets can be restricted to a functor $\mathcal{D}:$ CMon $\rightarrow$ CMon in a commuting diagram:


The vertical arrows 'forget' the monoid structure, by sending a monoid to its underlying set.

In Exercise2.7.8 we shall see that the restricted (or lifted) functor $\mathcal{D}: \mathbf{C M o n} \rightarrow$ CMon is also a monad. The same construction works for $\mathcal{D}_{\infty}$.

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,

$$
\omega+0=\omega+1|0\rangle=\sum_{a \in M} \omega(a) \cdot 1|a+0\rangle=\sum_{a \in M} \omega(a)|a\rangle=\omega .
$$

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 $f: M \rightarrow N$ is a homomorphism of monoids. We then have an equation $f \circ+=+\circ(f \times f)$, see Diagram (1.7). Hence:

$$
\begin{aligned}
\mathcal{D}(f)(\omega+\rho) & \stackrel{\sqrt{2.44}}{=}(\mathcal{D}(f) \circ \mathcal{D}(+))(\omega \otimes \rho) \\
& =\mathcal{D}(f \circ+)(\omega \otimes \rho) \\
& =\mathcal{D}(+\circ(f \times f))(\omega \otimes \rho) \quad f \text { is a homomorphism } \\
& =(\mathcal{D}(+) \circ \mathcal{D}(f \times f))(\omega \otimes \rho) \\
& =\mathcal{D}(+)(\mathcal{D}(f)(\omega) \otimes \mathcal{D}(f)(\rho)) \quad \text { by Lemma 2.3.2 3. } \\
& \stackrel{[2.44}{=} \mathcal{D}(f)(\omega)+\mathcal{D}(f)(\rho) .
\end{aligned}
$$

The next example illustrates how we sometimes have to stretch the underlying space a bit to make it into a monoid so that the convolution structure can be used for convolution.

Example 2.7.3. Recall that we write dice for the uniform (fair) dice distribution $\sum_{1 \leq i \leq 6} \frac{1}{6}|i\rangle \in \mathcal{D}(p i p s)$, 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: $\operatorname{supp}($ dice $)=$ pips.
A big benefit of this change of perspective 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:
$2 \cdot$ dice $=$ dice + dice $\stackrel{\sqrt{2.44}}{-} \mathcal{D}(+)($ dice $\otimes$ dice $)$

$$
\begin{aligned}
& =\mathcal{D}(+)\left(\sum_{i, j \in p i p s} \frac{1}{36}|i, j\rangle\right) \\
& = \\
& \sum_{i, j \in p i p s} \frac{1}{36}|i+j\rangle \\
& = \\
& \quad \frac{1}{36}|2\rangle+\frac{1}{18}|3\rangle+\frac{1}{12}|4\rangle+\frac{1}{9}|5\rangle+\frac{5}{36}|6\rangle+\frac{1}{6}|7\rangle \\
& \quad \quad \quad+\frac{5}{36}|8\rangle+\frac{1}{9}|9\rangle+\frac{1}{12}|10\rangle+\frac{1}{18}|11\rangle+\frac{1}{36}|12\rangle .
\end{aligned}
$$

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 (2) 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 2.7.4 contains another variation, with modular arithmetic.
The general structure of Diagram (2.43) 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 2.7.4.

1 The convolution sum of multiple fip distributions is a binomial distribution: for $r \in[0,1]$,

$$
\operatorname{flip}(r)+\operatorname{flip}(r)=b n[2](r) .
$$

In a diagram:


More generally, in K-ary form:

$$
K \cdot f l i p(r)=\underbrace{\text { flip }(r)+\cdots+\text { flip }(r)}_{K \text { times }}=b n[K](r) .
$$

2 Binomials themselves are closed under convolution:


This last equation can also be obtained by marginalising out the wire on the right in Theorem 2.5.1 2. We shall see later in Exercise 3.3.9 that multinomial distributions are also closed under convolution.

Proof. 1 We do the binary case:

$$
\begin{aligned}
& \text { flip }(r)+\text { flip }(r) \\
& =\mathcal{D}(+)(\text { flip }(r) \otimes \text { flip }(r)) \\
& =\mathcal{D}(+)((r|1\rangle+(1-r)|0\rangle) \otimes(r|1\rangle+(1-r)|0\rangle)) \\
& =\mathcal{D}(+)\left(r^{2}|1,1\rangle+r \cdot(1-r)|1,0\rangle+(1-r) \cdot r|0,1\rangle+(1-r)^{2}|0,0\rangle\right) \\
& =r^{2}|2\rangle+2 \cdot r \cdot(1-r)|1\rangle+(1-r)^{2}|0\rangle \\
& =\sum_{0 \leq i \leq 2}\binom{2}{i} \cdot r^{i} \cdot(1-r)^{2-i}|i\rangle=\operatorname{bn}[2](r) .
\end{aligned}
$$

2 For $r \in[0,1]$ and $K, L \in \mathbb{N}$ we compute, via Vandermonde's binary formula 1.44):

$$
\begin{aligned}
& \mathcal{D}(+)(b n[K](r) \otimes b n[L](r)) \\
&=\mathcal{D}(+)\left(\sum_{0 \leq i \leq K} \sum_{0 \leq j \leq L}\binom{K}{i} \cdot r^{i} \cdot(1-r)^{K-i} \cdot\binom{L}{j} \cdot r^{j} \cdot(1-r)^{L-j}|i, j\rangle\right) \\
&=\sum_{0 \leq i \leq K} \sum_{0 \leq j \leq L}\binom{K}{i} \cdot\binom{L}{j} \cdot r^{i+j} \cdot(1-r)^{(K+L)-(i+j)}|i+j\rangle \\
&=\sum_{0 \leq m \leq K+L}\left(\sum_{0 \leq i \leq K, 0 \leq j \leq L, i+j=m}\binom{K}{i} \cdot\binom{L}{j}\right) \cdot r^{m} \cdot(1-r)^{(K+L)-m}|m\rangle \\
& \frac{\square 1.44}{=} \sum_{0 \leq m \leq K+L}\binom{K+L}{m} \cdot r^{m} \cdot(1-r)^{(K+L)-m}|m\rangle \\
&=b n[K+L](r) .
\end{aligned}
$$

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 2.7.5. For a distribution $\omega \in \mathcal{D}(X)$ and numbers $K, L$,

$$
\mathcal{D}(++)(\operatorname{iid}[K](\omega) \otimes \operatorname{iid}[L](\omega))=\operatorname{iid}[K+L]
$$

This means that the following convolution-style diagram commutes.


Proof. Since:

$$
\begin{aligned}
(+\odot(\operatorname{iid}[K] \otimes \operatorname{iid}[L]) \odot \Delta)(\omega) & =\sum_{\vec{x} \in X^{K}, \vec{k} \in X^{L}} \omega^{K}(\vec{x}) \cdot \omega^{L}(\vec{y})|\vec{x}+\vec{y}\rangle \\
& =\sum_{\vec{x} \in X^{K}, \vec{y} \in X^{L}} \omega^{K+L}(\vec{x}+\vec{y})|\vec{x}+\vec{y}\rangle \\
& =\sum_{\vec{z} \in X^{K+L}} \omega^{K+L}(\vec{z})|\vec{z}\rangle \\
& =\omega^{K+L} \\
& =\operatorname{iid}[K+L](\omega) .
\end{aligned}
$$

We conclude with a basic property of Poisson distributions pois[ $\lambda$ ], see 2.3. This property is commonly expressed in terms of random variables $X_{i}$ as: if $X_{1} \sim \operatorname{pois}\left[\lambda_{1}\right]$ and $X_{2} \sim \operatorname{pois}\left[\lambda_{2}\right]$ then $X_{1}+X_{2} \sim \operatorname{pois}\left[\lambda_{1}+\lambda_{2}\right]$. 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 2.1.5, but that difference is immaterial here. We now use the mapping $\lambda \mapsto$ pois $[\lambda]$ as a function $\mathbb{R}_{\geq 0} \rightarrow \mathcal{D}_{\infty}(\mathbb{N})$ and as a $\mathcal{D}_{\infty}$-channel pois: $\mathbb{R}_{\geq 0} \hookrightarrow \mathbb{N}$.

Proposition 2.7.6. The Poisson channel pois: $\mathbb{R}_{\geq 0} \rightarrow \mathcal{D}_{\infty}(\mathbb{N})$ is a homomorphism of monoids. Indeed, for $\lambda_{1}, \lambda_{2} \in \mathbb{R}_{\geq 0}$,

$$
\operatorname{pois}\left[\lambda_{1}+\lambda_{2}\right]=\operatorname{pois}\left[\lambda_{1}\right]+\operatorname{pois}\left[\lambda_{2}\right] \quad \text { and } \quad \operatorname{pois}[0]=1|0\rangle .
$$

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.


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 $k \in \mathbb{N}$.

$$
\begin{aligned}
\left(\text { pois }\left[\lambda_{1}\right]+\operatorname{pois}\left[\lambda_{2}\right]\right)(k) & \stackrel{[2.44}{=} \mathcal{D}(+)\left(\operatorname{pois}\left[\lambda_{1}\right] \otimes \operatorname{pois}\left[\lambda_{2}\right]\right)(k) \\
& =\sum_{k_{1}, k_{2}, k_{1}+k_{2}=k}\left(\operatorname{pois}\left[\lambda_{1}\right] \otimes \operatorname{pois}\left[\lambda_{2}\right]\right)\left(k_{1}, k_{2}\right) \\
& =\sum_{0 \leq m \leq k} \operatorname{pois}\left[\lambda_{1}\right](m) \cdot \operatorname{pois}\left[\lambda_{2}\right](k-m) \\
& \stackrel{(2.3)}{=} \sum_{0 \leq m \leq k}\left(e^{-\lambda_{1}} \cdot \frac{\lambda_{1}^{m}}{m!}\right) \cdot\left(e^{-\lambda_{2}} \cdot \frac{\lambda_{2}^{k-m}}{(k-m)!}\right) \\
& =\sum_{0 \leq m \leq k} \frac{e^{-\left(\lambda_{1}+\lambda_{2}\right)}}{k!} \cdot \frac{k!}{m!\cdot(k-m)!} \cdot \lambda_{1}^{m} \cdot \lambda_{2}^{k-m} \\
& =\frac{e^{-\left(\lambda_{1}+\lambda_{2}\right)}}{k!} \cdot \sum_{0 \leq m \leq k}\binom{k}{m} \cdot \lambda_{1}^{m} \cdot \lambda_{2}^{k-m} \\
& \stackrel{1.39}{=} \frac{e^{-\left(\lambda_{1}+\lambda_{2}\right)}}{k!} \cdot\left(\lambda_{1}+\lambda_{2}\right)^{k} \\
\frac{2.3!}{=} & \operatorname{pois}\left[\lambda_{1}+\lambda_{2}\right](k) .
\end{aligned}
$$

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 $[0]=1|0\rangle$.

## Exercises

2.7.1 Describe $3 \cdot$ dice $\in \mathcal{D}(\mathbb{N})$ in detail, following Example 2.7.3
2.7.2 The set of natural numbers $\mathbb{N}$ has two commutative monoid structures, one additive with,+ 0 , and one multiplicative with $\cdot, 1$. Accordingly,

Definiton 2.7.1 gives two commutative monoid structures on $\mathcal{D}(\mathbb{N})$, namely:

$$
\omega+\rho=\mathcal{D}(+)(\omega \otimes \rho) \quad \text { and } \quad \omega \star \rho=\mathcal{D}(\cdot)(\omega \otimes \rho)
$$

Consider the following three distributions on $\mathbb{N}$.

$$
\rho_{1}=\frac{1}{2}|0\rangle+\frac{1}{3}|1\rangle+\frac{1}{6}|2\rangle \quad \rho_{2}=\frac{1}{2}|0\rangle+\frac{1}{2}|1\rangle \quad \omega=\frac{2}{3}|0\rangle+\frac{1}{3}|1\rangle .
$$

Show consecutively:

```
\(\rho_{1}+\rho_{2}=\frac{1}{4}|0\rangle+\frac{5}{12}|1\rangle+\frac{1}{4}|2\rangle+\frac{1}{12}|3\rangle ;\)
\(\omega \star\left(\rho_{1}+\rho_{2}\right)=\frac{3}{4}|0\rangle+\frac{5}{36}|1\rangle+\frac{1}{12}|2\rangle+\frac{1}{36}|3\rangle ;\)
\(\omega \star \rho_{1}=\frac{5}{6}|0\rangle+\frac{1}{9}|1\rangle+\frac{1}{18}|2\rangle ;\)
\(\omega \star \rho_{2}=\frac{5}{6}|0\rangle+\frac{1}{6}|1\rangle ;\)
\(\left(\omega \star \rho_{1}\right)+\left(\omega \star \rho_{2}\right)=\frac{25}{36}|0\rangle+\frac{25}{108}|1\rangle+\frac{7}{108}|2\rangle+\frac{1}{108}|3\rangle\).
```

Observe that $\star$ does not distribute over + on $\mathcal{D}(\mathbb{N})$. More generally, conclude that the convolution construction of Definition 2.7.1 does not extend to commutative semirings.
2.7.3 Show that negative binomials nbn, see Example 2.1.7 (3), are also closed under convolution:

$$
n b n[K](s)+\operatorname{nbn}[L](s)=n b n[K+L](s),
$$

where the sum + of distributions on the left-hand-side is the one from Definition 2.7.1 for $\mathcal{D}_{\infty}$.
2.7.4 Recall that for $N \in \mathbb{N}_{>0}$ we write $N=\{0,1, \ldots, N-1\}$ 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:

$$
u n i f_{N}+\text { unif }_{N}=\text { unif }_{N}, \quad \text { with }+ \text { from Definition 2.7.1 }
$$

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 2.7.3. See [162] for more info.
2.7.5 Consider a function $f: X \rightarrow M$ where $X$ is an ordinary set and $M$ is a commutative monoid. We can add noise to the function $f$ via a channel $c: X \leadsto M$. The result is a channel noise $(f, c): X \leadsto M$ given by pointwise convolution:

$$
\operatorname{noise}(f, c):=\langle f\rangle+c
$$

Check that we can concretely describe this noise channel as:

$$
\operatorname{noise}(f, c)(x)=\sum_{y \in Y} c(x)(y)|f(x)+y\rangle .
$$

2.7.6 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:

$$
\star:=(\mathcal{N}(M) \times \mathcal{N}(M) \xrightarrow{\otimes} \mathcal{N}(M \times M) \xrightarrow{\mathcal{N}(+)} \mathcal{N}(M))
$$

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:

$$
\begin{gathered}
(2|1\rangle+3|2\rangle) \star(1|0\rangle+1|1\rangle+1|2\rangle) \\
\left(2 x+3 x^{2}\right) \cdot\left(1+x+x^{2}\right) .
\end{gathered}
$$

Show in general that $\star$ corresponds to multiplication of polynomials.
2.7.7 We consider the binomial distribution from Example 2.1.2 (2) as a channel bn $[-](-): \mathbb{N} \times[0,1] \mapsto \mathbb{N}$ with two arguments / inputs. We further write prod: $[0,1] \times[0,1] \rightarrow[0,1]$ for the obvious multiplication function $\operatorname{prod}(r, s)=r \cdot s$, 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

$$
b n[-](s) \gg b n[K](r)=b n[K](r \cdot s) \quad b n[K](1)=1|K\rangle .
$$

2 Prove these equations.
3 Check that they express that $b n[-](-): \mathbb{N} \times[0,1] \rightsquigarrow[0,1]$ 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'.)
2.7.8 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 : $M \rightarrow \mathcal{D}(M)$ is a homomorphism of monoids, which can also be expressed via commutation of the diagram:


2 Check also that flat: $\mathcal{D}(\mathcal{D}(M)) \rightarrow \mathcal{D}(M)$ is a monoid homomorphism. This means that for $\Omega, \Theta \in \mathcal{D}(\mathcal{D}(M))$ one has:

$$
f f a t(\Omega+\Theta)=\text { flat }(\Omega)+\operatorname{flat}(\Theta) \quad \text { and } \quad \text { flat }(1|1| 0\rangle\rangle)=1|0\rangle
$$

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 $\mathcal{D}: \mathbf{C M o n} \rightarrow \mathbf{C M o n}$ in 2.46 is also a monad.

### 2.8 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 4.5 later on. At this stage we describe a different comparison, called divergence, or more fully Kullback-Leibler divergence, written as $D_{K L}$. It is a not a distance function since it is not symmetric: $D_{K L}(\omega, \rho) \neq D_{K L}(\rho, \omega)$, in general. But it does satisfy $D_{K L}(\omega, \rho)=0$ iff $\omega=\rho$, and other useful properties.

In the sequel we shall make frequent use of the Kullback-Leibler divergence $D_{K L}$. This section collects the definition and some basic facts. It assumes rudimentary familiarity with the logarithm function $\log : \mathbb{R}_{>0} \rightarrow \mathbb{R}$. The crucial property is $\log (x)=y$ iff $x=2^{y}$, for $\log =\log _{2}$. Sometimes it is more convenient to use the natural logarithm $\ln =\log _{e}$, with Euler's number $e=2.718 \ldots$
as base. These logarithms with different bases are related via a constant, as $\log _{b}(x)=\frac{\ln (x)}{\ln (b)}$. A crucial property is that the logarithm sends multiplications to sums, see Exercise 1.4.2

Definition 2.8.1. Let $\omega, \rho$ be two distributions/states on the same set $X$ with $\operatorname{supp}(\omega) \subseteq \operatorname{supp}(\rho)$. The Kullback-Leibler divergence, or KL-divergence, or simply divergence, of $\omega$ from $\rho$ is:

$$
\begin{equation*}
D_{K L}(\omega, \rho):=\sum_{x \in X} \omega(x) \cdot \log \left(\frac{\omega(x)}{\rho(x)}\right) . \tag{2.47}
\end{equation*}
$$

The convention is that $r \cdot \log (r)=0$ when $r=0$.
The inclusion $\operatorname{supp}(\omega) \subseteq \operatorname{supp}(\rho)$ is equivalent to: $\rho(x)=0$ implies $\omega(x)=$ 0 . This requirement immediately implies that divergence is not symmetric. But even when $\omega$ and $\rho$ do have the same support, the divergences $D_{K L}(\omega, \rho)$ and $D_{K L}(\rho, \omega)$ are different, in general, see Exercise 2.8.1 below for an easy illustration. To emphasise the difference, some people write $D_{K L}(\rho \| \omega)$ instead of $D_{K L}(\rho, \omega)$.

Whenever we write an expression $D_{K L}(\omega, \rho)$ we will implicitly assume an inclusion $\operatorname{supp}(\omega) \subseteq \operatorname{supp}(\rho)$.

We start with some easy properties of divergence.
Lemma 2.8.2. Let $\omega, \rho \in \mathcal{D}(X)$ and $\omega^{\prime}, \rho^{\prime} \in \mathcal{D}(Y)$ be distributions.
1 Zero-divergence is the same as equality:

$$
D_{K L}(\omega, \rho)=0 \Longleftrightarrow \omega=\rho .
$$

2 Divergence of tensor products is a sum of divergences:

$$
D_{K L}\left(\omega \otimes \omega^{\prime}, \rho \otimes \rho^{\prime}\right)=D_{K L}(\omega, \rho)+D_{K L}\left(\omega^{\prime}, \rho^{\prime}\right) .
$$

Proof. 1 The direction $(\Leftarrow)$ is easy, since $\log (1)=0$. For $(\Rightarrow)$, let $0=D_{K L}(\omega, \rho)=$ $\sum_{x} \omega(x) \cdot \log (\omega(x) / \rho(x))$. This means that if $\omega(x) \neq 0$, one has $\log (\omega(x) / \rho(x))=0$, and thus $\omega(x) / \rho(x)=1$ and $\omega(x)=\rho(x)$. In particular:

$$
\begin{equation*}
1=\sum_{x \in \operatorname{supp}(\omega)} \omega(x)=\sum_{x \in \operatorname{supp}(\omega)} \rho(x) \tag{*}
\end{equation*}
$$

By assumption we have $\operatorname{supp}(\omega) \subseteq \operatorname{supp}(\rho)$. Write $\operatorname{supp}(\rho)$ as disjoint union $\operatorname{supp}(\omega) \cup U$ for some $U \subseteq \operatorname{supp}(\rho)$. It suffices to show $U=\emptyset$. We have:

$$
1=\sum_{x \in \operatorname{supp}(\rho)} \rho(x)=\sum_{x \in \operatorname{supp}(\omega)} \rho(x)+\sum_{x \in U} \rho(x)=1+\sum_{x \in U} \rho(x) .
$$

Hence $U=\emptyset$.

2 By unwrapping the relevant definitions and using that log sends multiplications to sums:

$$
\begin{aligned}
& D_{K L}\left(\omega \otimes \omega^{\prime}, \rho \otimes \rho^{\prime}\right) \\
& =\sum_{x \in X, y \in Y}\left(\omega \otimes \omega^{\prime}\right)(x, y) \cdot \log \left(\frac{\left(\omega \otimes \omega^{\prime}\right)(x, y)}{\left(\rho \otimes \rho^{\prime}\right)(x, y)}\right) \\
& =\sum_{x \in X, y \in Y} \omega(x) \cdot \omega^{\prime}(y) \cdot \log \left(\frac{\omega(x)}{\rho(x)} \cdot \frac{\omega^{\prime}(y)}{\rho^{\prime}(y)}\right) \\
& =\sum_{x \in X, y \in Y} \omega(x) \cdot \omega^{\prime}(y) \cdot\left(\log \left(\frac{\omega(x)}{\rho(x)}\right)+\log \left(\frac{\omega^{\prime}(y)}{\rho^{\prime}(y)}\right)\right) \\
& =\sum_{x \in X, y \in Y} \omega(x) \cdot \omega^{\prime}(y) \cdot \log \left(\frac{\omega(x)}{\rho(x)}\right)+\sum_{x \in X, y \in Y} \omega(x) \cdot \omega^{\prime}(y) \cdot \log \left(\frac{\omega^{\prime}(y)}{\rho^{\prime}(y)}\right) \\
& =\sum_{x \in X} \omega(x) \cdot \log \left(\frac{\omega(x)}{\rho(x)}\right)+\sum_{y \in Y} \omega^{\prime}(y) \cdot \log \left(\frac{\omega^{\prime}(y)}{\rho^{\prime}(y)}\right) \\
& =D_{K L}(\omega, \rho)+D_{K L}\left(\omega^{\prime}, \rho^{\prime}\right) .
\end{aligned}
$$

A joint distribution $\omega$ is typically 'entwined', that is, it is different from the product $\omega[1,0] \otimes \omega[0,1]$ of its marginals, see Example 2.3.9 (2]. The KullbackLeibler divergence $D_{K L}(\omega, \omega[1,0] \otimes \omega[0,1])$ 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 2.8 .6

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 in learning.

Lemma 2.8.3 (Jensen's inequality). Let $f: \mathbb{R}_{>0} \rightarrow \mathbb{R}$ be a function whose second derivative is negative: $f^{\prime \prime}<0$. Then for all $a_{1}, \ldots, a_{n} \in \mathbb{R}_{>0}$ and $r_{1}, \ldots, r_{n} \in[0,1]$ with $\sum_{i} r_{i}=1$ there is an inequality:

$$
\begin{equation*}
f\left(\sum_{i} r_{i} \cdot a_{i}\right) \geq \sum_{i} r_{i} \cdot f\left(a_{i}\right) \tag{2.48}
\end{equation*}
$$

The inequality is strict, except in trivial cases.
The inequality holds in particular for logarithms, when $f=\log$, or $f=\ln$.
The proof is standard but is included, for convenience.
Proof. We shall provide a proof for $n=2$. The inequality is easily extended to $n>2$, by induction. So let $a, b \in \mathbb{R}_{>0}$ be given, with $r \in[0,1]$. We need to prove $f(r a+(1-r) b) \geq r f(a)+(1-r) f(b)$. The result is trivial if $a=b$ or
$r=0$ or $r=1$. So let, without loss of generality, $a<b$ and $r \in(0,1)$. Write $c:=r a+(1-r) b=b-r(b-a)$, so that $a<c<b$. By the mean value theorem we can find $a<u<c$ and $c<v<b$ with:

$$
\frac{f(c)-f(a)}{c-a}=f^{\prime}(u) \quad \text { and } \quad \frac{f(b)-f(c)}{b-c}=f^{\prime}(v)
$$

Since $f^{\prime \prime}<0$ we have that $f^{\prime}$ is strictly decreasing, so $f^{\prime}(u)>f^{\prime}(v)$ 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 $f^{\prime}(u)>f^{\prime}(v)$ we deduce inequalities:

$$
\frac{f(c)-f(a)}{(1-r)(b-a)}>\frac{f(b)-f(c)}{r(b-a)} \quad \text { i.e. } \quad r(f(c)-f(a))>(1-r)(f(b)-f(c) .
$$

By reorganising the latter inequality we get $f(c)>r f(a)+(1-r) f(b)$, as required.

We can now say a bit more about divergence. For instance, that it is nonnegative, as one expects.

Proposition 2.8.4. Let $\omega, \rho \in \mathcal{D}(X)$ be states on the same space $X$.
$1 D_{K L}(\omega, \rho) \geq 0$.
2 State transformation is $D_{K L}$-non-expansive: for a channel $c: X \leadsto Y$ and states $\omega, \rho \in \mathcal{D}(X)$ one has:

$$
D_{K L}(c \gg=\omega, c \gg=\rho) \leq D_{K L}(\omega, \rho) .
$$

The first item shows that Kullback-Leibler divergence is non-negative. The divergence is not bounded: it can become arbitrarily large, see Exercise 2.8.5 below.

Proof. 1 Via Jensen's inequality, we deduce for the minus of divergence:

$$
\begin{aligned}
-D_{K L}(\omega, \rho) & =\sum_{x \in X} \omega(x) \cdot \log \left(\frac{\rho(x)}{\omega(x)}\right) \\
& \leq \log \left(\sum_{x \in X} \omega(x) \cdot \frac{\rho(x)}{\omega(x)}\right)=\log \left(\sum_{x \in X} \rho(x)\right)=\log (1)=0 .
\end{aligned}
$$

2 Again Via Jensen's inequality:

$$
\begin{aligned}
D_{K L}(c \gg=\omega, c \gg=\rho) & =\sum_{y \in Y}(c \gg=\omega)(y) \cdot \log \left(\frac{(c \gg=\omega)(y)}{(c \gg=\rho)(y)}\right) \\
& =\sum_{x \in X, y \in Y} \omega(x) \cdot c(x)(y) \cdot \log \left(\frac{(c \gg=\omega)(y)}{(c \gg=\rho)(y)}\right) \\
& \leq \sum_{x \in X} \omega(x) \cdot \log \left(\sum_{y \in Y} c(x)(y) \cdot \frac{(c \gg=\omega)(y)}{(c \gg=\rho)(y)}\right) .
\end{aligned}
$$

Hence it suffices to prove:

$$
\begin{aligned}
\sum_{x \in X} \omega(x) \cdot \log \left(\sum_{y \in Y} c(x)(y) \cdot \frac{(c \gg=\omega)(y)}{(c \gg=\rho)(y)}\right) & \leq D_{K L}(\omega, \rho) \\
& =\sum_{x \in X} \omega(x) \cdot \log \left(\frac{\omega(x)}{\rho(x)}\right) .
\end{aligned}
$$

This inequality $\leq$ follows from another application of Jensen's inequality:

$$
\begin{aligned}
& \sum_{x \in X} \omega(x) \cdot \log \left(\sum_{y \in Y} c(x)(y) \cdot \frac{(c \gg=\omega)(y)}{(c \gg=\rho)(y)}\right)-\sum_{x \in X} \omega(x) \cdot \log \left(\frac{\omega(x)}{\rho(x)}\right) \\
& =\sum_{x \in X} \omega(x) \cdot\left[\log \left(\sum_{y \in Y} c(x)(y) \cdot \frac{(c \gg=\omega)(y)}{(c \gg=\rho)(y)}\right)-\log \left(\frac{\omega(x)}{\rho(x)}\right)\right] \\
& =\sum_{x \in X} \omega(x) \cdot \log \left(\sum_{y \in Y} c(x)(y) \cdot \frac{(c \gg=\omega)(y)}{(c \gg=\rho)(y)} \cdot \frac{\rho(x)}{\omega(x)}\right) \\
& \leq \log \left(\sum_{x \in X, y \in Y} \omega(x) \cdot c(x)(y) \cdot \frac{(c \gg=\omega)(y)}{(c \gg=\rho)(y)} \cdot \frac{\rho(x)}{\omega(x)}\right) \\
& =\log \left(\sum_{y \in Y}\left(\sum_{x \in X} c(x)(y) \cdot \rho(x)\right) \cdot \frac{(c \gg=\omega)(y)}{(c \gg=\rho)(y)}\right) \\
& =\log \left(\sum_{y \in Y}(c \gg=\omega)(y)\right) \\
& =\log (1)=0 .
\end{aligned}
$$

## Exercises

2.8.1 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:

$$
\begin{array}{ll}
1 & D_{K L}(\omega, \rho)=\frac{3}{4} \cdot \log (3)-1 \approx 0.19 . \\
2 & D_{K L}(\rho, \omega)=1-\frac{1}{2} \cdot \log (3) \approx 0.21 .
\end{array}
$$

2.8.2 Check that for $y \in \operatorname{supp}(\rho)$ one has:

$$
D_{K L}(1|y\rangle, \rho)=-\log (\rho(y))
$$

2.8.3 1 Show that the function $D_{K L}(\omega,-)$ yields an inequality on convex sums:

$$
\begin{aligned}
& D_{K L}\left(\omega, r_{1} \cdot \rho_{1}+\cdots+r_{n} \cdot \rho_{n}\right) \\
& \quad \leq r_{1} \cdot D_{K L}\left(\omega, \rho_{1}\right)+\cdots+r_{n} \cdot D_{K L}\left(\omega, \rho_{n}\right) .
\end{aligned}
$$

2 Prove the following "log sum inequality" from [31, Thm. 2.7.1]: for finite collections $a_{i}, b_{i} \in \mathbb{R}_{>0}$,

$$
\left(\sum_{i} a_{i}\right) \cdot \log \left(\frac{\sum_{i} a_{i}}{\sum_{i} b_{i}}\right) \leq \sum_{i} a_{i} \cdot \log \left(\frac{a_{i}}{b_{i}}\right) .
$$

Hint: Use Jensen's inequality with $f(x)=-x \cdot \log (x)$, for $x \in \mathbb{R}_{>0}$, in Lemma 2.8.3
3 Use this $\log$ sum inequality to show that $D_{K L}$ also yields an inequality on (the same) convex sums in both arguments (as in [31, Thm. 2.7.2]):

$$
\begin{aligned}
& D_{K L}\left(r_{1} \cdot \omega_{1}+\cdots+r_{n} \cdot \omega_{n}, r_{1} \cdot \rho_{1}+\cdots+r_{n} \cdot \rho_{n}\right) \\
& \leq r_{1} \cdot D_{K L}\left(\omega_{1}, \rho_{1}\right)+\cdots+r_{n} \cdot D_{K L}\left(\omega_{n}, \rho_{n}\right) .
\end{aligned}
$$

2.8.4 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:

$$
D_{K L}(\omega, \omega[1,0] \otimes \omega[0,1]) \approx 0.05
$$

2.8.5 This exercise demonstrates that Kullback-Leibler divergence is unbounded: it can become arbitrarily large. Consider the following two distributions on the set $\left\{0,1, \ldots, 2^{n}-1\right\}$ with $2^{n}$ elements, for $n \geq 1$.

$$
\omega=\frac{1}{2}|0\rangle+\frac{1}{2}|1\rangle \quad \rho_{n}=\frac{1}{2^{n}}|0\rangle+\cdots+\frac{1}{2^{n}}\left|2^{n}-1\right\rangle .
$$

Check that $D_{K L}\left(\omega, \rho_{n}\right)=n-1$.
2.8.6 In a similar way one can show that mutual information is unbounded (with multiple products). Let $v=f l i p\left(\frac{1}{2}\right)=\frac{1}{2}|0\rangle+\frac{1}{2}|1\rangle \in \mathcal{D}(\mathbf{2})$ and let $\rho_{n} \in \mathcal{D}\left(2^{n}\right)$, for $n \geq 1$, be defined by:

$$
\rho_{n}=\frac{1}{2}|\underbrace{0, \ldots, 0}_{n \text { times }}\rangle+\frac{1}{2}|\underbrace{1, \ldots, 1}_{n \text { times }}\rangle .
$$

1 Show that each $i$-th marginal $\mathcal{D}\left(\pi_{i}\right)\left(\rho_{n}\right)=\rho_{n}[0, \ldots, 0,1,0, \ldots, 0]$ is equal to $v$.

2 Deduce from Exercise 2.3.3 that:

$$
v^{n}=v \otimes \cdots \otimes v=\text { unif }_{2^{n}}=\frac{1}{2^{n}}|0, \ldots, 0\rangle+\cdots+\frac{1}{2^{n}}|1, \ldots, 1\rangle .
$$

3 Conclude, like in the previous exercise, that the $n$-fold mutual information $D_{K L}\left(\rho_{n}, v^{n}\right)$ is equal to $n-1$.
2.8.7 Use Jensen's inequality to prove what is known as the inequality of arithmetic and geometric means: for $r_{i}, a_{i} \in \mathbb{R}_{\geq 0}$ with $\sum_{i} r_{i}=1$,

$$
\sum_{i} r_{i} \cdot a_{i} \geq \prod_{i} a_{i}^{r_{i}} .
$$

### 2.9 Exchangeability for positions and elements

When we have a list of items in $X^{K}$, there are two ways to shuffle them: (1) we can keep the elements unchanged, but change their positions, 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 can be combined. 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: $X^{K} \rightarrow \mathcal{M}[K](X)$ and mat: $X^{K} \rightarrow S P(K)$ 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 2.9.1. Let $\vec{x}=\left(x_{0}, \ldots, x_{K-1}\right) \in X^{K}$ be a list of length $K$, with elements from an arbitrary set $X$.

1 A transposition is a bijective endofunction $t: \boldsymbol{K} \xlongequal{\approx} \boldsymbol{K}$ on the set of indices $\boldsymbol{K}=\{0, \ldots, K-1\}$. It gives rise to a bijection $\underline{t}:=t^{K}: X^{K} \xlongequal{\approx} X^{K}$ that reorders the elements in the list $\vec{x}$ as:

$$
t(\vec{x})=\left(x_{t(0)}, \ldots, x_{t(K-1)}\right) .
$$

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 $s: X \xlongequal{\cong} X$ on the set of elements in the list. It induces a bijection $s^{\star}: X^{K} \xlongequal{\leftrightharpoons} X^{K}$ given by:

$$
s^{\star}(\vec{x})=\mathcal{L}(s)(\vec{x})=\left(s\left(x_{0}\right), \ldots, s\left(x_{K-1}\right)\right.
$$

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 $f: X^{K} \rightarrow Y$ is called:

- stable under transposition when $f(\underline{t}(\vec{x}))=f(\vec{x})$, for all transpositions $t: \boldsymbol{K} \xlongequal{\cong} \boldsymbol{K}$ and sequences $\vec{x} \in X^{K}$; this is equivalent to: $f\left(\underline{t_{1}}(\vec{x})\right)=f\left(\underline{t_{2}}(\vec{x})\right)$, for all transpositions $t_{1}, t_{2}$, since the identity is a transposition itself;
- stable under substitution when $f$ satisfies $f\left(s^{\star}(\vec{x})\right)=f(\vec{x})$, for all substitutions $s: X \xlongequal{\cong} X$ and sequences $\vec{x} \in X^{K}$; this is again equivalent to $f\left(s_{1}^{\star}(\vec{x})\right)=f\left(s_{2}^{\star}(\vec{x})\right)$, for all substitutions $s_{1}, s_{2}$.
4 A joint distribution $\omega \in \mathcal{D}\left(X^{K}\right)$ is called transposition-exhangeable (resp. substitution-exchangeable) if, considered as a function $\omega: X^{K} \rightarrow[0,1]$, it is stable under transposition (resp. substitution).

We call a channel $c: Y \rightsquigarrow X^{K}$ transposition- / substitution-exchangeable if each distribution $c(y)$ 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: $X^{K} \rightarrow \mathcal{N}[K](X)$ from 1.35) is the archetypical function that is stable under transposition; similarly, the matching function mat: $X^{K} \rightarrow S P(X)$ from 1.21 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 2.9.2. Let $A$ be an object in an arbitrary category, with parallel endomaps $f_{1}, \ldots, f_{n}: A \rightarrow A$, for $n \geq 2$.

1 An equaliser of these parallel maps $f_{i}$ is a map $e: E \rightarrow A$ 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 $g: B \rightarrow A$ satisfies $f_{i} \circ g=f_{j} \circ g$ for $i, j$, then there is a unique map $h: B \rightarrow E$ with $e \circ h=g$. This is expressed in the following diagram, where the dashed map expresses uniqueness.


2 Dually, a coequaliser of these parallel maps $f_{i}$ is a morphism $c: A \rightarrow C$ with:

- $c \circ f_{i}=c \circ f_{j}$ for all $i, j$;
- if $g: A \rightarrow B$ satisfies $g \circ f_{i}=g \circ f_{j}$ for all $i, j$, then there is a unique $h: C \rightarrow B$ with $h \circ c=g$.


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 $E=\left\{a \in A \mid \forall i, j . f_{i}(a)=f_{j}(a)\right\}$. The coequaliser $A \rightarrow C$ is more difficult to describe and is best understood as the quotient (or collaps) of $A$ that forces all elements $f_{i}(a)$ to be equal, for each $a \in A$, see Exercise 2.9.2.

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 2.9.3.

1 The accumulation function acc: $X^{K} \rightarrow \mathcal{N}[K](X)$ is the coequaliser of all functions $\underline{t}: X^{K} \rightarrow X^{K}$ induced by the $K!$ transpositions $t: \boldsymbol{K} \xlongequal{\cong} \boldsymbol{K}$, in:

$$
\begin{equation*}
X^{K} \xrightarrow[\text { transpositions } \underline{ }]{ } X^{K} \xrightarrow{\text { acc }} \mathcal{N}[K](X) \tag{2.51}
\end{equation*}
$$

2 When the set $X$ has at least $K$ elements, the match function mat: $X^{K} \rightarrow$ $S P(K)$ is the coequaliser of the functions $s^{\star}: X^{K} \rightarrow X^{K}$ obtained from all the $|X|$ ! substitutions $s: X \xlongequal{\leftrightharpoons} X$, in:

$$
\begin{equation*}
X^{K} \xlongequal{\text { substitutions } s^{\star}} X^{K} \xrightarrow{\text { mat }} S P(K) \tag{2.52}
\end{equation*}
$$

Proof. 1 If $\ell^{\prime}$ is a transposition of $\ell \in X^{K}$, then $\operatorname{acc}\left(\ell^{\prime}\right)=\operatorname{acc}(\ell)$. Hence acc is stable under transposition. Let function $g: X^{K} \rightarrow Y$ be stable under transposition as well. For a multiset $\varphi=\sum_{1 \leq i \leq L} n_{i}\left|x_{i}\right\rangle \in \mathcal{N}[K](X)$, one can choose a list of elements, say $\vec{x}=\left\langle x_{1}, \ldots, x_{1}, \ldots, x_{L}, \ldots x_{L}\right\rangle \in X^{K}$, where $x_{i} \in X$ occurs $n_{i}$ many times. Then $\operatorname{acc}(\vec{x})=\varphi$, by construction. We can
now define $h(\varphi):=g(\vec{x})$. Since $g$ is stable under transposition, any list that accumulates to $\varphi$ gives the same outcome. Clearly, $h \circ \operatorname{acc}=g$. Moreover, $h$ is unique with this property.
2 Let $\ell^{\prime}$ be a substitution of $\ell \in X^{K}$. If elements at positions $i, j$ in $\ell$ are equal, then the elements at positions $i, j$ in $\ell^{\prime}$ are also equal. Thus $\operatorname{mat}(\ell)=\operatorname{mat}\left(\ell^{\prime}\right)$, making the match function stable under substitution. Next, let $g: X^{K} \rightarrow Y$ be stable under substitution. Given a set partition $P \in S P(K)$, say with $|P|=$ $n \leq K$ 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 $\operatorname{mat}(\vec{x})=P$. 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, $h \circ m a t=g$. 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 (2.27). Also for matching there is a probabilistic inverse mat ${ }^{\sim 1}$, see Exercise 2.4.12 We have not given this inverse a separate name. In general, $f \odot f^{\sim 1}=i d$ holds for probabilistic inverses $f^{\sim 1}$, see Definition 2.4.6. As a result, $f^{\sim 1} \odot f$ is a split idempotent with respect to channel composition $\odot$, that is, in the category Chan $=\operatorname{Chan}(\mathcal{D})$ 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 2.9.4. Fix a set $X$ and a number $K$.

1 Abbreviate transp $:=\operatorname{acc}^{\sim 1} \odot \operatorname{acc}: X^{K} \rightarrow X^{K}$. This is the transposition idempotent sending a list $\vec{x} \in X^{K}$ to all its transpositions:

2 Now assume that $X$ is finite; write subst $:=m a t^{\wedge 1} \odot$ mat $: X^{K} \leadsto X^{K}$. Then:

$$
\begin{equation*}
\operatorname{subst}(\vec{x}):=\sum_{s: X \underset{\cong}{\rightrightarrows} X} \frac{1}{|X|!}\left|s^{\star}(\vec{x})\right\rangle=\sum_{s: X \underset{\cong}{\rightrightarrows} X} \frac{1}{|X|!}\left|s\left(x_{0}\right), \ldots, s\left(x_{K-1}\right)\right\rangle . \tag{2.54}
\end{equation*}
$$

3 Let $\omega \in \mathcal{D}\left(X^{K}\right)$ be a joint distribution. The following points are equivalent:
(a) $\omega$ is transposition-exchangeble, see Definition 2.9.1 (4);
(b) $\mathcal{D}(\underline{t})(\omega)=\omega$, for all $t: K \xlongequal{\cong} \boldsymbol{K}$;
(c) transp $>=\omega=\omega$.

4 Similarly, the next three points are equivalent, for $\omega \in \mathcal{D}\left(X^{K}\right)$, when $X$ is finite:
（a）$\omega$ is substitution－exchangeble，see Definition 2．9．1］（4）；
（b） $\mathcal{D}\left(s^{\star}\right)(\omega)=\omega$ ，for all $s: X \xlongequal{\cong} X$ ；
（c）subst $>=\omega=\omega$ ．
Proof． 1 We have：

$$
\begin{aligned}
\operatorname{transp}(\vec{x})=\left(\operatorname{acc}^{\sim 1} \odot \operatorname{acc}\right)(\vec{x}) & =\sum_{\vec{y} \in \operatorname{acc} c^{-1}(\operatorname{acc}(\vec{x}))} \frac{1}{(\operatorname{acc}(\vec{x}))}|\vec{y}\rangle \\
& =\sum_{\vec{y} \text { is a permutation of } \vec{x}} \frac{1}{K!}|\vec{y}\rangle \\
& =\sum_{t: K^{\cong} \vec{\rightarrow} \boldsymbol{K}} \frac{1}{K!}\left|x_{t(0)}, \ldots, x_{t(K-1)}\right\rangle .
\end{aligned}
$$

2 Similarly．
3 The equation $\mathcal{D}(\underline{t})(\omega)(\vec{x})=\omega\left(\underline{t^{-1}}(\vec{x})\right)$ gives the equivalence a$) \quad \mathrm{b}$ in item（3）．Further，the pushforward transp $\gg \omega$ in（C）boils down to

$$
\begin{aligned}
& (\text { transp } \gg=\omega)(\vec{y})=\sum_{\vec{x} \in X^{K}} \operatorname{transp}(\vec{x})(\vec{y}) \cdot \omega(\vec{x})=\sum_{t: K \stackrel{\cong}{\rightrightarrows} K} \frac{1}{K!} \cdot \omega\left(\underline{\left.t^{-1}(\vec{y})\right)}\right. \\
& =\sum_{t: \boldsymbol{K} \underline{\underline{\underline{n}}}_{\boldsymbol{G}}} \frac{1}{K!} \cdot \mathcal{D}(\underline{t})(\omega)(\vec{y}) \text {. }
\end{aligned}
$$

Clearly，the latter equals $\omega(\vec{y})$ if（b）holds．This gives $(\mathrm{b}) \Rightarrow(\mathrm{c})$ ．For the reverse，take an arbitrary transposition $r: K \xlongequal{\leftrightharpoons} \boldsymbol{K}$ ．Then：

$$
\begin{aligned}
& \mathcal{D}(\underline{r})(\omega)(\vec{x})=\omega\left(\underline{r^{-1}}(\vec{x})\right) \stackrel{\text { 区- }}{=}(\text { transp } \gg=\omega)\left(\underline{r^{-1}}(\vec{x})\right) \\
& =\sum_{t: K \underset{\cong}{\cong} K} \frac{1}{K!} \cdot \mathcal{D}(\underline{t})(\omega)\left(\underline{r^{-1}}(\vec{x})\right)
\end{aligned}
$$

$$
\begin{aligned}
& =\sum_{t: K \underset{\underline{\underline{\imath}} \boldsymbol{K}}{ }} \frac{1}{K!} \cdot \mathcal{D}(\underline{r} \circ \underline{t})(\omega)(\vec{x}) \\
& =\sum_{t: \boldsymbol{K} \xlongequal{\underline{\star}} \boldsymbol{K}} \frac{1}{K!} \cdot \mathcal{D}(\underline{t})(\omega)(\vec{x}) \\
& =\stackrel{t: K \rightarrow K}{(\text { transp } \gg=\omega)(\vec{x})} \stackrel{\text { 区 }}{=} \omega(\vec{x}) \text {. }
\end{aligned}
$$

4 By the same reasoning．
The situation gets more interesting if we move to the category Chan $=$ $\operatorname{Chan}(\mathcal{D})$ of probabilistic channels．We are now going to instantiate the notion of（co）equaliser from Definition 2.9 .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 observations 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 2.9.5.

1 When we view the maps in Diagrams (2.51) and 2.52) 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 $\mathrm{acc}^{\sim 1}=$ arr is the equaliser of all transpositions (as deterministic channels), in:

$$
\begin{equation*}
\mathcal{N}[K](X) \xrightarrow{\text { acc-1 }^{-1}} X^{K} \xrightarrow{\text { transpositions } \underline{t}} X^{K} \tag{2.55}
\end{equation*}
$$

3 If a set $X$ has at least $K$ elements, then the probabilistic inverse match channel mat ${ }^{\sim 1}$ is the equaliser of all substitutions (as deterministic channels), in:

$$
\begin{equation*}
S P(K) \xrightarrow{\text { mat }{ }^{-1}} X^{K} \xrightarrow{\text { substitutions } s^{\star} \longrightarrow} X^{K} \tag{2.56}
\end{equation*}
$$

Proof. 1 The proofs of Proposition 2.9.3 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 $\rightarrow$ Chan is a left adjoint and thus preserves colimits, including coequalisers.)
2 For a transposition $t: \boldsymbol{K} \xlongequal{\cong} \boldsymbol{K}$ we have, for an arbitrary multiset $\varphi \in \mathcal{N}[K](X)$,

$$
\begin{aligned}
\left(\underline{t} \odot \operatorname{acc}^{\sim 1}\right)(\varphi) & =\mathcal{D}(\underline{t})\left(\operatorname{acc}^{\sim 1}(\varphi)\right) \\
\stackrel{[2.27}{=} & \sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{(\varphi)}|\underline{t}(\vec{x})\rangle \quad \text { since } \operatorname{acc}^{\sim 1}=\operatorname{arr} \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{(\varphi)}|\vec{x}\rangle=\operatorname{acc}^{\sim 1}(\varphi)
\end{aligned}
$$

Next, let $g: Y \leadsto X^{K}$ be a channel with $\underline{t} \odot g=g$ for each $t: K \xlongequal{\cong} \boldsymbol{K}$. This means $\mathcal{D}(t)(g(y))=g(y)$, for each $t$, and thus transp $\gg g(y)=g(y)$ by Lemma 2.9.4 (3), so that transp $\odot g=g$. We take $h=\operatorname{acc} \odot g: Y \leadsto \mathcal{N}[K](X)$. Then:

$$
\operatorname{acc}^{\sim 1} \odot h=\operatorname{acc}^{\sim 1} \odot \operatorname{acc} \odot g=\text { transp } \odot g=g .
$$

If also $k: Y \leftrightarrow \mathcal{N}[K](X)$ satisfies acc $^{\sim 1} \odot k=g$, then $k=h$ since:

$$
h=\operatorname{acc} \odot g=\operatorname{acc} \odot \operatorname{acc}^{\sim} \odot k=i d \odot k=k .
$$

3 Along the same lines, with mat in place of acc.
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 2.9.6. 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:


Explicitly:

- For a channel $f: Y \leadsto X^{K}$ with transp $\odot f=f$, the unique mediating channel $\bar{f}: Y \multimap \mathcal{N}[K](X)$ with arr $\odot \bar{f}=f$ is $\bar{f}:=\operatorname{acc} \odot f$.
- For a channel $g: X^{K} \leadsto Y$ with $g \odot$ transp $=g$, the unique mediating channel $\bar{g}: \mathcal{N}[K](X) \hookrightarrow Y$ with $\bar{g} \odot$ acc $=g$ is $\bar{g}:=g \odot \operatorname{arr}$.
2 Matching and its probabilistic inverse are coequaliser and equaliser of the substitution-idempotent subst and the identity:


For this we assume that the set $X$ has at least $K$ elements. In this case:

- For $f: Y \leadsto X^{K}$ with subst $\odot f=f$, one has $\bar{f}:=$ mat $\odot f$ uniquely satisfying mat ${ }^{\sim 1} \odot \bar{f}=f$;
- And for $g: X^{K} \rightsquigarrow Y$ with subst $\odot g=g$, the unique mediating map $\bar{g}$ with $\bar{g} \odot m a t=g$ is $\bar{g}:=g \odot m a t^{\sim 1}$.

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 2.9.7. Let $f: A \rightarrow A$ be a split idempotent in an arbitrary category. This means that $f$ can be written as $f=s \circ r$ where $r \circ s=i d$. Then:

1 s is the equaliser of $f, \mathrm{id}: A \rightrightarrows A$;
$2 r$ is their coequaliser.

In this situation $f$ is called a split idempotent, with (monic) section s and (epic) retraction $r$, in a situation:


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 $=\mathrm{acc}^{\sim 1} \odot$ acc and subst $=$ mat ${ }^{\sim} \odot$ mat .

Proof. Let $r: A \rightarrow B$ and $s: B \rightarrow A$ 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 $f \circ g=g=i d \circ g: C \rightarrow A$, then there is $\bar{g}:=r \circ g: C \rightarrow B$ satisfying:

- $s \circ \bar{g}=s \circ r \circ g=f \circ g=g$.
- if also $s \circ h=g$, then $h=r \circ s \circ h=r \circ g=\bar{g}$.

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 $: A \rightarrow C$, then $\bar{g}:==g \circ s: B \rightarrow C$ satisfies:

- $\bar{g} \circ r=g \circ s \circ r=g \circ f=g$.
- if also $h \circ r=g$, then $h=h \circ r \circ s=g \circ s=\bar{g}$.

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 (1.48), together with the muliplicity count and size count maps $m c$ 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 2.9.8. Fix a finite set $X$ and a number $K \geq 1$.
1 In the category Chan of probabilistic channels, the multiplicity count function mc: $\mathcal{N}[K](X) \rightarrow M P(K)$ is a coequaliser and its probabilistic inverse $\mathrm{mc}^{\sim 1}$ is an equaliser:

$$
\begin{equation*}
\operatorname{MP}(K) \xrightarrow[\rightarrow \rightarrow-1]{\substack{-1}} \mathcal{N}[K](X) \xrightarrow[\text { substitutions }]{ } \mathcal{N}[K](X) \xrightarrow[\rightarrow \rightarrow]{\mathrm{mc}} M P(K) \tag{2.60}
\end{equation*}
$$

These substitutions are of the form $\mathcal{M}(s): \mathcal{N}[K](X) \xlongequal{\cong} \mathcal{N}[K](X)$, for $s: X \xlongequal{\cong}$
X. The probabilistic inverse $m c^{\wedge 1}: M P(K) \rightsquigarrow \mathcal{N}[K](X)$ takes the following form, via Lemma 1.9.6.

$$
\begin{equation*}
m c^{\sim 1}(\alpha):=\sum_{\varphi \in \mathrm{mc}^{-1}(\alpha)} \frac{1}{\binom{X X}{\alpha}}|\varphi\rangle=\sum_{\varphi \in \mathrm{mc}^{-1}(\alpha)} \frac{\alpha \rrbracket \cdot(N-\|\alpha\|)!}{N!}|\varphi\rangle . \tag{2.61}
\end{equation*}
$$

2 Similarly, the size count function sc: $S P(K) \rightarrow M P(K)$ forms a coequaliser, and its probabilistic inverse $\mathrm{Sc}^{\sim 1}$ an equaliser, in the diagram of channels:

These transpositions in the middle are of the form $\mathcal{P}(\mathcal{P}(t)): S P(K) \stackrel{\cong}{\Rightarrow} S P(K)$, for $t:\{1, \ldots, K\} \stackrel{\cong}{\Rightarrow}\{1, \ldots, K\}$, see Exercise 1.5.4. The probabilistic inverse $\mathrm{sc}^{\sim 1}: M P(K) \rightsquigarrow S P(K)$ is of the following form.

$$
\begin{equation*}
\operatorname{sc}^{\sim 1}(\alpha):=\sum_{P \in \operatorname{sc}^{-1}(\alpha)} \frac{\alpha \rrbracket}{(\alpha)_{p}}|P\rangle . \tag{2.63}
\end{equation*}
$$

This is justified by Lemma 1.9.7.
The multiple substitution and transposition maps in 2.60 and 2.62 may be replaced by just two maps, like in 2.59, one of which is the identity.

The equaliser in (2.62) is used implicitly for what Pitman [151, 152] calls an exchangeable partition probability function (EPPF) of a distribution $\sigma \in$ $\mathcal{D}(S P(K))$ on set partitions. When such a distribution $\sigma$ is exchangeable, the EPPF arises via the equaliser $s c^{\sim 1}$ in:


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 2.9.9.

1 The following two rectangles commute.



For the diagram on the right we have to assume $|X| \geq K$.

2 The probabilistic inverses commute:


Proof. 1 We use Lemma 1.9 .8 for both rectangles. First, for $\varphi \in \mathcal{N}[K](X)$,

$$
\begin{aligned}
\left(\mathrm{sc}^{\sim 1} \odot m c\right)(\varphi) & \stackrel{\sqrt{2.63}}{=} \sum_{P \in c^{-1}(\operatorname{mc}(\varphi))} \frac{m c(\varphi) \rrbracket}{(\operatorname{mc}(\varphi))_{p}}|P\rangle \\
& \stackrel{[1.49}{=} \sum_{P \in \operatorname{sc}^{-1}(\operatorname{mc}(\varphi))} \sum_{\vec{x} \in \operatorname{mat}^{-1}(P) \cap \mathrm{acc}^{-1}(\varphi)} \frac{1}{\operatorname{mc}(\varphi) \rrbracket} \cdot \frac{\operatorname{mc}(\varphi) \rrbracket}{(\varphi)}|P\rangle \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{(\varphi)}|\operatorname{mat}(\vec{x})\rangle \\
& =\left(\operatorname{mat}^{2} \operatorname{acc}^{\sim 1}\right)(\varphi) .
\end{aligned}
$$

Similary, when $N=|X| \geq K$,

$$
\begin{aligned}
& \left(m c^{\sim 1} \odot s c\right)(P) \\
& \stackrel{\sqrt{2.61}}{=} \sum_{\varphi \in m c^{-1}(s c(P))} \frac{s c(P) \rrbracket \cdot(N-\|s c(P)\|)!}{N!}|\varphi\rangle \\
& =\sum_{\varphi \in m^{-1}(s c(P))} \sum_{\vec{x} \in \operatorname{mat}^{-1}(P) \operatorname{nacc}^{-1}(\varphi)} \frac{1}{s c(P) \rrbracket} \cdot \frac{s c(P) \rrbracket \cdot(N-|P|)!}{N!}|\varphi\rangle \\
& =\sum_{\vec{x} \in \operatorname{mat}^{-1}(P)} \frac{(N-|P|)!}{N!}|\operatorname{acc}(\vec{x})\rangle \\
& =\left(\operatorname{acc} \odot \operatorname{mat}^{\sim 1}\right)(P) .
\end{aligned}
$$

2 By unpacking the relevant definitions, for $\alpha \in \operatorname{MP}(K)$,

$$
\begin{aligned}
&\left(\mathrm{acc}^{\sim 1} \odot \mathrm{mc}^{\sim 1}\right)(\alpha)=\sum_{\varphi \in \mathrm{mc}^{-1}(\alpha)} \sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{\alpha \rrbracket \cdot(N-\|\alpha\|)!}{N!} \cdot \frac{1}{(\varphi)}|\vec{x}\rangle \\
& \frac{\| 1.49)}{=} \sum_{\varphi \in \mathrm{mc}^{-1}(\alpha)} \sum_{\vec{x} \in \mathrm{acc}^{-1}(\varphi)} \frac{\alpha \rrbracket \cdot(N-\|\alpha\|)!}{N!} \cdot \frac{1}{(\alpha)_{p}}|\vec{x}\rangle \\
&=\sum_{\vec{x} \in \operatorname{acc}^{-1}\left(m c^{-1}(\alpha)\right)} \frac{(N-\|\alpha\|)!}{N!} \cdot \frac{\alpha \rrbracket}{(\alpha)_{p}}|\vec{x}\rangle \\
&=\sum_{\vec{x} \in \operatorname{mat}^{-1}\left(\mathrm{sc}^{-1}(\alpha)\right)} \frac{(N-\|\alpha\|)!}{N!} \cdot \frac{\alpha \rrbracket}{(\alpha)_{p}}|\vec{x}\rangle \\
&=\sum_{P \in \mathrm{sc}^{-1}(\alpha)} \sum_{\vec{x} \in \operatorname{mat}^{-1}(P)} \frac{(N-|P|)!}{N!} \cdot \frac{\alpha \rrbracket}{(\alpha)_{p}}|\vec{x}\rangle \\
&=\left(\operatorname{mat}^{\sim 1} \odot \operatorname{sc}^{\sim 1}\right)(\alpha) .
\end{aligned}
$$

## Exercises

2.9.1 We have seen that the match function is stable under substitution. Prove the following slichtly more general statement. Let $X, Y$ be isomorphic sets, say via the isomorphism $f: X \xlongequal{\cong} Y$. Show that for a list $\ell \in X^{K}$,

$$
\operatorname{mat}(\ell)=\operatorname{mat}\left(f^{K}(\ell)\right)
$$

2.9.2 This exercises sketches how coequalisers can be obtained in the category Sets. Let $f, g: A \rightarrow B$ two arbitrary functions. Define the relation $R \subseteq B \times B$ as:

$$
R:=\{(f(a), g(a)) \mid a \in A\}
$$

Now let $\bar{R} \supseteq R$ be the least equivalenct relation containing $R$. Show that the quotient map $q: B \rightarrow B / \bar{R}$ is a coequaliser of $f, g$.
2.9.3 Recall the uniform projection channel unpr $[K]: X^{K} \leadsto X$ from Exercise 2.4.7. Check that it can be used to give an axiomatic definition of the frequentist learning map, as in [82], via the coequaliser property of accumulation:

2.9.4 Fix a set $X$ and a number $K \geq 1$.

1 Show that the composite $\mathcal{D}($ acc $) \circ \bigotimes: \mathcal{D}(X)^{K} \rightarrow \mathcal{D}(\mathcal{N}[K](X))$ is stable under transposition.
2 Deduce that there is a map from multisets over distributions to distributions over multisets, in:


In the next chapter, Sections 3.6-3.8 are devoted to this map.
2.9.5 Show that the transposition idempotent transp $=$ arr $\odot$ acc commutes with (big) tensors, as in the following diagram of channels.

2.9.6 Use Proposition 2.9.9 to show that the following diagram of split idempotents commutes.

2.9.7 Show also that the transposition idempotent transp interacts with zip as in the following diagram, in which zip occurs as deterministic channel.


Informally, this says that transposing a zip of already transposed inputs does not add anything.
2.9.8 Recall the triangular prism diagram (1.48) and define the equality count map ec: $X^{K} \rightarrow M P(K)$ as:

$$
\mathrm{ec}:=\mathrm{mc} \circ \mathrm{acc}=\mathrm{sc} \circ \mathrm{mat} .
$$

Develop a theory for ec, as we have done for acc and mat in this section, involving both transposition and substitution.

## 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. [103, 150, 159, 126] 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 pricture in 0.1 . The frequentist learning operation gives the probability distribution $\operatorname{Flrn}(v)$ associated with an urn $v$ and determines the probability of drawing a ball of colour $x$, see Remark 2.2.1. This probability is the ratio $\operatorname{Flrn}(v)(x)=\frac{v(x)}{\|v\|}$ of the number of balls $v(x)$ of colour $x$ and the total number of balls $\|v\|$ 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 $>=m n[K](\omega)=\omega$, see Theorem 3.3.3. 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 3.6 - 3.9 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 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, namely by considering draws of single balls. 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 3.3-3.5 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: $h g[K] \odot m n[K+L]=m n[K]$. 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 $\odot$ in this equation ensures that all probabilities in this statement interact in the appropriate manner.
The subsequent three sections 3.6-3.8deal 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 3.6 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 $\mathcal{D N}$. Similarly, there is a distributive law of lists over distributions, giving a monad $\mathcal{D} \mathcal{L}$. 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 | + | - | + |

### 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 $X^{K} \times Y^{K} \rightarrow(X \times Y)^{K}$ 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 $=\operatorname{acc}^{\sim 1}$. The multizip was first described in [80]. 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 $K$, we first look at all lists $\vec{x} \in X^{K}$ and $\vec{y} \in Y^{K}$ that accumulate to $\varphi$ and $\psi$, that is, with $\operatorname{acc}(\vec{x})=\varphi$ and $\operatorname{acc}(\vec{y})=\psi$. We then zip the two lists $\vec{x} \in X^{K}$ and $\vec{y} \in Y^{K}$ into a single list $\operatorname{zip}(\vec{x}, \vec{y}) \in$ $(X \times Y)^{K}$. Finally we apply accumulation acc: $(X \times Y)^{K} \rightarrow \mathcal{M}[K](X \times Y)$, giving a combined multiset $\operatorname{mzip}(\varphi, \psi) \in \mathcal{M}[K](X \times Y)$, still of size $K$. Diagrammatically we may describe mzip thus as the following composite.

$$
\begin{align*}
& \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \xrightarrow{\text { arrهarr }} \mathcal{D}\left(X^{K} \times Y^{K}\right) \\
& \cong \not \mathcal{D}^{(z i p)}  \tag{3.1}\\
& \mathcal{D}\left((X \times Y)^{K}\right) \xrightarrow[\mathcal{D}(\text { acc })]{ } \mathcal{D}(\mathcal{N}[K](X \times Y))
\end{align*}
$$

Definition 3.1.1. For sets $X, Y$ and a number $K$ we define the multizip channel $\operatorname{mzip}[K]: \mathcal{N}[K](X) \times \mathcal{N}[K](Y) \mapsto \mathcal{N}[K](X \times Y)$ as the composite:

$$
\begin{equation*}
\operatorname{mzip}[K]:=\operatorname{acc} \odot \operatorname{zip} \odot(\operatorname{arr} \otimes \operatorname{arr}) . \tag{3.2}
\end{equation*}
$$

This means that mzip $[K]$ makes the following rectangle commute.


Explicitly, for $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[K](Y)$,

$$
\begin{equation*}
\operatorname{mzip}[K](\varphi, \psi):=\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \sum_{\vec{y} \in \operatorname{acc}^{-1}(\psi)} \frac{1}{(\varphi) \cdot(\psi)}|\operatorname{acc}(\operatorname{zip}(\vec{x}, \vec{y}))\rangle . \tag{3.3}
\end{equation*}
$$

We drop the parameter $K$ in $\operatorname{mzip}[K]$ 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:

$$
\varphi=1|a\rangle+2|b\rangle \quad \text { and } \quad \psi=2|0\rangle+1|1\rangle
$$

Then:

$$
\mathbf{(} \varphi)=\binom{3}{1,2}=3 \quad(\psi)=\binom{3}{2,1}=3 .
$$

The sequences in $X^{3}$ and $Y^{3}$ that accumulate to $\varphi$ and $\psi$ are:

$$
\left\{\begin{array} { l } 
{ a , b , b } \\
{ b , a , b } \\
{ b , b , a }
\end{array} \quad \text { and } \quad \left\{\begin{array}{l}
0,0,1 \\
0,1,0 \\
1,0,0
\end{array}\right.\right.
$$

Zipping them together gives the following nine sequences in $(X \times Y)^{3}$.

| $(a, 0),(b, 0),(b, 1)$ | $(b, 0),(a, 0),(b, 1)$ | $(b, 0),(b, 0),(a, 1)$ |
| :--- | :--- | :--- |
| $(a, 0),(b, 1),(b, 0)$ | $(b, 0),(a, 1),(b, 0)$ | $(b, 0),(b, 1),(a, 0)$ |
| $(a, 1),(b, 0),(b, 0)$ | $(b, 1),(a, 0),(b, 0)$ | $(b, 1),(b, 0),(a, 0)$. |

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 $\frac{1}{(\varphi) \cdot(\psi)}=\frac{1}{9}$ gives:

$$
\begin{aligned}
& \operatorname{mzip}[3](1|a\rangle+2|b\rangle, 2|0\rangle+1|1\rangle) \\
& \left.\left.\left.\left.\quad=\frac{1}{3}|1| a, 1\right\rangle+2|b, 0\rangle\right\rangle+\frac{2}{3}|1| a, 0\right\rangle+1|b, 0\rangle+1|b, 1\rangle\right\rangle
\end{aligned}
$$

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: $\mathcal{N}[K](X) \times \mathcal{N}[K](Y) \leadsto \mathcal{N}[K](X \times Y)$, 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 $f: X \rightarrow U$ and $g: Y \rightarrow V$ the following diagram commutes.


2 For $\varphi \in \mathcal{N}[K](X)$ and $y \in Y$,

$$
\operatorname{mzip}(\varphi, K|y\rangle)=1|\varphi \otimes 1| y\rangle\rangle
$$

And similarly in symmetric form.
3 Multizip commutes with projections in the following sense.


This means that $\left\langle\mathcal{N}\left(\pi_{1}\right), \mathcal{N}\left(\pi_{2}\right)\right\rangle \odot$ mzip $=$ id, and thus that in the other order mzip $\odot\left\langle\mathcal{N}\left(\pi_{1}\right), \mathcal{N}\left(\pi_{2}\right)\right\rangle: \mathcal{N}[K](X \times Y) \rightsquigarrow \mathcal{N}[K](X \times Y)$ is a split idempotent channel.
4 Arrangement arr relates zip and mzip as in:


5 Multizip is associative, as given by:


Here we take associativity of $\times$ for granted.

Proof. 1 Easy, via the diagrammatic formulation (3.1), using naturality of arr (Exercise 2.4.11), of zip (Exercise 1.11.4), and of acc (Exercise 1.7.12).

2 Since:

$$
\begin{aligned}
\operatorname{mzip}(\varphi, K|y\rangle) & =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{\mathbf{( \varphi )}}|\operatorname{acc}(\operatorname{zip}(\vec{x},\langle y, \ldots, y\rangle))\rangle \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{\mathbf{( \varphi )}}\left|\operatorname{acc}\left(\overrightarrow{x_{i}, y}\right)\right\rangle \\
& \left.\left.=\sum_{\vec{x} \in \operatorname{aacc}^{-1}(\varphi)} \frac{1}{\mathbf{( \varphi )}}|\operatorname{acc}(\vec{x}) \otimes 1| y\right\rangle\right\rangle \\
& \left.\left.\left.\left.=\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{\mathbf{( \varphi )}}|\varphi \otimes 1| y\right\rangle\right\rangle=1|\varphi \otimes 1| y\right\rangle\right\rangle .
\end{aligned}
$$

3 By naturality of acc and zip:

$$
\begin{aligned}
& \mathcal{D}\left(\mathcal{N}\left(\pi_{1}\right)\right)(\operatorname{mzip}(\varphi, \psi)) \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \sum_{\vec{y} \in \operatorname{acc}^{-1}(\psi)} \frac{1}{(\varphi) \cdot(\psi)}\left|\mathcal{N}\left(\pi_{1}\right)(\operatorname{acc}(\operatorname{zip}(\vec{x}, \vec{y})))\right\rangle \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \sum_{\overrightarrow{\vec{e} \operatorname{acc}^{-1}(\psi)}} \frac{1}{(\varphi) \cdot(\psi)}\left|\operatorname{acc}\left(\left(\pi_{1}\right)^{K}(\operatorname{zip}(\vec{x}, \vec{y}))\right)\right\rangle \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \sum_{\vec{y} \operatorname{accc}^{-1}(\psi)} \frac{1}{(\varphi) \cdot(\psi)}|\operatorname{acc}(\vec{x})\rangle=\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{(\varphi)}|\varphi\rangle=1|\varphi\rangle .
\end{aligned}
$$

4 Directly from Exercise 2.9.7
5 Consider the following diagram chase, obtained by unpacking the (diagram-
matic) definition of mzip on the right-hand side.


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 2.28.

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:

$$
\operatorname{Flrn} \gg \operatorname{mzip}(\varphi, \psi)=\operatorname{Flrn}(\varphi \otimes \psi)=\operatorname{Flrn}(\varphi) \otimes \operatorname{Flrn}(\psi) .
$$

The second equation is known from Lemma 2.3.2 (5) but is added for completeness. In diagrammatic form the first equation reads:


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:

$$
(\text { Flrn } \gg \operatorname{mzip}(\varphi, \psi))(a, b)=\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \sum_{\vec{y} \in \operatorname{acc}^{-1}(\psi)} \frac{\operatorname{acc}(\operatorname{zip}(\vec{x}, \vec{y}))(a, b)}{K \cdot(\varphi) \cdot(\psi)}
$$

is the same as the probability:

$$
\operatorname{Flrn}(\varphi \otimes \psi)(a, b)=\frac{\varphi(a) \cdot \psi(a)}{K \cdot K}
$$

We reason informally, as follows. For arbitrary $\vec{x} \in \operatorname{acc}^{-1}(\varphi)$ and $\vec{y} \in \operatorname{acc}^{-1}(\psi)$ we need to find the fraction of occurrences $(a, b)$ in $\operatorname{zip}(\vec{x}, \vec{y})$. The fraction of
occurrences of $a$ in $\vec{x}$ is $\frac{\varphi(a)}{K}=\operatorname{Flrn}(\varphi)(a)$, and the fraction of occurrences of $b$ in $\vec{y}$ is $\frac{\psi(b)}{K}=F \operatorname{lrn}(\psi)(b)$. Hence the fraction of occurrences of $(a, b)$ in $\operatorname{zip}(\vec{x}, \vec{y})$ is $\operatorname{Flrn}(\varphi)(a) \cdot \operatorname{Flrn}(\psi)(b)=\operatorname{Flrn}(\varphi \otimes \psi)(a, b)$.

We continue with an alternative way to describe multizip ${ }^{11}$ It uses the concept of a coupling for natural multisets. We have described couplings of distributions $\sigma, \tau$ in Definition 2.3.8 (3) as a joint distribution that marginalises to $\sigma$ and to $\tau$. 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 $F$ : Sets $\rightarrow$ 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 $F\left(\pi_{1}\right)(c)=a$ and $F\left(\pi_{2}\right)(c)=b$.
Let's introduce the decouple map dcpl as:

$$
\begin{equation*}
\text { dcpl }:=\left(F(X \times Y) \xrightarrow{\left\langle F\left(\pi_{1}\right), F\left(\pi_{2}\right)\right\rangle} F(X) \times F(Y)\right) . \tag{3.4}
\end{equation*}
$$

A coupling of $(a, b)$ is then an element $c$ with $\operatorname{dcpl}(c)=(a, b)$, that is, with $c \in \operatorname{dcpl}^{-1}(a, b)$.

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 $\operatorname{mzip}(\varphi, \psi)$ is a distribution of couplings of $\varphi, \psi$. In Proposition 3.1.7 it will be shown that $\operatorname{mzip}(\varphi, \psi)$ 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: $X^{K} \times Y^{K} \xrightarrow{\cong}(X \times Y)^{K}$ from Exercise 1.3.7. As shown there, its inverse is the tuple $\left\langle\left(\pi_{1}\right)^{K},\left(\pi_{2}\right)^{K}\right\rangle:(X \times$ $Y)^{K} \xrightarrow{\cong} X^{K} \times Y^{K}$.

Lemma 3.1.6. Consider the setting of Definition 3.1.5
1 There is a commuting triangle of the form:


As a result, the decouple function dcpl: $\mathcal{N}[K](X \times Y) \rightarrow \mathcal{N}[K](X) \times \mathcal{N}[K](Y)$ is surjective.
${ }^{1}$ Developed together with Dario Stein.

2 There is the following equality of subsets of $\mathcal{N}[K](X \times Y)$.

$$
d c p I^{-1}(\varphi, \psi)=\left\{\operatorname{acc}(\operatorname{zip}(\vec{x}, \vec{y})) \mid \vec{x} \in \operatorname{acc}^{-1}(\varphi), \vec{y} \in \operatorname{acc}^{-1}(\psi)\right\} .
$$

In particular, the subset $\operatorname{dcpl}^{-1}(\varphi, \psi) \subseteq \mathcal{N}[K](X \times Y)$ is finite.
3 Further:

$$
\sum_{\chi \in d c p l^{-1}(\varphi, \psi)}(\chi)=(\varphi) \cdot(\psi)
$$

Proof. 1 Via the naturality of acc, see Exercise 1.7.12 for $\vec{x} \in \operatorname{acc}^{-1}(\varphi)$ and $\vec{y} \in \operatorname{acc}^{-1}(\psi)$,

$$
\begin{aligned}
& \operatorname{dcpl}(\operatorname{acc}(\operatorname{zip}(\vec{x}, \vec{y}))) \stackrel{\text { 3.4 }}{-}\left(\mathcal{N}\left(\pi_{1}\right)(\operatorname{acc}(\operatorname{zip}(\vec{x}, \vec{y}))), \mathcal{N}\left(\pi_{2}\right)(\operatorname{acc}(\operatorname{zip}(\vec{x}, \vec{y})))\right) \\
&=\left(\operatorname{acc}\left(\left(\pi_{1}\right)^{K}(\operatorname{zip}(\vec{x}, \vec{y}))\right), \operatorname{acc}\left(\left(\pi_{2}\right)^{K}(\operatorname{zip}(\vec{x}, \vec{y}))\right)\right) \\
&=\left(\operatorname{acc}\left(\pi_{1}(\vec{x}, \vec{y})\right), \operatorname{acc}\left(\pi_{2}(\vec{x}, \vec{y})\right)\right) \\
&=(\operatorname{acc}(\vec{x}), \operatorname{acc}(\vec{y})) \\
&=(\varphi, \psi) .
\end{aligned}
$$

This allows us to show that dcpl: $\mathcal{N}[K](X \times Y) \rightarrow \mathcal{N}[K](X) \times \mathcal{N}[K](Y)$ 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 $\operatorname{acc}(\vec{x})=\varphi$ and $\operatorname{acc}(\vec{y})=\psi$. The above argument shows that $\chi:=$ $\operatorname{acc}(\operatorname{zip}(\vec{x}, \vec{y})) \in \mathcal{N}[K](X \times Y)$ satisfies $d c p l(\chi)=(\operatorname{acc}(\vec{x}), \operatorname{acc}(\vec{y}))=(\varphi, \psi)$.

2 The previous point gives the inclusion (Э). For ( $\subseteq$ ), let $\chi \in \operatorname{dcpl}^{-1}(\varphi, \psi)$ be given. There is a sequence $\vec{z} \in(X \times Y)^{K}$ with $\operatorname{acc}(\vec{z})=\chi$. Write $\vec{z}=\operatorname{zip}(\vec{x}, \vec{y})$, for $\vec{x}=\left(\pi_{1}\right)^{K}(\vec{z})$ and $\vec{y}=\left(\pi_{2}\right)^{K}(\vec{z})$. We claim that $\operatorname{acc}(\vec{x})=\varphi$ and $\operatorname{acc}(\vec{y})=\psi$. We elaborate only the first equation, since the second one is obtained in the same way. By assumption, $\chi$ is coupling of $\varphi, \psi$, so:

$$
\varphi=\mathcal{N}\left(\pi_{1}\right)(\chi)=\mathcal{N}\left(\pi_{1}\right)(\operatorname{acc}(z i p(\vec{x}, \vec{y})))=\operatorname{acc}(\vec{x})
$$

The last equation is obtained as in the previous point.

3 By:

$$
\begin{aligned}
& \sum_{\chi \in d c p 1^{-1}(\varphi, \psi)}(\chi) \\
& =\sum_{\chi \in \operatorname{dcpl}^{-1}(\varphi, \psi)}\left|\operatorname{acc}^{-1}(\chi)\right| \quad \text { by Proposition 1.7.2 } \\
& =\left|\bigcup_{\chi \in d c p l^{-1}(\varphi, \psi)} \operatorname{acc}^{-1}(\chi)\right| \quad \text { since } \bigcup \text { is a disjoint union } \\
& =\left|(d c p l \circ \operatorname{acc})^{-1}(\varphi, \psi)\right| \\
& \left.=\mid(\text { dcpl } \circ \operatorname{acc} \circ \operatorname{zip})^{-1}\right)(\varphi, \psi) \mid \quad \text { since zip is an isomorphism } \\
& \stackrel{3.5}{-}\left|(\operatorname{acc} \times \operatorname{acc})^{-1}(\varphi, \psi)\right| \\
& =\left|(a c c)^{-1}(\varphi) \times(a c c)^{-1}(\psi)\right| \\
& =\left|(\operatorname{acc})^{-1}(\varphi)\right| \cdot\left|(\operatorname{acc})^{-1}(\psi)\right| \\
& =(\varphi) \cdot(\psi) \\
& \text { by Proposition 1.7.2 again. }
\end{aligned}
$$

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:

$$
\begin{equation*}
\operatorname{mzip}(\varphi, \psi)=\sum_{\chi \in d c p l^{-1}(\varphi, \psi)} \frac{(\chi)}{\mathbf{( \varphi ) \cdot ( \psi )}}|\chi\rangle \tag{3.6}
\end{equation*}
$$

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.

$$
\begin{aligned}
\sum_{\chi \in d c p l^{-1}(\varphi, \psi)} \frac{(\chi)}{(\varphi) \cdot(\psi)}|\chi\rangle & =\sum_{\vec{z} \in(X \times Y)^{K}} \frac{1}{(\varphi) \cdot(\psi)}|\operatorname{acc}(\vec{z})\rangle \\
& =\sum_{\vec{x} \in X^{K}} \sum_{\overrightarrow{\vec{y}} \in Y^{K}} \frac{1}{(\varphi) \cdot(\psi)}|\operatorname{acc}(z i p(\vec{x}, \vec{y}))\rangle \\
& =\operatorname{mzip}(\varphi, \psi) .
\end{aligned}
$$

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:

$$
\begin{aligned}
\mathcal{N}[K](X) \times \mathcal{N}[L](X) \xrightarrow{\text { arr®arr }} \mathcal{D}( & \left.X^{K} \times X^{L}\right) \\
& \cong \nmid \mathcal{D}(+) \\
& \mathcal{D}\left(X^{K+L}\right) \xrightarrow[\mathcal{D}(a c c)]{ } \mathcal{D}(\mathcal{N}[K+L](X))
\end{aligned}
$$

It is easy to see that this yields addition of multisets, as a deterministic channel.
In Section 2.9 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 $S P(K) \times S P(K) \rightsquigarrow S P(K)$ in:

$$
S P(K) \times S P(L) \xrightarrow{\text { mat }^{-1} \otimes \otimes m a t^{-1}} X^{K} \times Y^{K} \xrightarrow{\text { zip }} \underset{\cong}{o \rightarrow}(X \times Y)^{K} \xrightarrow{\text { mat }} S P(K)
$$

is deterministic; it sends set partitions $P, Q \in S P(K)$ to their 'intersection':

$$
P \& Q:=\{B \cap C \mid B \in P, C \in Q, B \cap C \neq \emptyset\} .
$$

Proof. Let $D \in \operatorname{mat}(\operatorname{zip}(\vec{x}, \vec{y}))$ be a block, for sequences $\vec{x} \in \operatorname{mat}^{-1}(P)$ and $\vec{y} \in \operatorname{mat}^{-1}(Q)$, say $D=\left\{i_{1}, \ldots, i_{n}\right\}$ for $n \geq 2$. Then $\operatorname{zip}(\vec{x}, \vec{y})_{i_{j}}=\operatorname{zip}(\vec{x}, \vec{y})_{i_{k}}$ 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 $\operatorname{mat}(\operatorname{zip}(\vec{x}, \vec{y}))$. The argument does not depend on the choice over $\vec{x}$ or $\vec{y}$, and thus gives a deterministic channel.

We conclude with several useful observations about accumulation in two dimensions.

Lemma 3.1.9. Fix a number $K \in \mathbb{N}$ and a set $X$.
1 We can mix K-ary and 2-ary accumulation in the following way.


2 When we generalise from 2 to $L \geq 2$ we get:


Proof. 1 Commutation of the diagram is 'obvious', so we provide only an exemplary proof. The two paths in the diagram yield the same outcomes in:

$$
\begin{aligned}
& (+\circ(\operatorname{acc}[K] \times \operatorname{acc}[K]))([a, b, c, a, c],[b, b, a, a, c]) \\
& =(2|a\rangle+1|b\rangle+2|c\rangle)+(2|a\rangle+2|b\rangle+1|c\rangle) \\
& =4|a\rangle+3|b\rangle+3|c\rangle . \\
& \left(+\circ \operatorname{acc}[2]^{K} \circ z i p\right)([a, b, c, a, c],[b, b, a, a, c]) \\
& =\left(+\circ \operatorname{acc}[2]^{K}\right)([(a, b),(b, b),(c, a),(a, a),(c, c)]) \\
& =+([1|a\rangle+1|b\rangle, 2|b\rangle, 1|a\rangle+1|c\rangle, 2|a\rangle, 2|c\rangle]) \\
& =4|a\rangle+3|b\rangle+3|c\rangle .
\end{aligned}
$$

## 2 Similarly.

We now transfer these results to multizip.
Proposition 3.1.10. In binary form one has:


More generally, we have for $L \geq 2$,


The map mzip $_{L}$ is the L-ary multizip, obtained via:

$$
\operatorname{mzip}_{2}:=\text { mzip } \quad \text { and } \quad \text { mzip }_{L+1}:=\text { mzip } \odot\left(m_{z i p} \otimes \text { id }\right)
$$

Via the associativity of Proposition 3.1.3 (5) the actual arrangement of these multiple multizips does not matter.

Proof. We only do the binary case, via an equational proof:

```
\(\mathcal{D}(f l a t) \circ \mathcal{D}(\mathcal{N}[K](\operatorname{acc}[2])) \circ\) mzip
    \(\stackrel{3.1}{-} \mathcal{D}(\) flat \() \circ \mathcal{D}(\mathcal{N}[K](\operatorname{acc}[2])) \circ \mathcal{D}(\operatorname{acc}[K]) \circ \mathcal{D}(\) zip \() \circ(\operatorname{arr} \otimes \operatorname{arr})\)
    \(=\mathcal{D}(\) flat \() \circ \mathcal{D}(\operatorname{acc}[K]) \circ \mathcal{D}\left(\operatorname{acc}[2]^{K}\right) \circ \mathcal{D}(\) zip \() \circ(\operatorname{arr} \otimes \operatorname{arr})\)
            by naturality of acc \([K]: X^{K} \rightarrow \mathcal{N}[K](X)\)
    \(=\mathcal{D}(+) \circ \mathcal{D}\left(\operatorname{acc}[2]^{K}\right) \circ \mathcal{D}(\) zip \() \circ(\operatorname{arr} \otimes \operatorname{arr}) \quad\) by Exercise 1.8.3
    \(=\mathcal{D}(+) \circ \mathcal{D}(\operatorname{acc}[K] \times \operatorname{acc}[K]) \circ(\operatorname{arr} \otimes \operatorname{arr}) \quad\) by Lemma3.1.9 1
    \(=\mathcal{D}(+) \circ((\mathcal{D}(\operatorname{acc}[K]) \circ \operatorname{arr}) \otimes(\mathcal{D}(\operatorname{acc}[K]) \circ \operatorname{arr}))\)
    \(\stackrel{[2.28}{=} \mathcal{D}(+) \circ(\) unit \(\otimes\) unit)
    \(=\mathcal{D}(+) \circ\) unit
    \(=\) unit \(\circ+\).
```


## Exercises

3.1.1 Show that:

$$
\begin{aligned}
& \operatorname{mzip}[4](1|a\rangle+2|b\rangle+1|c\rangle, 3|0\rangle+1|1\rangle) \\
& \left.\left.=\frac{1}{4}|1| a, 1\right\rangle+2|b, 0\rangle+1|c, 0\rangle\right\rangle \\
& \left.\left.\quad+\frac{1}{2}|1| a, 0\right\rangle+1|b, 0\rangle+1|b, 1\rangle+1|c, 0\rangle\right\rangle \\
& \left.\left.\quad+\frac{1}{4}|1| a, 0\right\rangle+2|b, 0\rangle+1|c, 1\rangle\right\rangle
\end{aligned}
$$

3.1.2 Show in the context of the previous exercise that:

$$
\begin{aligned}
& \text { Flrn } \gg \operatorname{mzip}[4](1|a\rangle+2|b\rangle+1|c\rangle, 3|0\rangle+1|1\rangle) \\
& =\frac{3}{16}|a, 0\rangle+\frac{1}{16}|a, 1\rangle+\frac{3}{8}|b, 0\rangle+\frac{1}{8}|b, 1\rangle+\frac{3}{16}|c, 0\rangle+\frac{1}{16}|c, 1\rangle .
\end{aligned}
$$

Compute also: $\operatorname{Flrn}(1|a\rangle+2|b\rangle+1|c\rangle) \otimes \operatorname{Flrn}(3|0\rangle+1|1\rangle)$, 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 $\operatorname{zip}(\vec{x}, \vec{y}) \in(X \times Y)^{K}$.
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 $\varphi$ and $\psi$.
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 $\operatorname{Flrn}(\tau) \in \mathcal{D}(X \times Y)$ 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.


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:


Hint: Take sequences $[a, b, b],[0,0,1]$ and re-use Example 3.1.2
3.1.7 Compute the intersection of the two set partitions of 10 below.

$$
\{\{1,2,3,4,5,9,10\},\{6,7,8\}\} \quad\{\{1,6\},\{2,3,10\},\{4,8,9\},\{5\},\{7\}\} .
$$

### 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 $D D$ and draw-add $D A$, 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: $X^{K+1} \rightarrow X^{K}$ is defined as:

$$
\begin{equation*}
\operatorname{PD}\left(x_{1}, \ldots, x_{K+1}\right):=\sum_{1 \leq i \leq K+1} \frac{1}{K+1}\left|x_{1}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{K+1}\right\rangle . \tag{3.7}
\end{equation*}
$$

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 $D D: \mathcal{N}[K+1](X) \leadsto \mathcal{N}[K](X)$ is defined on a multiset $\psi \in \mathcal{N}[K+1](X)$ as:

$$
\begin{align*}
D D(\psi) & \left.\left.:=\sum_{x \in \operatorname{supp}(\psi)} \frac{\psi(x)}{K+1}|\psi-1| x\right\rangle\right\rangle \\
& \left.\left.=\sum_{x \in \operatorname{supp}(\psi)} F \operatorname{lrn}(\psi)(x)|\psi-1| x\right\rangle\right\rangle . \tag{3.8}
\end{align*}
$$

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.


As a result, projection-delete commutes with the transposition channel transp $=$ arr $\odot$ acc from (2.53), in:

$$
\text { transp } \odot P D=P D \odot \text { transp }
$$

Proof. For $\vec{x}=\left(x_{1}, \ldots, x_{K+1}\right) \in X^{K+1}$ with $\operatorname{acc}(\vec{x})=\psi$ we have:

$$
\begin{aligned}
(\operatorname{acc} \odot P D)(\vec{x}) & =\sum_{1 \leq i \leq K+1} \frac{1}{K+1}\left|\operatorname{acc}\left(x_{1}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{K+1}\right)\right\rangle \\
& \left.\left.=\sum_{1 \leq i \leq K+1} \frac{1}{K+1}|\psi-1| x_{i}\right\rangle\right\rangle \\
& \left.\left.=\sum_{y \in \operatorname{supp}(\varphi)} \frac{\psi(y)}{K+1}|\psi-1| y\right\rangle\right\rangle=D D(\psi)=(D D \odot \operatorname{acc})(\vec{x}) .
\end{aligned}
$$

Similarly, for $\psi \in \mathcal{N}[K+1](X)$,

$$
\begin{aligned}
(\operatorname{arr} \odot D D)(\psi) & =\sum_{\vec{x} \in X^{K}} \sum_{y \in \operatorname{supp}(\psi)} \operatorname{arr}(\psi-1|y\rangle)(\vec{x}) \cdot \frac{\psi(y)}{K+1}|\vec{x}\rangle \\
& =\sum_{y \in \operatorname{supp}(\psi)} \sum_{\vec{x} \in \operatorname{aacc}^{-1}(\psi-1|y\rangle)} \frac{1}{(\psi-1|y\rangle)} \cdot \frac{\psi(y)}{K+1}|\vec{x}\rangle \\
& =\sum_{y \in \operatorname{supp}(\psi)} \sum_{\vec{x} \in \operatorname{aacc}^{-1}(\psi-1|y\rangle)} \frac{(\psi(y)-1)!\cdot \prod_{z \neq y} \psi(z)!}{K!} \cdot \frac{\psi(y)}{K+1}|\vec{x}\rangle \\
& =\sum_{y \in \operatorname{supp}(\psi)} \sum_{\vec{x} \in \operatorname{acc}^{-1}(\psi-1|y\rangle)} \frac{\psi \rrbracket}{(K+1)!}|\vec{x}\rangle \\
& =\sum_{\vec{z} \in \operatorname{acc}-1} \sum_{1 \leq i \leq K} \frac{1}{(\psi)} \cdot \frac{1}{K+1}\left|z_{1}, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{K}\right\rangle \\
& =\sum_{\vec{x} \in X^{K}} \sum_{\vec{z} \in X^{K+1}} \operatorname{arr}(\psi)(\vec{z}) \cdot P D(\vec{z})(\vec{x})|\vec{x}\rangle \\
& =(P D \odot \operatorname{arr})(\psi) .
\end{aligned}
$$

We now use commutation of these two rectangles in:

$$
\begin{aligned}
\text { transp } \odot P D=\operatorname{arr} \odot \operatorname{acc} \odot P D & =\operatorname{arr} \odot D D \odot \operatorname{acc} \\
& =P D \odot \operatorname{arr} \odot \operatorname{acc}=P D \odot \operatorname{transp} .
\end{aligned}
$$

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 $\left(x_{1}, \ldots, x_{K}\right)$ 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: $X^{K} \leadsto X^{K+1}$ is:

$$
\begin{align*}
& P A\left(x_{1}, \ldots, x_{K}\right) \\
& :=\sum_{1 \leq i \leq K} \sum_{1 \leq j \leq K+1} \frac{1}{K(K+1)}\left|x_{1}, \ldots, x_{j-1}, x_{i}, x_{j}, \ldots, x_{K}\right\rangle . \tag{3.9}
\end{align*}
$$

2 The draw-add channel $D A: \mathcal{N}[K](X) \rightsquigarrow \mathcal{N}[K+1](X)$ is defined on a multiset $\varphi \in \mathcal{N}[K](X)$ as:

$$
\begin{align*}
D A(\varphi) & \left.\left.:=\sum_{x \in \operatorname{supp}(\varphi)} \frac{\varphi(x)}{K}|\varphi+1| x\right\rangle\right\rangle  \tag{3.10}\\
& \left.\left.=\sum_{x \in \operatorname{supp}(\varphi)} \operatorname{Flrn}(\varphi)(x)|\varphi+1| x\right\rangle\right\rangle .
\end{align*}
$$

For instance,

$$
\begin{aligned}
\left.P A(a, b, c)=\frac{1}{6} \right\rvert\, & |a, a, b, c\rangle+\frac{1}{12}|a, b, a, c\rangle+\frac{1}{6}|a, b, b, c\rangle \\
& +\frac{1}{12}|a, b, c, a\rangle+\frac{1}{12}|a, b, c, b\rangle+\frac{1}{6}|a, b, c, c\rangle \\
& +\frac{1}{12}|a, c, b, c\rangle+\frac{1}{12}|b, a, b, c\rangle+\frac{1}{12}|c, a, b, c\rangle .
\end{aligned}
$$

Lemma 3.2.4. The projection-add channel commutes with transposition:

$$
\text { transp } \odot P A=P A \odot \text { transp }
$$

In addition, both projection- and draw-add maps commute with accumulation and aggregation:


Proof. The equation transp $\odot P A=P A \odot$ 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 projectadd to a permutation of the original sequence $x_{1}, \ldots, x_{K}$.

For commutation with accumulation, let $\vec{x} \in X^{K}$,

$$
\begin{aligned}
(\operatorname{acc} \odot P A)(\vec{x}) & =\sum_{1 \leq i \leq K} \sum_{1 \leq j \leq K+1} \frac{1}{K(K+1)}\left|\operatorname{acc}\left(x_{1}, \ldots, x_{j-1}, x_{i}, x_{j}, \ldots, x_{K}\right)\right\rangle \\
& \left.\left.=\sum_{1 \leq i \leq K} \frac{1}{K}\left|\operatorname{acc}\left(x_{1}, \ldots, x_{K}\right)+1\right| x_{i}\right\rangle\right\rangle \\
& \left.\left.=\sum_{y \in \operatorname{supp}(\operatorname{acc}(\vec{x}))} \frac{\operatorname{acc}(\vec{x})(y)}{K}|\operatorname{acc}(\vec{x})+1| y\right\rangle\right\rangle \\
& =(D A \odot \operatorname{acc})(\vec{x}) .
\end{aligned}
$$

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:

$$
\begin{aligned}
(P A \odot a r r) \circ \operatorname{acc}=P A \odot t r a n s p & =\operatorname{transp} \odot P A \\
& =\operatorname{arr} \odot \operatorname{acc} \odot P A \\
& =\operatorname{arr} \odot D A \odot \operatorname{acc} \quad \text { as just shown } \\
& =(\operatorname{arr} \odot D A) \circ \text { acc. }
\end{aligned}
$$

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 $D D$ and $D A$ acting on multisets.

Remark 3.2.5. The delete and add channels are not each other's inverses, in a situation:


For instance:

$$
\begin{aligned}
& \left.\left.\left.\left.D D(3|a\rangle+1|b\rangle)=\frac{3}{4}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{4}|3| a\right\rangle\right\rangle \\
& \left.\left.\left.\left.D A(2|a\rangle+1|b\rangle)=\frac{2}{3}|3| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{3}|2| a\right\rangle+2|b\rangle\right\rangle .
\end{aligned}
$$

In a next step we see that neither $D A \odot D D$ nor $D D \odot D A$ is the identity.

$$
\begin{aligned}
& (D A \odot D D)(3|a\rangle+1|b\rangle) \\
& \left.\left.\left.\left.=D A \gg\left(\frac{3}{4}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{4}|3| a\right\rangle\right\rangle\right) \\
& \left.\left.\left.\left.\left.\left.=\frac{3}{4} \cdot\left(\frac{2}{3}|3| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{3}|2| a\right\rangle+2|b\rangle\right\rangle\right)+\frac{1}{4} \cdot 1|4| a\right\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.=\frac{1}{2}|3| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{4}|2| a\right\rangle+2|b\rangle\right\rangle+\frac{1}{4}|4| a\right\rangle\right\rangle \\
& (D D \odot D A)(2|a\rangle+1|b\rangle) \\
& \left.\left.\left.\left.=D D \gg\left(\frac{2}{3}|3| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{3}|2| a\right\rangle+2|b\rangle\right\rangle\right) \\
& \left.\left.\left.\left.\left.\left.\left.\left.=\frac{2}{3} \cdot\left(\frac{3}{4}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{4}|3| a\right\rangle\right\rangle\right)+\frac{1}{3} \cdot\left(\frac{1}{2}|1| a\right\rangle+2|b\rangle\right\rangle+\frac{1}{2}|2| a\right\rangle+1|b\rangle\right\rangle\right) \\
& \left.\left.\left.\left.\left.\left.\left.\left.=\frac{1}{2}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{6}|3| a\right\rangle\right\rangle+\frac{1}{6}|1| a\right\rangle+2|b\rangle\right\rangle+\frac{1}{6}|2| a\right\rangle+1|b\rangle\right\rangle .
\end{aligned}
$$

Since the drawing of a single ball from an urn takes the multiplicities into account, the urns after $D D$ and $D A$ have the same frequentist distribution as before, but only if we interpret 'after' as channel composition $\odot$. This is the content of the following basic result.

Theorem 3.2.6. One has both:

$$
\text { Flrn } \odot D D=\text { Flrn } \quad \text { and } \quad \text { Flrn } \odot D A=\text { Flrn } .
$$

Equivalently, the following two triangles of channels commute.


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:

$$
\begin{aligned}
(F l r n \odot D D)(\psi)(y) & =\sum_{\varphi \in \mathcal{N}[K](X)} D D(\psi)(\varphi) \cdot \operatorname{Flrn}(\varphi)(y) \\
& \stackrel{\text { 3.8: }}{=} \sum_{x \in \operatorname{supp}(\psi)} \frac{\psi(x)}{K+1} \cdot \operatorname{Flrn}(\psi-1|x\rangle)(y) \\
& =\frac{\psi(y)}{K+1} \cdot \frac{\psi(y)-1}{K}+\sum_{x \neq y} \frac{\psi(x)}{K+1} \cdot \frac{\psi(y)}{K} \\
& =\frac{\psi(y)}{K(K+1)} \cdot\left(\psi(y)-1+\sum_{x \neq y} \psi(x)\right) \\
& =\frac{\psi(y)}{K(K+1)} \cdot\left(\left(\sum_{x} \psi(x)\right)-1\right) \\
& =\frac{\psi(y)}{K(K+1)} \cdot((K+1)-1)=\frac{\psi(y)}{K+1}=\operatorname{Flrn}(\psi)(y) .
\end{aligned}
$$

Similarly, for $\varphi \in \mathcal{N}[K](X)$, where $K>0$, and $y \in X$,

$$
\begin{aligned}
(F l r n \odot D A)(\varphi)(y) & \stackrel{\boxed{3.10}}{=} \sum_{x \in \operatorname{supp}(\varphi)} \frac{\varphi(x)}{K} \cdot \operatorname{Flrn}(\varphi+1|x\rangle)(y) \\
& =\frac{\varphi(y)}{K} \cdot \frac{\varphi(y)+1}{K+1}+\sum_{x \neq y} \frac{\varphi(x)}{K} \cdot \frac{\varphi(y)}{K+1} \\
& =\frac{\varphi(y)}{K(K+1)} \cdot\left(\varphi(y)+1+\sum_{x \neq y} \varphi(x)\right) \\
& =\frac{\varphi(y)}{K(K+1)} \cdot(K+1)=\frac{\varphi(y)}{K}=\operatorname{Flrn}(\varphi)(y) .
\end{aligned}
$$

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 3.4.3 and 3.5.3. The iteration of draws leads to subtraction and addition of the draws $\varphi$ 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:


On $v \in \mathcal{N}[L+K](X)$ and $\psi \in \mathcal{N}[L](X)$ they satisfy:

$$
\begin{equation*}
D D^{K}(v)=\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{L+K}{K}}|v-\varphi\rangle \quad D A^{K}(\psi)=\sum_{\varphi \leq K \psi} \frac{\left(\binom{\psi}{\varphi}\right)}{\left(\binom{L}{K}\right)}|\psi+\varphi\rangle . \tag{3.11}
\end{equation*}
$$

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 $\varphi$ in $\mathcal{N}[0](X)$ is the empty multiset $\mathbf{0}$, so that the sums in (3.11) are equal to $1|v\rangle$ and $1|\psi\rangle$, and thus equal to the zero-th iteration $D D^{0}(v):=1|v\rangle$ and $D A^{0}(\psi):=1|\psi\rangle$.

For the induction steps we make extensive use of the equations in Exercise 1.8.12. In those cases we shall put ' $E$ ' above the equation. We start with the induction step for draw-delete. For $v \in \mathcal{N}[L+K+1](X)$,

$$
\begin{aligned}
D D^{K+1}(v) & =D D^{K} \gg=D D(v) \\
& \left.\left.=D D^{K} \gg=\left(\sum_{x \in \operatorname{supp}(v)} \frac{v(x)}{L+K+1}|v-1| x\right\rangle\right\rangle\right) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in \operatorname{supp}(v)} \frac{v(x)}{L+K+1} \cdot D D^{K}(v-1|x\rangle)(\varphi)|\varphi\rangle \\
& \left.\left.\stackrel{(\mathrm{IH})}{=} \sum_{x \in \operatorname{supp}(v)} \sum_{\varphi \leq K v-1|x\rangle} \frac{v(x)}{L+K+1} \cdot \frac{\binom{v-1|x\rangle}{\varphi}}{\binom{L+K}{K}} \right\rvert\,(v-1|x\rangle)-\varphi\right\rangle \\
& \left.\left.\stackrel{(\mathrm{E})}{=} \sum_{x \in \operatorname{supp}(v)} \sum_{\varphi \leq K^{v} v-1|x\rangle} \frac{\varphi(x)+1}{K+1} \cdot \frac{\binom{v}{\varphi+1|x\rangle}}{\binom{L+K+1}{K+1}} \right\rvert\, v-(\varphi+1|x\rangle)\right\rangle \\
& =\sum_{\varphi \leq K+1} \sum_{x \in \operatorname{supp}(v)} \frac{\varphi(x)}{K+1} \cdot \frac{\binom{v}{\varphi}}{\binom{L+K+1}{K+1}}|v-\varphi\rangle \\
& =\sum_{\varphi \leq K+1 v} \frac{\binom{v}{\varphi}}{\binom{L+K+1}{K+1}}|v-\varphi\rangle, \quad \text { since }\|\varphi\|=K+1 .
\end{aligned}
$$

We reason basically in the same way for draw-add, and now also use Exer-
cise 1.8.11. For $\psi \in \mathcal{N}[L](X)$,

$$
\begin{aligned}
D A^{K+1}(\psi) & \left.\left.=D A^{K} \gg=\left(\sum_{x \in \operatorname{supp}(\psi)} \frac{\psi(x)}{L}|\psi+1| x\right\rangle\right\rangle\right) \\
& \left.\left.\stackrel{(\mathrm{IH})}{=} \sum_{x \in \operatorname{supp}(\psi)} \sum_{\varphi \leq{ }_{K} \psi+1|x\rangle} \frac{\psi(x)}{L} \cdot \frac{\left(\binom{\psi+1|x\rangle}{\varphi}\right)}{\left(\binom{L+1}{K}\right)} \right\rvert\,(\psi+1|x\rangle)+\varphi\right\rangle \\
& \left.\left.\stackrel{(\mathrm{E})}{=} \sum_{x \in \operatorname{supp}(\psi)} \sum_{\varphi \leq K_{K} \psi+1|x\rangle} \frac{\varphi(x)+1}{K+1} \cdot \frac{\left(\binom{\psi}{\varphi+1|x\rangle}\right)}{\left(\binom{L}{K+1}\right)} \right\rvert\, \psi+(\varphi+1|x\rangle)\right\rangle \\
& =\sum_{\varphi \leq K+1} \sum_{x \in \operatorname{supp}(\psi)} \frac{\varphi(x)}{K+1} \cdot \frac{\left(\binom{\psi}{\varphi}\right)}{\left(\binom{L}{K+1}\right)}|\psi+\varphi\rangle \\
& =\sum_{\varphi \leq K+1} \psi \sum_{x \in \operatorname{supp}(\psi)} \frac{\left(\binom{\psi}{\varphi}\right)}{\left(\binom{L}{K+1}\right)}|\psi+\varphi\rangle .
\end{aligned}
$$

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.

Proof. We use the formulation of mzip from Proposition 3.1.7 in terms of couplings. We will uses the following equation, for multisets $\varphi \in \mathcal{N}[K+1](X)$, $\psi \in \mathcal{N}[K+1](Y)$.

$$
\begin{align*}
& \left\{\chi \in \operatorname{dcpl}^{-1}(\varphi-1|x\rangle, \psi-1|y\rangle) \mid x \in \operatorname{supp}(\varphi), y \in \operatorname{supp}(\psi)\right\} \\
& =\left\{\rho-1|x, y\rangle \mid \rho \in \operatorname{dcpl}^{-1}(\varphi, \psi),(x, y) \in \operatorname{supp}(\rho)\right\} \tag{*}
\end{align*}
$$

For the inclusion ( $\subseteq$ ) we observe that if $\chi \in \operatorname{dcpl}^{-1}(\varphi-1|x\rangle, \psi-1|y\rangle)$, then $\mathcal{M}\left(\pi_{1}\right)(\chi)=\varphi-1|x\rangle$ and $\mathcal{M}\left(\pi_{2}\right)(\chi)=\psi-1|y\rangle$. We take $\rho:=\chi+1|x, y\rangle$, so that $(x, y) \in \operatorname{supp}(\rho)$ obviously holds. Moreover,

$$
\begin{aligned}
\mathcal{M}\left(\pi_{1}\right)(\rho) & =\mathcal{M}\left(\pi_{1}\right)(\chi+1|x, y\rangle) \\
& =\mathcal{M}\left(\pi_{1}\right)(\chi)+\mathcal{M}\left(\pi_{1}\right)(1|x, y\rangle) \quad \text { by Lemma 1.6.3 (3) } \\
& =\varphi-1|x\rangle+1|x\rangle \\
& =\varphi
\end{aligned}
$$

Similarly, $\mathcal{M}\left(\pi_{2}\right)(\rho)=\psi$, so that $\rho \in \operatorname{dcpl}^{-1}(\varphi, \psi)$. For the inclusion (Э), let $\rho \in$ $d c p l^{-1}(\varphi, \psi)$ have $(x, y) \in \operatorname{supp}(\rho)$. Then, clearly, $x \in \operatorname{supp}(\varphi)$ and $y \in \operatorname{supp}(\psi)$. 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 $\varphi \in \mathcal{N}[K+1](X)$, $\psi \in \mathcal{N}[K+1](Y)$.

$$
\begin{aligned}
& \text { mzip } \gg=(D D(\varphi) \otimes D D(\psi)) \\
&=\sum_{x \in \operatorname{supp}(\varphi)} \sum_{y \in \operatorname{supp}(\psi)} \frac{\varphi(x) \cdot \psi(y)}{(K+1)^{2}} \cdot \operatorname{mzip}(\varphi-1|x\rangle, \psi-1|y\rangle) \\
& \stackrel{3.6}{=} \sum_{x \in \operatorname{supp}(\varphi)} \sum_{y \in \operatorname{supp}(\psi)} \sum_{\chi \in d c p l^{-1}(\varphi-1|x\rangle, \psi-1|y\rangle)} \frac{\varphi(x) \cdot \psi(y) \cdot(\chi)}{(K+1)^{2} \cdot(\varphi-1|x\rangle) \cdot(\psi-1|y\rangle)}|\chi\rangle \\
&\left.\left.\stackrel{(\stackrel{*}{2})}{=} \sum_{\rho \in d c p l^{-1}(\varphi, \psi)} \sum_{(x, y) \in \operatorname{supp}(\rho)} \frac{(\rho-1|x, y\rangle)}{(\varphi) \cdot(\psi)}|\rho-1| x, y\right\rangle\right\rangle \\
&\left.\left.=\sum_{\rho \in d c p l^{-1}(\varphi, \psi)} \sum_{(x, y) \in \operatorname{supp}(\rho)} \frac{\rho(x, y)}{K+1} \cdot \frac{\mathbf{( \rho )}}{(\varphi) \cdot(\psi)}|\rho-1| x, y\right\rangle\right\rangle \\
&=\sum_{\rho \in \mathcal{M}[K+1](X \times Y)} \sum_{x \in \mathcal{M}[K](X \times Y)} D D(\rho)(\chi) \cdot \operatorname{mzip}(\varphi, \psi)(\rho)|\chi\rangle \\
&=D D \gg=\operatorname{mzip}(\varphi, \psi) .
\end{aligned}
$$

The draw-delete and draw-add channels both form an infinite chain. We are especially interested in the draw-delete chain, of the form:

A collection of distributions $\left(\sigma_{K} \in \mathcal{D}(\mathcal{N}[K](X))\right)_{K \in \mathbb{N}}$ will be called a drawdelete cone, or simply a $D D$-cone, if $D D \gg=\sigma_{K+1}=\sigma_{K}$ for each $K$. Such cones play an important role in De Finetti's work on exchangeability, see [41] and e.g. [91, 169]. 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' as 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 arbitrary set $X$ and for all $K \in \mathbb{N}$, the following two triangles commute.



Proof. For the triangle on the left, let $\omega \in \mathcal{D}(X)$ be an arbitrary distribution.

$$
\begin{aligned}
& (D D \odot m n[K+1])(\omega) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[K+1](X)} m n[K+1](\omega)(\psi) \cdot D D[K](\psi)(\varphi)|\varphi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in X} m n[K+1](\omega)(\varphi+1|x\rangle) \cdot \frac{\varphi(x)+1}{K+1}|\varphi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in X} \frac{(K+1)!}{\prod_{y}(\varphi+1|x\rangle)(y)!} \cdot \prod_{y} \omega(y)^{(\varphi+1|x\rangle)(y)} \cdot \frac{\varphi(x)+1}{K+1}|\varphi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in X} \frac{K!}{\prod_{y} \varphi(y)!} \cdot \omega(x) \cdot \prod_{y} \omega(y)^{\varphi(y)}|\varphi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K][(X)}\left(\sum_{x \in X} \omega(x)\right) \cdot(\varphi) \cdot \prod_{y} \omega(y)^{\varphi(y)}|\varphi\rangle \\
& =m n[K](\omega) .
\end{aligned}
$$

In the proof below, of commutation of the above triangle on the right, the marked equation $\stackrel{(*)}{=}$ indicates the use of Exercises 1.8.11 and 1.8.12 For urn $v \in \mathcal{N}[L](X)$ of size $L>0$, and for $\varphi \in \mathcal{N}[K](X)$,

$$
\begin{aligned}
& (D D \odot p l[K+1])(\psi) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{x \in \mathcal{N}[K+1](\operatorname{supp}(\psi))} D D(\chi)(\varphi) \cdot p l[K+1](\psi)(\chi)|\varphi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K] \operatorname{supp}(\psi))} \sum_{x \in \operatorname{supp}(\psi)} \frac{\varphi(x)+1}{K+1} \cdot p l[K+1](\psi)(\varphi+1|x\rangle)|\varphi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K] \operatorname{supp}(\psi))} \sum_{x \in \operatorname{supp}(\psi)} \frac{\varphi(x)+1}{K+1} \cdot \frac{\left(\binom{\psi}{\varphi+1|x\rangle}\right)}{\left(\binom{L}{K+1}\right)}|\varphi\rangle \\
& \stackrel{(*)}{=} \sum_{\varphi \in \mathcal{N}[K] \operatorname{supp}(\psi))} \sum_{x \in \operatorname{supp}(\psi)} \frac{\psi(x)+\varphi(x)}{L+K} \cdot \frac{\left(\begin{array}{l}
\psi \\
\varphi \\
\hline
\end{array}\right)}{\left(\binom{L}{K}\right)}|\varphi\rangle \\
& =p l[K](\psi) .
\end{aligned}
$$

In the literature (transposition-)exchangeable distributions play an important role, notably in relation to De Finetti's result, see [41, 99, 14]. 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:

$$
\frac{\text { transposition-exchangeable }\left(\tau_{K} \in \mathcal{D}\left(X^{K}\right)\right)_{K \in \mathbb{N}} \text { with } \mathcal{D}(\pi)\left(\tau_{K+1}\right)=\tau_{K}}{\text { draw-delete cones }\left(\sigma_{K} \in \mathcal{D}(\mathcal{N}[K](X))\right)_{K \in \mathbb{N}}}
$$

Here, $\pi: X^{K+1} \rightarrow X^{K}$ is the function that projects away the last element.
Proof. ( $\downarrow$ ) Let $\tau_{K} \in \mathcal{D}\left(X^{K}\right)$, for $K \in \mathbb{N}$, be a collection of transpositionexchangeable distributions such that $\tau_{K+1}$ marginalises to $\tau_{K}$, for each $K \in$ $\mathbb{N}$, that is, $\mathcal{D}(\pi)\left(\tau_{K+1}\right)=\tau_{K}$. We first show that:

$$
\begin{equation*}
P D \gg=\tau_{K+1}=\tau_{K} . \tag{*}
\end{equation*}
$$

Indeed, by using the assumptions about the $\tau^{\prime} s$ we get, for each sequence $\vec{x}=\left(x_{1}, \ldots, x_{K}\right) \in X^{K}$,

$$
\begin{aligned}
\left(P D \gg \tau_{K+1}\right)(\vec{x}) & \stackrel{\sqrt[3]{3} 7}{=} \sum_{y \in X} \sum_{1 \leq i \leq K+1} \frac{1}{K+1} \cdot \tau_{K+1}\left(x_{1}, \ldots, x_{i-1}, y, x_{i}, \ldots, x_{K}\right) \\
& =\sum_{y \in X} \sum_{1 \leq i \leq K+1} \frac{1}{K+1} \cdot \tau_{K+1}\left(x_{1}, \ldots, x_{K}, y\right) \\
& =\sum_{y \in X} \tau_{K+1}\left(x_{1}, \ldots, x_{K}, y\right) \\
& =\mathcal{D}(\pi)\left(\tau_{K+1}\right)\left(x_{1}, \ldots, x_{K}\right) \\
& =\tau_{K}(\vec{x})
\end{aligned}
$$

The transposition-exchangeability of $\tau_{K}$ means that transp $\gg \tau_{K}=\tau_{K}=$ id $\gg \tau_{K}$ by Lemma 2.9.4 (3). When we view $\tau_{K}$ as a channel $1 \rightarrow X^{K}$, we can apply the equaliser property of arrangement in Proposition 2.9.5 (2). It gives us a unique map $\mathbf{1} \leadsto \mathcal{N}[K](X)$, that is a distribution $\bar{\tau}_{K} \in \mathcal{D}(\mathcal{N}[K](X))$, with arr $\gg \bar{\tau}_{K}=\tau_{K}$. Then:

$$
\mathcal{D}(\operatorname{acc})\left(\tau_{K}\right)=\operatorname{acc} \gg=\left(\operatorname{arr} \gg=\bar{\tau}_{K}\right)=(\operatorname{acc} \odot a r r) \gg=\bar{\tau}_{K} \stackrel{2.28}{=} \bar{\tau}_{K} . \quad(* *)
$$

We can now show that these $\bar{\tau}_{K}$ form a draw-delete cone:

$$
\begin{aligned}
D D \gg=\bar{\tau}_{K+1} & \stackrel{(* *)}{=}(D D \odot a c c) \gg=\tau_{K+1} \\
& =(\operatorname{acc} \odot P D) \gg \tau_{K+1} \quad \text { by Lemma 3.2.2 } \\
& =\operatorname{acc} \gg\left(P D \gg \tau_{K+1}\right) \\
& \stackrel{(*)}{=} \operatorname{acc} \gg=\tau_{K}=\mathcal{D}(\operatorname{acc})\left(\tau_{K}\right) \stackrel{(* *)}{=} \bar{\tau}_{K} .
\end{aligned}
$$

$(\Uparrow)$ Now assume we have a draw-delete cone $\sigma_{K} \in \mathcal{D}(\mathcal{N}[K](X))$, for $K \in$ $\mathbb{N}$. We define $\bar{\sigma}_{K}:=$ arr $\gg=\sigma_{K} \in \mathcal{D}\left(X^{K}\right)$. This $\bar{\sigma}_{K}$ is exchangeable by Lemma 2.9.4 (3):

$$
\text { transp } \gg=\bar{\sigma}_{K}=(\operatorname{arr} \odot \operatorname{acc} \odot a r r) \gg \sigma_{K}=\operatorname{arr} \gg=\sigma_{K}=\bar{\sigma}_{K} .
$$

Moreover, Exercise 3.2.3 gives us:

$$
\mathcal{D}(\pi)\left(\bar{\sigma}_{K+1}\right)=(\pi \odot \operatorname{arr}) \gg \sigma_{K+1}=(\operatorname{arr} \odot D D) \gg \sigma_{K+1}=\operatorname{arr} \gg=\sigma_{K}=\bar{\sigma}_{K} .
$$

Thus, these distributions $\bar{\sigma}_{K}$ 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,

$$
\overline{\bar{\tau}}_{K}=\text { arr } \gg=\bar{\tau}_{K}=\tau_{K}, \quad \text { by definition of } \bar{\tau}_{K} .
$$

Next $\overline{\bar{\sigma}}_{K}$ is by definition the unique distribution with arr $\gg=\overline{\bar{\sigma}}_{K}=\bar{\sigma}_{K}$. But also, again by definition, arr $\gg \sigma_{K}=\bar{\sigma}_{K}$. Hence $\overline{\bar{\sigma}}_{K}=\sigma_{K}$.

## Exercises

3.2.1 Check that we could have introduced $D D$ via the coequaliser property of accumulation (from Proposition 2.9.3 (1) as:

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:

$$
\begin{aligned}
P A \gg \operatorname{arr}(\varphi)= & \frac{1}{18}|a, a, b, b\rangle+\frac{1}{18}|a, b, a, b\rangle+\frac{1}{18}|a, b, b, a\rangle \\
& +\frac{1}{6}|a, b, b, b\rangle+\frac{1}{18}|b, a, a, b\rangle+\frac{1}{18}|b, a, b, a\rangle \\
& +\frac{1}{6}|b, a, b, b\rangle+\frac{1}{18}|b, b, a, a\rangle \\
& +\frac{1}{6}|b, b, a, b\rangle+\frac{1}{6}|b, b, b, a\rangle \\
= & \operatorname{arr} \gg=D A(\varphi) .
\end{aligned}
$$

3.2.3 Write $\pi: X^{K+1} \rightarrow X^{K}$ for the (standard, non-probabilistic) projection function that throws away the last element.

1 Show that the following diagram of channels commutes.


2 Check that accumulation does not commute with draw-and-delete, as in:

3.2.4 Show that the draw-and-delete and draw-and-add maps $D D$ and $D A$ 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 2.4.3 where the set $X$ is finite.

1 Prove that these uniform distributions form a draw-delete cone, that is, satisfy:

$$
D D \gg u^{n i f_{\mathcal{N}[K+1](X)}}=\operatorname{unif}_{\mathcal{N}[K](X)} .
$$

One can calculate this by expanding the definitions, or by combining Proposition 3.2 .9 with Exercise 2.6.3
2 Check that the draw-add maps do not preserve uniform distributions. Show for instance that for $X=\{a, b\}$,

$$
\begin{gathered}
\left.\left.\left.\left.\left.\left.D A \gg \text { unif }_{\mathcal{N}[3](X)}=\frac{1}{4}|4| a\right\rangle\right\rangle+\frac{1}{6}|3| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{6}|2| a\right\rangle+2|b\rangle\right\rangle \\
\left.\left.\left.\left.+\frac{1}{6}|1| a\right\rangle+3|b\rangle\right\rangle+\frac{1}{4}|4| b\right\rangle\right\rangle .
\end{gathered}
$$

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:

$$
\begin{aligned}
& \left.\left.\left.\left.\left.\left.\operatorname{mn}[2](\omega)=\frac{1}{9}|2| a\right\rangle\right\rangle+\frac{4}{9}|1| a\right\rangle+1|b\rangle\right\rangle+\frac{4}{9}|2| b\right\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.\left.\left.\operatorname{mn}[3](\omega)=\frac{1}{27}|3| a\right\rangle\right\rangle+\frac{2}{9}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{4}{9}|1| a\right\rangle+2|b\rangle\right\rangle+\frac{8}{27}|3| b\right\rangle\right\rangle .
\end{aligned}
$$

Now show that $D A \gg m n[2](\omega)$ differs from the last distribution. (Pólya channels do commute with DA, see Exercise 3.4.1)
3.2.7 Draw-add $D A$ does not commute with mzip, like $D D$ 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:

$$
\begin{aligned}
& \text { DA } \gg=\operatorname{mzip}(\varphi, \psi) \\
& \begin{aligned}
&=\left.\left.\left.\left.\frac{1}{4}|2| a, 1\right\rangle+1|b, 0\rangle\right\rangle+\frac{1}{4}|1| a, 1\right\rangle+2|b, 0\rangle\right\rangle \\
&\left.\left.\left.\left.\quad+\frac{1}{4}|2| a, 0\right\rangle+1|b, 1\rangle\right\rangle+\frac{1}{4}|1| a, 0\right\rangle+2|b, 1\rangle\right\rangle \\
& \text { mzip } \gg(D A(\varphi) \otimes D A(\psi)) \\
&=\left.\left.\left.\left.\frac{1}{6}|1| a, 0\right\rangle+1|a, 1\rangle+1|b, 0\rangle\right\rangle+\frac{1}{12}|2| a, 1\right\rangle+1|b, 0\rangle\right\rangle \\
&\left.\left.\left.\left.\quad+\frac{1}{12}|1| a, 1\right\rangle+2|b, 0\rangle\right\rangle+\frac{1}{12}|2| a, 0\right\rangle+1|b, 1\rangle\right\rangle \\
& \quad\left.\left.\left.\left.+\frac{1}{6}|1| a, 0\right\rangle+1|a, 1\rangle+1|b, 1\rangle\right\rangle+\frac{1}{6}|1| a, 0\right\rangle+1|b, 0\rangle+1|b, 1\rangle\right\rangle \\
& \quad\left.\left.\left.\left.+\frac{1}{6}|1| a, 1\right\rangle+1|b, 0\rangle+1|b, 1\rangle\right\rangle+\frac{1}{12}|1| a, 0\right\rangle+2|b, 1\rangle\right\rangle .
\end{aligned}
\end{aligned}
$$

3.2.8 Let $X$ be a finite set, say with $n$ elements, of the form $X=\left\{x_{1}, \ldots, x_{n}\right\}$.

Define for $K \geq 1$,

$$
\left.\left.\sigma_{K}:=\sum_{1 \leq i \leq n} \frac{1}{n}|K| x_{i}\right\rangle\right\rangle \in \mathcal{D}(\mathcal{N}[K](X))
$$

Thus:

$$
\begin{aligned}
& \left.\left.\left.\left.\sigma_{1}=\frac{1}{n}|1| x_{1}\right\rangle\right\rangle+\cdots+\frac{1}{n}|1| x_{n}\right\rangle\right\rangle \\
& \left.\left.\left.\left.\sigma_{2}=\frac{1}{n}|2| x_{1}\right\rangle\right\rangle+\cdots+\frac{1}{n}|2| x_{n}\right\rangle\right\rangle, \quad \text { etc. }
\end{aligned}
$$

Show that these $\sigma_{K}$ form a cone, both for draw-delete and for drawadd:

$$
D D \gg=\sigma_{K+1}=\sigma_{K} \quad \text { and } \quad D A \gg=\sigma_{K}=\sigma_{K+1}
$$

Thus, the whole sequence $\left(\sigma_{K}\right)_{K \in \mathbb{N}_{>0}}$ can be generated from $\sigma_{1}=$ unif $f_{\mathcal{N}[1](X)}$ by repeated application of $D A$.
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:

$$
\left.\left.\frac{(D D \gg=\sigma)(\varphi)}{\mathbf{( \varphi )}}=\sum_{x \in X} \frac{\sigma(\varphi+1|x\rangle)}{(\varphi+1|x\rangle)}|\varphi+1| x\right\rangle\right\rangle .
$$

3.2.10 Show that the probabilistic projection channel PD makes the following diagram commute.

3.2.11 Recall the uniform projection channel unpr $[K]: X^{K} \mapsto X$ from Exercise 2.4.7 Prove that the following rectangles commute.

3.2.12 We define a draw-store-delete channel $D S D$ of the form:

$$
\mathcal{N}[K+1](X) \xrightarrow{D S D} \mathcal{D}(X \times \mathcal{N}[K](X))
$$

as:

$$
\left.\left.\operatorname{DSD}(\psi):=\sum_{x \in \operatorname{supp}(\psi)} \frac{\psi(x)}{\|\psi\|}|x, \psi-1| x\right\rangle\right\rangle .
$$

Thus, frequentist learning is the first marginal of $D S D$ and drawdelete is the second.

We also define a multiset-cons function mcons: $X \times \mathcal{N}[K](X) \rightarrow$ $\mathcal{N}[K+1](X)$ as $\operatorname{mcons}(x, \varphi)=\varphi+1|x\rangle$.

1 Check that mcons $\odot D S D=i d$, so that $D S D \odot$ mcons is split idempotent in Chan.

2 Prove the equality below, expressed in terms of string diagrams.


We shall see an extended version in Theorem 3.4.4
3.2.13 Check that the following diagram does not commute.


### 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 Example 2.1.2 (2). 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:

$$
\mathcal{D}(X) \xrightarrow{m n[K]} \mathcal{D}(\mathcal{N}[K](X)) \quad \text { written as } \quad \mathcal{D}(X) \xrightarrow{m n[K]} \mathcal{M}[K](X) .
$$

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 $m n[K](\omega)$ 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 $m n[K]: \mathcal{D}(X) \rightsquigarrow \mathcal{N}[K](X)$ with associated sequential and parallel composition $\odot$ 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 $\omega \in \mathcal{D}(X)$.

$$
\begin{aligned}
& \operatorname{mn}[K](\omega) \stackrel{[2.40}{=} \sum_{\varphi \in \mathcal{N}[K](X)}(\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)}|\varphi\rangle \stackrel{\sqrt{2.41}}{=} \mathcal{D}(\operatorname{acc})(\operatorname{iid}[K](\omega)) \\
&=\mathcal{D}(\operatorname{acc})\left(\omega^{K}\right) \\
&=\mathcal{D}(\operatorname{acc})(\underbrace{\omega \otimes \cdots \otimes \omega}_{K \text { times }})
\end{aligned}
$$

Thus, the probability assigned to a draw / multiset $\varphi \in \mathcal{N}[K](X)$ is the sum over all probabilities $\omega^{K}(\vec{x})=\prod_{i} \omega\left(x_{i}\right)$ for sequence $\vec{x} \in X^{K}$ that accumulate to $\varphi$, that is, with $\operatorname{acc}(\vec{x})=\varphi$. Recall, from Proposition 1.7.2, that there are ( $\varphi$ )


We start with a result from [88] (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 7.6 . One can recognise this result in [15], §2.2], where it is mentioned that a particular sum - amounting to accumulation - is sufficient as statistic for iid.

In the first point below we emphasise the crucial property of accumulation
that is used. This property occurs in more general form in Proposition 7.6 .5 later on. It is formulated in terms of kernel relations $\operatorname{ker}(f)$, for a function $f$, where $\operatorname{ker}(f)=\left\{\left(x, x^{\prime}\right) \mid f(x)=f\left(x^{\prime}\right)\right\}$.

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 $\operatorname{ker}(\operatorname{acc}) \subseteq \operatorname{ker}(\operatorname{iid}[K](\omega))$. More explicitly, for all sequences $\vec{x}, \vec{y} \in X^{K}$,

$$
\operatorname{acc}(\vec{x})=\operatorname{acc}(\vec{y}) \Longrightarrow \omega^{K}(\vec{x})=\omega^{K}(\vec{y}) .
$$

2 The accumulation function acc: $X^{K} \rightarrow \mathcal{N}[K](X)$ and its probabilistic 'arrangement' inverse arr $=\operatorname{acc}^{\sim 1}: \mathcal{N}[K](X) \leadsto X^{K}$ from 2.27) satisfy the equation between the following two string diagrams.


By discarding the wires on the left on both sides we get as consequence the equation:

$$
\begin{equation*}
\operatorname{iid}[K]=\operatorname{arr} \odot \operatorname{mn}[K] . \tag{3.15}
\end{equation*}
$$

Proof. We choose an arbitrary distribution $\omega \in \mathcal{D}(X)$.

1 Let $\vec{x}, \vec{y} \in X^{K}$ be given with $\operatorname{acc}(\vec{x})=\varphi=\operatorname{acc}(\vec{y})$. 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.

$$
\omega^{K}(\vec{x})=\prod_{i} \omega\left(x_{i}\right)=\prod_{z \in X} \omega(z)^{\varphi(z)}=\prod_{i} \omega\left(y_{i}\right)=\omega^{K}(\vec{y}) .
$$

2 We need to prove an equation 3.14 between tuple channels, see Defini-
tion 2.4.4 (3). We use Proposition 1.7.2 and Theorem 2.6.7.

$$
\begin{aligned}
\langle\mathrm{acc}, \text { id }\rangle \gg=\operatorname{iid}[K](\omega) & =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in X^{K}}\langle\operatorname{acc}\rangle(\vec{x})(\varphi) \cdot \operatorname{iid}[K](\omega)(\vec{x})|\varphi, \vec{x}\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \omega^{K}(\vec{x})|\varphi, \vec{x}\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in \operatorname{accc}^{-1}(\varphi)} \sum_{\vec{y} \in \operatorname{acc}^{-1}(\varphi)} \frac{\omega^{K}(\vec{y})}{\left|\operatorname{acc}^{-1}(\varphi)\right|}|\varphi, \vec{x}\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in \operatorname{aacc}^{-1}(\varphi)} \frac{\operatorname{Dacc})\left(\omega^{K}\right)(\varphi)}{(\varphi)}|\varphi, \vec{x}\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in X^{K}} \operatorname{arr}(\varphi)(\vec{x}) \cdot \operatorname{mn}[K](\omega)(\varphi)|\varphi, \vec{x}\rangle \\
& =\langle\operatorname{id}, \operatorname{arr}\rangle \gg \operatorname{mn}[K](\omega) .
\end{aligned}
$$

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 3.3.2. Fix a distribution $\omega \in \mathcal{D}(X)$ and a number $K$. For each $y \in X$,

$$
\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi(y)=K \cdot \omega(y) .
$$

Proof. The equation holds for $K=0$, since then $\varphi(y)=0$. Hence we may assume $K>0$. Then:

$$
\begin{aligned}
& \sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi(y) \\
= & \sum_{\varphi \in \mathcal{N}[K](X), \varphi(y) \neq 0} \varphi(y) \cdot \frac{K!}{\prod_{x} \varphi(x)!} \cdot \prod_{x} \omega(x)^{\varphi(x)} \\
= & \sum_{\varphi \in \mathcal{N}[K](X), \varphi(y) \neq 0} K \cdot \frac{(K-1)!}{\prod_{x}(\varphi-1|y\rangle)(x)!} \cdot \omega(y) \cdot \prod_{x} \omega(x)^{(\varphi-1 \mid y))(x)} \\
= & K \cdot \omega(y) \cdot \sum_{\varphi \in \mathcal{N}[K-1](X)}(\varphi) \cdot \prod_{x} \omega(x)^{\varphi(x)} \\
= & K \cdot \omega(y) \cdot \sum_{\varphi \in \mathcal{N}[K-1](X)} m n[K-1](\omega)(\varphi) \\
= & K \cdot \omega(y) .
\end{aligned}
$$

Theorem 3.3.3. Frequentist learning from a multinomial gives the original distribution:

$$
\begin{equation*}
\text { Flrn } \gg m n[K](\omega)=\omega . \tag{3.16}
\end{equation*}
$$

This means that the following diagram of channels commutes.


The channel sam: $\mathcal{D}(X) \mapsto X$ is the identity function $\mathcal{D}(X) \rightarrow \mathcal{D}(X)$, considered as channel; it is called the sample channel.

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 $\operatorname{Flrn}(\varphi)$ is close to $\omega$. Operationally, such sampling involves a random number generator and (mostly) produces different $\varphi$ 's in different runs. Hence such sampling is a randomised algorithm, not a mathematical function.
The multinomial distribution $m n[K](\omega)$ 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 $m n[K](\omega)=$ $\sum_{i} r_{i}\left|\varphi_{i}\right\rangle$. The above theorem says that frequentist learning of all these samples, while taking the probabilities $r_{i}$ into account, yields the original distribution $\omega$, that is:

$$
\text { Flrn >> } m n[K](\omega)=\sum_{i} r_{i} \cdot \operatorname{Flrn}\left(\varphi_{i}\right)=\omega
$$

We read this as a correctness result for multinomial sampling. We shall come accross similar results, see especially Theorem 5.5.4

Proof. By Lemma 3.3.2.

$$
\begin{aligned}
(F l r n \odot \operatorname{mn}[K])(\omega)(y) & =\sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \operatorname{Flrn}(\varphi)(y) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \frac{\varphi(y)}{K} \\
& =K \cdot \omega(y) \cdot \frac{1}{K} \\
& =\omega(y) .
\end{aligned}
$$

An alternative proof is suggested in Exercise 3.3.5
We turn to the multizip channel mzip, from Section 3.1, in relation to multinomial distributions. Our main result involves the decouple function $d c p l=$ $\left\langle\mathcal{N}\left(\pi_{1}\right), \mathcal{N}\left(\pi_{2}\right)\right\rangle: \mathcal{N}[K](X \times Y) \rightarrow \mathcal{N}[K](X) \times \mathcal{N}[K](Y)$, 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 3.3.4. Decouple and multizip are related to multinomial distributions via the following equation between string diagrams.


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 (3.18).

$$
\begin{aligned}
& \langle d c p l, i d\rangle \gg=m n[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\operatorname{dcpl}\rangle(\chi)(\varphi, \psi) \cdot \operatorname{mn}[K](\omega \otimes \rho)(\chi)|\varphi, \psi, \chi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[K](Y)} \sum_{\chi \in d c p l^{-1}(\varphi, \psi)}(\chi) \cdot \prod_{x \in X, y \in Y}((\omega \otimes \rho)(x, y))^{\chi(x, y)}|\varphi, \psi, \chi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[K](Y)} \sum_{\chi \in d c p l^{-1}(\varphi, \psi)} \\
& \text { ( } \chi \text { ) } \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)}|\varphi, \psi, \chi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[K](Y)} \sum_{\chi \in d c p l^{-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)}|\varphi, \psi, \chi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[K](Y)} \sum_{\chi \in d c p l^{-1}(\varphi, \psi)} \\
& (\chi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \cdot \prod_{y \in Y} \rho(y)^{\psi(y)}|\varphi, \psi, \chi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[K](Y)} \sum_{\chi \in d c p l^{-1}(\varphi, \psi)} \\
& \frac{(\chi)}{(\varphi) \cdot(\psi)} \cdot \operatorname{mn}[K](\omega)(\varphi) \cdot \operatorname{mn}[K](\rho)(\psi)|\varphi, \psi, \chi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[K](Y)} \sum_{\chi \in \mathcal{N}[K](X \times Y)} \\
& \operatorname{mzip}(\varphi, \psi)(\chi) \cdot(\operatorname{mn}[K](\omega) \otimes \operatorname{mn}[K](\rho))(\varphi, \psi)|\varphi, \psi, \chi\rangle \\
& =\langle i d, \operatorname{mzip}\rangle \geqslant=(m n[K](\omega) \otimes m n[K](\rho)) \text {. }
\end{aligned}
$$

This completes the proof.
Equation in 3.18) 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 (3.18), we get the result below. It demonstrates that the multinomial channel interacts nicely with multizip and tensors (of distributions). Later in Theorey 3.7.12 a categorical interpretation will be given: multinomial is a monoidal transformation.

Corollary 3.3.5. Multinomial channels commute with tensor and multizip:

$$
\begin{equation*}
m z i p \gg=(m n[K](\omega) \otimes m n[K](\rho))=m n[K](\omega \otimes \rho) . \tag{3.19}
\end{equation*}
$$

Diagrammatically this amounts to:


Proof. As mentioned, Equation (3.19) follows directly from (3.18, 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.


The three subdiagrams commute by (3.15), 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 4.1.3. We can describe it via a flatten map flat: $\mathcal{M}(\mathcal{M}(X)) \rightarrow \mathcal{M}(X)$, using inclusions $\mathcal{D}(X) \hookrightarrow \mathcal{M}(X)$ and $\mathcal{N}[K](X) \hookrightarrow \mathcal{M}(X)$. This is another consequence of Lemma 3.3.2.

Proposition 3.3.6. 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, $\omega$ can be understood as an element of $\mathcal{M}(X)$. With these inclusions in mind one has:

$$
\operatorname{flat}(m n[K](\omega))=\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi=K \cdot \omega \in \mathcal{M}(X) .
$$

The first equation expands the definition of 'flat'. The second equation is the relevant new fact.

Proof. Since:

$$
\begin{aligned}
\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi & =\sum_{x \in X}\left(\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi(x)\right)|x\rangle \\
& =\sum_{x \in X} K \cdot \omega(x)|x\rangle \quad \text { by Lemma 3.3.2 } \\
& =K \cdot \sum_{x \in X} \omega(x)|x\rangle \\
& =K \cdot \omega .
\end{aligned}
$$

Since a multinomial $m n[K](\omega)$ 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 $m n[K](\omega)$ 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 3.3.7.

1 Multinomial channels compose, with a bit of help of the (fixed-size) flatten operation for multisets, as in:


2 The following diagram commutes:


Proof. 1 We use the following diagram chase:


The triangle on the left commutes by Exercise 2.3.10 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: $\mathcal{L} \Rightarrow \mathcal{N}$ is a map of monads, see Exercise 1.11.7.
2 For $\omega \in \mathcal{D}(\mathcal{N}[K](X))$ and $x \in X$ we have:

$$
\begin{aligned}
& (\text { flat } \circ \mathcal{D}(\text { Flrn }) \circ \mathcal{D}(\text { flat }) \circ \operatorname{mn}[L])(\omega)(x) \\
& =((\text { Flrn } \circ \text { flat }) \gg m n[L](\omega))(x) \\
& =\sum_{\Psi \in \mathcal{N}[L](\mathcal{N}[K](X))} \operatorname{Flrn}(\text { flat }(\Psi))(x) \cdot \operatorname{mn}[L](\omega)(\Psi) \\
& =\sum_{\Psi \in \mathcal{N}[L](\mathcal{N}[K](X))} \sum_{\varphi \in \mathcal{N}[K](X)} \frac{\Psi(\varphi) \cdot \varphi(x)}{L \cdot K} \cdot \operatorname{mn}[L](\omega)(\Psi) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \frac{\varphi(x)}{L \cdot K} \cdot \sum_{\Psi \in \mathcal{N}[L](\mathcal{N}[K](X))} \Psi(\varphi) \cdot \operatorname{mn}[L](\omega)(\Psi) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \frac{\varphi(x)}{L \cdot K} \cdot L \cdot \omega(\varphi) \quad \text { by Lemma 3.3.2 } \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} F \operatorname{Flrn}(\varphi)(x) \cdot \omega(\varphi) \\
& =(F l r n \gg=\omega)(x) \\
& =(\text { flat } \circ \mathcal{D}(F l r n))(\omega)(x) .
\end{aligned}
$$

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 (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:

$$
\begin{array}{cc}
\left(\mathbb{R}_{\geq 0}\right)^{m} \xrightarrow{\text { Freq }} \mathcal{\cong} \mathcal{M}(\boldsymbol{m}) & \begin{array}{l}
\mathbb{N}^{m} \xrightarrow{\text { Freq }} \mathcal{\cong} \\
\\
\vec{r} \longmapsto \\
\sum_{0 \leq i<m} r_{i}|i\rangle
\end{array}  \tag{3.20}\\
\vec{n} \longmapsto
\end{array}
$$

We have seen these isomorphisms in Exercise 1.6.5
Theorem 3.3.8. Poisson and multinomial channels are related as expressed by the following equation between channels of type $\left(\mathbb{R}_{>0}\right)^{m} \leadsto \mathbb{N} \times \mathcal{N}(\boldsymbol{m})$.


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 2.2.3.

By marginalising out the wires on the right in (3.21) one obtains that Poisson distributions commute with sums, as we have seen in Proposition 2.7.6

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 3.21).

$$
\begin{aligned}
& \sum_{\vec{n} \in \mathbb{N} \mathbb{N}^{m}} \prod_{0 \leq i<m} \operatorname{pois}\left[\lambda_{i}\right]\left(n_{i}\right)|\operatorname{sum}(\vec{n}), \operatorname{Freq}(\vec{n})\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](m)} \prod_{0 \leq i<m} e^{-\lambda_{i}} \cdot \frac{\lambda_{i}^{\varphi(i)}}{\varphi(i)!}|K, \varphi\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](m)} e^{-\lambda} \cdot \frac{\prod_{0 \leq i<m} \lambda_{i}^{\varphi(i)}}{\varphi \rrbracket}|K, \varphi\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](m)} e^{-\lambda} \cdot \frac{\lambda^{K}}{K!} \cdot \frac{K!}{\varphi!} \cdot \prod_{0 \leq i<m}\left(\frac{\lambda_{i}}{\lambda}\right)^{\varphi(i)}|K, \varphi\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](m)} \operatorname{pois}[\lambda](K) \cdot \operatorname{mn}[K]\left(\sum_{i} \frac{\lambda_{i}}{\lambda}|i\rangle\right)(\varphi)|K, \varphi\rangle
\end{aligned}
$$

The latter expression is the right-hand-side of 3.21 .

Since $m n[K](\omega)$ is a distribution, the sum over all draws $\sum_{\varphi} m n[K](\omega)(\varphi)$ equals one. But what if we restrict this sum to draws $\varphi$ of certain colours only, that is, with $\operatorname{supp}(\varphi) \subseteq S$, for a proper subset $S \subseteq \operatorname{supp}(\omega)$ ? 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 3.5.6.

Proposition 3.3.9. Let $\omega \in \mathcal{D}(X)$ be given with a proper, non-empty subset $S \subseteq \operatorname{supp}(\omega)$, so $S \neq \emptyset$ and $S \neq \operatorname{supp}(\omega)$. For $K \in \mathbb{N}$, write:

$$
M_{K}:=\sum_{\varphi \in \mathcal{M}[K](S)} m n[K](\omega)(\varphi) .
$$

Then $M_{K}>M_{K+1}$ and $\lim _{K \rightarrow \infty} M_{K}=0$.
Proof. Write $r:=\sum_{x \in S} \omega(x)$ in:

$$
M_{K}=\sum_{\varphi \in \mathcal{M}[K](S)} m n[K](\omega)(\varphi)=\sum_{\varphi \in \mathcal{M}[K](S)}(\varphi) \cdot \prod_{x \in S} \omega(x)^{\varphi(x)} \stackrel{1.40]}{=} r^{K}
$$

Since $0<r<1$ we get $M_{K}=r^{K}>r^{K+1}=M_{K+1}$ and $\lim _{K \rightarrow \infty} M_{K}=\lim _{K \rightarrow \infty} r^{K}=0$.

We like to conclude with a result with a learning flavour. Suppose we have a single draw $\varphi \in \mathcal{N}[K](X)$ and we ask ourselves the question: which distribution $\omega \in \mathcal{D}(X)$ makes this draw most likely, that is, for which $\omega$ is the probability $m n[K](\omega)(\varphi)$ maximal?

Possibly not entirely unsurprising, the answer is $\omega=\operatorname{Flrn}(\varphi)$, the frequentist learning of $\varphi$. The proof requires some basic analysis, in particular Lagrange's multiplier method, see e.g. [15, §2.2]. The result below is standard, see e.g. [115, Ex. 17.5], but sometimes formulated differently.

Proposition 3.3.10. Let a natural multiset $\varphi \in \mathcal{N}[K](X)$ be given, with size $K>0$. The distribution $\omega \in \mathcal{D}(X)$ that gives the highest multinomial probability $m n[K](\omega)(\varphi)$ is Flrn $(\varphi)$. That is:

$$
\operatorname{Flrn}(\varphi) \in \underset{\omega \in \mathcal{D}(X)}{\operatorname{argmax}} \operatorname{mn}[K](\omega)(\varphi) .
$$

Proof. We seek the maximum of the function $\omega \mapsto \operatorname{mn}[K](\omega)$ by taking the derivative with respect to $\omega$. It turns out to be convenient to look at the maximum of the natural logarithm $\ln$ of the multinomial. This gives the same outcome, since the logarithm is monotone. It reduces all products in the definition of the multinomial distribution to sums, see Exercise 1.4.2.

Let's write $\varphi=\sum_{i} k_{i}\left|x_{i}\right\rangle$, with support $\operatorname{supp}(\varphi)=\left\{x_{1}, \ldots, x_{n}\right\}$. We look at distributions $\omega \in \mathcal{D}\left(\left\{x_{1}, \ldots, x_{n}\right\}\right)$, which we identify with numbers $v_{1}, \ldots, v_{n} \in$ $\mathbb{R}_{\geq 0}$ with $\sum_{i} v_{i}=1$. Lagrange's method takes care of this condition.

We thus seek the maximum of the log-validity function:

$$
\begin{aligned}
h(\vec{v}):=\ln \left(m n[K]\left(\sum_{i} v_{i}\left|x_{i}\right\rangle\right)(\varphi)\right) & =\ln ((\varphi))+\ln \left(\prod_{i} v_{i}^{k_{i}}\right) \\
& =\ln ((\varphi))+\sum_{i} k_{i} \cdot \ln \left(v_{i}\right) .
\end{aligned}
$$

The Lagrange multiplier method for finding the maximum prescribes that we take an additional parameter $\lambda$ in a new function:

$$
\begin{aligned}
H(\vec{v}, \lambda) & :=h(\vec{v})-\lambda \cdot\left(\left(\sum_{i} v_{i}\right)-1\right) \\
& =\ln ((\varphi))+\sum_{i} k_{i} \cdot \ln \left(v_{i}\right)-\lambda \cdot\left(\left(\sum_{i} v_{i}\right)-1\right) .
\end{aligned}
$$

The partial derivatives of $H$ are:

$$
\frac{\partial H}{\partial v_{i}}(\vec{v}, \lambda)=\frac{k_{i}}{v_{i}}-\lambda \quad \frac{\partial H}{\partial \lambda}(\vec{v}, \lambda)=1-\sum_{i} v_{i} .
$$

Setting all of these to 0 and solving gives the required maximum:

$$
1=\sum_{i} v_{i}=\sum_{i} \frac{k_{i}}{\lambda}=\frac{\sum_{i} k_{i}}{\lambda}=\frac{\|\varphi\|}{\lambda} .
$$

Hence $\lambda=\|\varphi\|$ and thus:

$$
v_{i}=\frac{k_{i}}{\lambda}=\frac{k_{i}}{\|\varphi\|} \stackrel{[2.5}{=} \operatorname{Flrn}(\varphi)\left(x_{i}\right) .
$$

There is another optimisation result that we introduce at this stage, which has a different flavor. Suppose we have a distribution $\sigma \in \mathcal{D}(\mathcal{N}[K](X))$ and we like to find the distribution $\omega \in \mathcal{D}(X)$ that is diverges minimally from $\sigma$. If $\sigma$ is already a multinomial distribution, say $\sigma=m n[K](\rho)$, then we can find $\rho$ as Flrn $\gg=\sigma$, via Theorem 3.3.3. The result below, from [94], shows that the least divergent distribution is obtained also in this way. Such kind of approximations are the topic of the area of 'variational inference', see e.g. [125, 137].

Proposition 3.3.11. Let $\sigma \in \mathcal{D}(\mathcal{N}[K](X))$ be given. The distribution $\omega \in \mathcal{D}(X)$ with minimal Kullback-Leibler divergence $D_{K L}(\sigma, m n[K](\omega))$ is Flrn $\gg=\sigma \in$ $\mathcal{D}(X)$, that is:

$$
\text { Flrn >>= } \sigma \in \underset{\omega \in \mathcal{D}(X)}{\operatorname{argmin}} D_{K L}(\sigma, m n[K](\omega)) .
$$

Proof. We first note that Flrn $\gg \sigma \in \mathcal{D}(X)$ is given by:

$$
\begin{equation*}
(F l r n \gg=\sigma)(x)=\sum_{\varphi \in \mathcal{M}[K](X)} \operatorname{Flrn}(\varphi)(x) \cdot \sigma(x)=\sum_{\varphi \in \mathcal{M}[K](X)} \frac{\varphi(x) \cdot \sigma(x)}{K} . \tag{*}
\end{equation*}
$$

Then, for an arbitrary $\omega \in \mathcal{D}(X)$, we unravel the divergence in the following manner, where Const is an irrelevant constant that depends only on $\sigma$, not on $\omega$.

$$
\begin{aligned}
& D_{K L}(\sigma, \operatorname{mn}[K](\omega)) \\
& \stackrel{\sqrt[2.47]{-}}{\stackrel{2.40)}{=}} \sum_{\varphi \in \mathcal{M}[K](X)} \sigma(\varphi) \cdot \ln \left(\frac{\sigma(\varphi)}{\operatorname{mn}[K](\omega)(\varphi)}\right) \\
& =\text { Const }-\sum_{x \in X}\left(\sum_{\varphi(\varphi)} \sigma(\varphi) \cdot \ln (\sigma(\varphi))-\sigma(\varphi) \cdot \ln ((\varphi))-\sigma(\varphi) \cdot \sum_{x \in X} \varphi(x) \cdot \ln (\omega(x))\right. \\
& \stackrel{(*)}{=} \operatorname{Const}-K \cdot \sum_{x \in X}(\text { Flrn 》= } \sigma)(x) \cdot \ln (\omega(x)) \\
& =\text { Const }-K \cdot \ln \left(\prod_{x \in X} \omega(x)^{(F \operatorname{lrn} \gg \sigma)(x)}\right) .
\end{aligned}
$$

Thus, in order to minimise the original divergence $D_{K L}(\sigma, m n[K](\omega))$ we have to maximise the latter log-expression $\ln (\cdots)$. This is precisely what happens in the above (proof of) Proposition 3.3.10 The log expression is maximal for $\omega=$ Flrn $\gg=\sigma$.

## 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 Use Theorem 3.3.3 and Proposition 2.8.4 (2) to prove that for distributions $\omega_{1}, \omega_{2} \in \mathcal{D}(X)$ and $K \in \mathbb{N}$,

$$
D_{K L}\left(\omega_{1}, \omega_{2}\right) \leq D_{K L}\left(m n[K]\left(\omega_{1}\right), \operatorname{mn}[K]\left(\omega_{2}\right)\right) .
$$

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:

$$
m n[K+L](\omega)(\varphi+\psi)=\frac{\binom{K+L}{K}}{\binom{\varphi+\psi}{\varphi}} \cdot \operatorname{mn}[K](\omega)(\varphi) \cdot \operatorname{mn}[L](\omega)(\psi),
$$

and if $\varphi, \psi$ have disjoint support, then:

$$
m n[K+L](\omega)(\varphi+\psi)=\binom{K+L}{K} \cdot m n[K](\omega)(\varphi) \cdot m n[L](\omega)(\psi)
$$

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:

- $\quad \sum_{x \in X} \sigma_{1}(1|x\rangle)=1$;
- $\sigma_{K}(K|x\rangle)=\sigma_{1}(1|x\rangle)^{K}$, 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 $m n[K](\omega)$, 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 $\sigma_{K}=m n[K](\omega)$ for a distribution $\omega \in \mathcal{D}(X)$.
3.3.5 Give an alternative proof of the equation Flrn $\gg m n[K](\omega)=\omega$ in Theorem 3.3.3 via Theorem 2.6.7 and Exercise2.4.7
3.3.6 Let $X$ be a set with a subset $A \subseteq X$. Consider the sum-evaluation function $\operatorname{sum}_{A}$, defined as $\operatorname{sum}_{A}(f)=\sum_{x \in A} f(x)$. We use it as map with the following two types.

$$
\mathcal{D}(X) \xrightarrow{\text { sum }_{A}}[0,1] \quad \mathcal{N}[K](X) \xrightarrow{\text { sum }_{A}}\{0,1, \ldots, K\}
$$

Show that the following diagram commutes.

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:

$$
m n[K](\omega)(\varphi)=\sum_{x \in \operatorname{supp}(\varphi)} \omega(x) \cdot \operatorname{mn}[K-1](\omega)(\varphi-1|x\rangle) .
$$

Here are two possible avenues:
1 use the recurrence relations 1.38 for multiset coefficients ( $\varphi$ );
2 show first:

$$
\omega(x) \cdot m n[K-1](\omega)(\varphi-1|x\rangle)=\operatorname{Flrn}(\varphi)(x) \cdot m n[K](\omega)(\varphi)
$$

Follow-up both avenues.
3.3.8 Prove the following items, in the style of Lemma 3.3.2, for a distribution $\omega \in \mathcal{D}(X)$ and a number $K>1$.

1 For two elements $y \neq z$ in $X$,

$$
\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi(y) \cdot \varphi(z)=K \cdot(K-1) \cdot \omega(y) \cdot \omega(z)
$$

2 For a single element $y \in X$,

$$
\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi(y) \cdot(\varphi(y)-1)=K \cdot(K-1) \cdot \omega(y)^{2} .
$$

3 Again, for a single element $y \in X$,

$$
\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi(y)^{2}=K \cdot(K-1) \cdot \omega(y)^{2}+K \cdot \omega(y) .
$$

Hint: Write $a^{2}=a \cdot(a-1)+a$ and then use the previous item and Lemma3.3.2
3.3.9 1 Generalise Proposition 2.7.4 (2) from binomials to multinomials: multinomials are closed under convolution. For $K, L \in \mathbb{N}$,


One can use Exercise 1.7.12 and Lemma 2.7.5 for a proof. One can also wait for Theorem 3.4.4, since this closure under convolution is a consequence.
2 Generalise these convolutions to $K$-ary form and conclude that $K$ sized draws can be reduced to parallel single draws, as in:


Notice that this is essentially Theorem 2.6.7, via the isomorphism $\mathcal{N}[1](X) \cong X$.
3.3.10 Show that for a natural multiset $\varphi$ of size $K$ one has:

$$
\operatorname{flat}(\operatorname{mn}[K](\operatorname{Flrn}(\varphi)))=\varphi
$$

3.3.11 Check that the following diagram does not commute, in general:


Take for instance as distribution of distributions $\left.\left.\frac{3}{4}\left|\frac{1}{3}\right| a\right\rangle+\frac{2}{3}|b\rangle\right\rangle+$ $\left.\left.\frac{1}{3}|1| a\right\rangle\right\rangle$, with $K=2$.
3.3.12 Check that multinomial channels do not commute with tensors, as in:


### 3.4 The hypergeometric channel

A multinomial distribution assigns probabilities to draws, with replacement. Two variations can be distinguished without 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:

$$
\begin{equation*}
\mathcal{N}[L](X) \xrightarrow{h g[K]} \mathcal{N}[K](X) \tag{3.22}
\end{equation*}
$$

We recall the definition with draws of size $K \leq L$ from an urn $v \in \mathcal{N}[L](X)$ of size $L$.

$$
\operatorname{hg}[K](v) \stackrel{\sqrt{2.34}}{-} \sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{L}{K}}|\varphi\rangle=\sum_{\varphi \leq K \psi} \frac{\prod_{x}\binom{v(x)}{\varphi(x)}}{\binom{L}{K}}|\varphi\rangle .
$$

Recall that we write $\varphi \leq_{K} v$ for: $\|\varphi\|=K$ and $\varphi \leq v$, 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 $\varphi$ should be multisets over the support $\operatorname{supp}(v)$ of the urn $v$. Indeed, only balls with colours that occur in the urn can be drawn. This is handled implicitly via the requirement $\varphi \leq_{K} v$. It ensures an inclusion of supports $\operatorname{supp}(\varphi) \subseteq \operatorname{supp}(v)$.

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
$D D$ from Definition 3.2.1 (2). As we shall see soon afterwards, this fact has many consequences.

Theorem 3.4.1. For $L, K \in \mathbb{N}$, the hypergeometric channel hg[K]: $\mathcal{N}[K+$ $L](X) \longrightarrow \mathcal{N}[K](X)$ equals L consecutive draw-and-delete's $D D^{L}$ in:


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 3.4.3 below.

The diagram 3.23 involves iterations $D D^{L}=D D \odot \cdots \odot D D$ as channels, given explicitly by:

$$
D D^{0}=\text { unit } \quad D D^{n+1}=D D \odot D D^{n}=D D^{n} \odot D D
$$

Proof. Write $v \in \mathcal{N}[K+L](X)$ for the urn. The proof proceeds by induction on the number of iterations $L$, starting with $L=0$. Then $\varphi \leq_{K} v$ means $\varphi=v$. Hence:

$$
h g[K](v)=\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{K+0}{K}}|\varphi\rangle=\frac{\binom{v}{v}}{\binom{K}{K}}|v\rangle=1|v\rangle=\operatorname{unit}(v)=D^{0}(v) .
$$

For the induction step we use $v \in \mathcal{N}[K+(L+1)](X)$ in:

$$
\begin{aligned}
& D D^{L+1}(v)=\sum_{\varphi \leq K^{v}}\left(D D \gg D D^{L}(v)\right)(\varphi)|\varphi\rangle \\
& =\sum_{\varphi \leq K^{v}} \sum_{\psi \in \mathcal{N}[K+1](X)} D D^{L}(v)(\psi) \cdot D D(\psi)(\varphi)|\varphi\rangle \\
& =\sum_{y \in X} \sum_{\varphi \in \mathcal{N}[K](X)} D D^{L}(v)(\varphi+1|y\rangle) \cdot \frac{\varphi(y)+1}{K+1}|\varphi\rangle \\
& \stackrel{(\mathrm{IH})}{=} \sum_{y \in X} \sum_{\varphi \in \mathcal{N}[K](X), \varphi+1 \mid y) \leq v} \frac{\binom{v}{\varphi+1|y\rangle}}{\binom{K+L+1}{K+1}} \cdot \frac{\varphi(y)+1}{K+1}|\varphi\rangle \\
& =\sum_{\varphi \leq K^{v}} \sum_{y, \varphi(y)<v(y)} \frac{(v(y)-\varphi(y)) \cdot\binom{v}{\varphi}}{(L+1) \cdot\binom{K+L+1}{K}}|\varphi\rangle \quad \text { by Exercise } 1.8 .12 \\
& =\sum_{\varphi \leq K^{v}} \frac{((K+L+1)-K) \cdot\binom{v}{\varphi}}{(L+1) \cdot\binom{K+L+1}{K}}|\varphi\rangle \\
& =\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{K+L+1}{K}}|\varphi\rangle=\operatorname{hg}[K](v) .
\end{aligned}
$$

From this result we can deduce many additional facts about hypergeometric distributions.

## Corollary 3.4.2.

1 Hypergeometric channels (3.22) are natural in $X$.
2 Frequentist learning form hypergeometric draws is like learning from the urn: for $L \geq K$,


3 For $L \geq K$ one has:


4 Also, for $L \geq K+1$,


5 Hypergeometric channels compose, as in:


6 Hypergeometric and multinomial channels commute, as in:


7 Hypergeometric channels commute with multizip: for $L \geq K$,


Proof. 1 By naturality of draw-and-delete, see Exercise 3.2.4
2 By Theorem 3.2.6
3 By Theorem 3.4.1
4 Idem.
5 Similarly, since $D D^{L+M}=D D^{L} \odot D D^{M}$.
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 [93] 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: $\mathcal{M}[K+1](X) \rightsquigarrow X \times \mathcal{M}[K](X)$. We can extend it to the hypergeometric case, for which we introduce the ad hoc name hypergeometric-store, abbreviated as $h g s[K]: \mathcal{N}[K+L](X) \leadsto \mathcal{N}[K](X) \times \mathcal{N}[L](X)$. It keeps track both of the
draw and the remaining urn:

$$
\begin{equation*}
\operatorname{hgs}[K](v):=\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{K+L}{K}}|\varphi, v-\varphi\rangle . \tag{3.24}
\end{equation*}
$$

We consider the two marginals.
Proposition 3.4.3. The two marginals of the above hypergeometric-store channel 3.24) are iterated draw-delete's in:


Proof. The triangle on the left commutes by Theorem 3.4.1

$$
\begin{aligned}
& \pi_{1} \gg=\operatorname{hg}_{S}[K](v) \stackrel{\left[\frac{3.24}{=}\right.}{=} \mathcal{D}\left(\pi_{1}\right)\left(\sum_{\varphi_{\leq K^{v}}} \frac{\binom{v}{\varphi}}{\binom{K+L}{K}}|\varphi, v-\varphi\rangle\right) \\
&=\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{K+L}{K}}|\varphi\rangle=h g[K](v)=D D^{L}(v) .
\end{aligned}
$$

For the triangle on the right we use Theorem 3.2.7

$$
\begin{aligned}
\pi_{2} \gg h g S[K](v) & \stackrel{\sqrt{3.24}}{-} \mathcal{D}\left(\pi_{2}\right)\left(\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{K+L}{K}}|\varphi, v-\varphi\rangle\right) \\
& =\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{K+L}{K}}|v-\varphi\rangle \stackrel{3.11]}{=} D D^{K}(v) .
\end{aligned}
$$

Theorem 3.4.4. The sum of multisets function and the hypergeometric-store channel are related in the following way:


The string diagram equation that we have seen earlier in Exercise 3.2.12for draw-store-delete is a special case of this theorem, for $K=1$. A consequence
of (3.25), obtained by marginalising out the outputs on the right, is the closure of multinomials under convolution, that we saw earlier in Exercise 3.3.9

Proof. For a distribution $\omega \in \mathcal{D}(X)$ one has:

$$
\begin{aligned}
& \langle s u m, i d\rangle \gg=(m n[K](\omega) \otimes m n[L](\omega)) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[L](X)} m n[K](\omega)(\varphi) \cdot m n[L](\omega)(\psi)|\varphi+\psi, \varphi, \psi\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[L](X)} \mathbf{( \varphi )} \cdot \prod_{x} \omega(x)^{\varphi(x)} \cdot \mathbf{( \psi )} \cdot \prod_{x} \omega(x)^{\psi(x)}|\varphi+\psi, \varphi, \psi\rangle \\
& =\sum_{\psi \in \mathcal{N}[L](X)} \sum_{\varphi \in \mathcal{N}[K](X)}(\varphi) \cdot(\psi) \cdot \prod_{x} \omega(x)^{(\varphi+\psi)(x)}|\varphi+\psi, \varphi, \psi\rangle \\
& =\sum_{v \in \mathcal{N}[K+L](X)} \sum_{\varphi \leq K^{v}} \frac{(\varphi) \cdot(v-\varphi)}{\mathbf{( v )}} \cdot(v) \cdot \prod_{x} \omega(x)^{v(x)}|v, \varphi, v-\varphi\rangle \\
& =\sum_{v \in \mathcal{N}[K+L](X)} \sum_{\varphi \leq K^{v}} \frac{K!\cdot L!\cdot v \rrbracket}{\varphi \rrbracket \cdot(v-\varphi) \rrbracket \cdot(K+L)!} \cdot m n[K+L](\omega)(v)|v, \varphi, v-\varphi\rangle \\
& =\sum_{v \in \mathcal{N}[K+L](X)} \sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\binom{K+L}{K}} \cdot m n[K+L](\omega)(v)|v, \varphi, v-\varphi\rangle \\
& =\langle i d, h g s[K]\rangle \gg m n[K+L](\omega) \text {. }
\end{aligned}
$$

There is the following analogue of Lemma3.3.2

Lemma 3.4.5. Let $v \in \mathcal{N}(X)$ be a non-empty urn of size $L=\|v\|>0$.

1 For a fixed element $y \in X$,

$$
\begin{equation*}
\sum_{\varphi \leq K^{v}} \operatorname{hg}[K](v)(\varphi) \cdot \varphi(y)=K \cdot \operatorname{Flrn}(v)(y) \tag{3.26}
\end{equation*}
$$

2 For $L \geq K \geq 1$,

$$
\operatorname{flat}(h g[K](v))=\sum_{\varphi \leq K v} h g[K](v)(\varphi) \cdot \varphi=\frac{K}{L} \cdot v=K \cdot \operatorname{Flrn}(v)
$$

Proof. 1 We use Exercises 1.8.11 and 1.8.12 in the marked equation $\stackrel{(*)}{=}$ below.

Let $K \leq L$ in:

$$
\begin{aligned}
\sum_{\varphi \leq K^{v}} h g[K](v)(\varphi) \cdot \varphi(y) & =\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi} \cdot \varphi(y)}{\binom{L}{K}} \\
& \stackrel{(*)}{=} \sum_{\varphi \leq K^{v}} \frac{v(y) \cdot\binom{v-1|y\rangle}{\varphi-1|y\rangle}}{\frac{L}{K} \cdot\binom{L-1}{K-1}} \\
& =K \cdot \frac{v(y)}{L} \cdot \sum_{\varphi \leq K-1 v-1|y\rangle} \frac{\binom{v-1|y\rangle}{\varphi}}{\binom{L-1}{K-1}} \\
& =K \cdot \operatorname{Flrn}(v)(y) .
\end{aligned}
$$

2 We use the previous point in:

$$
\begin{aligned}
& \sum_{\varphi \leq K^{v}} \operatorname{hg}[K](v)(\varphi) \cdot \varphi=\sum_{x \in X}\left(\sum_{\varphi \leq K^{v}} \operatorname{hg}[K](v)(\varphi) \cdot \varphi(x)\right)|x\rangle \\
& \stackrel{\text { B.26 }}{=} \sum_{x \in X} K \cdot \operatorname{Flrn}(v)(x)|x\rangle \\
&=K \cdot \operatorname{Flrn}(v) .
\end{aligned}
$$

## Exercises

3.4.1 Draw-delete $D D$ commutes with hypergeometric and Pólya distributions, see Corollary 3.4.2 (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 $\varphi=3|a\rangle+1|b\rangle$,

$$
\left.\left.\left.\left.\left.\left.\left.\left.\begin{array}{rl}
\operatorname{hg}[2](\varphi) & \left.\left.\left.\left.=\frac{1}{2}|2| a\right\rangle\right\rangle+\frac{1}{2}|1| a\right\rangle+1|b\rangle\right\rangle \\
& h g[3](\varphi)
\end{array}=\frac{1}{4}|3| a\right\rangle\right\rangle+\frac{3}{4}|2| a\right\rangle+1|b\rangle\right\rangle, \frac{1}{4}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{4}|1| a\right\rangle+2|b\rangle\right\rangle,
$$

And:

$$
\begin{aligned}
& p l[2](\varphi) \\
& \left.\left.\left.\left.\left.\left.=\frac{3}{5}|2| a\right\rangle\right\rangle+\frac{3}{10}|1| a\right\rangle+1|b\rangle\right\rangle+\frac{1}{10}|2| b\right\rangle\right\rangle \\
& p l[3](\varphi) \\
& \left.\left.\left.\left.\left.\left.\left.\left.=\frac{1}{2}|3| a\right\rangle\right\rangle+\frac{3}{10}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{3}{20}|1| a\right\rangle+2|b\rangle\right\rangle+\frac{1}{20}|3| b\right\rangle\right\rangle \\
& D A \geqslant=p l[2](\varphi) \\
& \left.\left.\left.\left.\left.\left.\left.\left.=\frac{3}{5}|3| a\right\rangle\right\rangle+\frac{3}{20}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{3}{20}|1| a\right\rangle+2|b\rangle\right\rangle+\frac{1}{10}|3| b\right\rangle\right\rangle
\end{aligned}
$$

3.4.2 1 Give a direct proof of Theorem 3.4 .1 for $L=1$.

2 Elaborate also the case $K=1$ in Theorem 3.4.1 and show that in that case $D D^{L}(\psi)=\operatorname{Flrn}(\psi)$, at least when we identify the (isomorphic) sets $\mathcal{N}[1](X)$ and $X$.
3.4.3 Let unif $\mathcal{L}_{L} \in \mathcal{D}(\mathcal{N}[L](X))$ be the uniform distribution on multisets of size $L$, from Exercise 3.2.5, where $X$ is a finite set. Show that $h g[K] »=$ unif $_{L}=$ unif $_{K}$, for $K \leq L$.
3.4.4 Prove in analogy with Exercise 3.3.8, for an urn $v \in \mathcal{N}[L](X)$ and elements $y, z \in X$, the following points.
1 When $y \neq z$ and $K \leq L$,

$$
\begin{aligned}
& \sum_{\varphi \in \mathcal{N}[K](X)} h g[K](v)(\varphi) \cdot \varphi(y) \cdot \varphi(z) \\
= & K \cdot(K-1) \cdot \operatorname{Flrn}(v)(y) \cdot \frac{v(z)}{L-1} .
\end{aligned}
$$

2 When $K \leq L$,

$$
\begin{aligned}
& \sum_{\varphi \in \mathcal{N}[K](X)} \lg [K](v)(\varphi) \cdot \varphi(y)^{2} \\
= & K \cdot \operatorname{Flrn}(v)(y) \cdot \frac{(K-1) \cdot v(y)+(L-K)}{L-1} .
\end{aligned}
$$

3.4.5 Use Theorem 3.4.1 to prove the following two recurrence relations for hypgeometric distributions.

$$
\begin{aligned}
h g[K](v) & =\sum_{x \in \operatorname{supp}(v)} \operatorname{Flrn}(v)(x) \cdot h g[K-1](v-1|x\rangle) \\
h g[K](v)(\varphi) & =\sum_{x} \frac{\varphi(x)+1}{K+1} \cdot h g[K-1](v)(\varphi+1|x\rangle)
\end{aligned}
$$

3.4.6 Fix numbers $N, M \in \mathbb{N}$ and write $v=N|0\rangle+M|1\rangle$ for an urn with $N$ balls of colour 0 and $M$ of colour 1 . Let $n \leq N$. Show that:

$$
\sum_{0 \leq m \leq M} h g[n+m](v)(n|0\rangle+m|1\rangle)=\frac{N+M+1}{N+1}
$$

Hint: Use Exercise 1.8.9. Notice that the right-hand side does not depend on $n$.

### 3.5 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 [67, 119, 16].

The Pólya distribution for draws of size $K$ can be described as a channel of the form:

$$
\begin{equation*}
\mathcal{N}_{*}(X) \xrightarrow{p l[K]} \mathbb{N}[K](X) \tag{3.27}
\end{equation*}
$$

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 $\varphi$ are multisets over the support of the urn $v$.

$$
p l[K](v) \stackrel{2.36}{=} \sum_{\varphi \in \mathcal{N}[K](\operatorname{supp}(v))} \frac{\left(\binom{v}{\varphi}\right)}{\left(\binom{L}{K}\right)}|\varphi\rangle=\sum_{\varphi \in \mathcal{N}[K](\operatorname{supp}(v))} \frac{\prod_{x \in \operatorname{supp}(v)}\left(\binom{v(x)}{\varphi(x)}\right)}{\left(\binom{L}{K}\right)}|\varphi\rangle
$$

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. [126]. 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 3.4.5.
Lemma 3.5.1. Fix a non-empty urn $v \in \mathcal{N}(X)$.

1 For an arbitrary element $y \in X$,

$$
\begin{equation*}
\sum_{\varphi \in \mathcal{M}[K](\operatorname{supp}(v))} p l[K](v)(\varphi) \cdot \varphi(y)=K \cdot \operatorname{Flrn}(v)(y) \tag{3.28}
\end{equation*}
$$

2 For $K \geq 1$,

$$
\operatorname{flat}(p l[K](v))=\sum_{\varphi \in \mathcal{N}[K] \operatorname{supp}(v))} p l[K](v)(\varphi) \cdot \varphi=\frac{K}{L} \cdot v=K \cdot \operatorname{Flrn}(v)
$$

Proof. 1 We use Exercises 1.8.11 and 1.8.12 in the marked equation $\stackrel{(*)}{=}$ in:

$$
\begin{aligned}
\sum_{\varphi \in \mathcal{M}[K](\operatorname{supp}(v))} p l[K](v)(\varphi) \cdot \varphi(y) & =\sum_{\varphi \in \mathcal{M}[K] \operatorname{supp}(v))} \frac{\left(\binom{v}{\varphi}\right) \cdot \varphi(y)}{\left(\binom{L}{K}\right)} \\
& \stackrel{(*)}{=} \sum_{\varphi \in \mathcal{M}[K](\operatorname{supp}(v))} \frac{v(y) \cdot\binom{v+1|y\rangle}{\varphi-1|y\rangle}}{\frac{L}{K} \cdot\binom{L+1}{K-1}} \\
& =K \cdot \frac{v(y)}{L} \cdot \sum_{\varphi \in \mathcal{M}[K-1](\operatorname{supp}(v+1|y\rangle))} \frac{\binom{v+1|y\rangle}{\varphi}}{\binom{L+1}{K-1}} \\
& =K \cdot \operatorname{Flrn}(v)(y) .
\end{aligned}
$$

2 We proceed as follows.

$$
\begin{aligned}
& \sum_{\varphi \in \mathcal{N}[K](\operatorname{supp}(v))} p l[K](v)(\varphi) \cdot \varphi \\
&= \sum_{x \in X}\left(\sum_{\varphi \in \mathcal{N}[K] \operatorname{supp}(v))} p l[K](v)(\varphi) \cdot \varphi(x)\right)|x\rangle \\
& \stackrel{3.28}{=} \sum_{x \in X} K \cdot \operatorname{Flrn}(v)(x)|x\rangle=K \cdot \operatorname{Flrn}(v) .
\end{aligned}
$$

In the previous section we have seen that hypergeometric channels are in fact iterated draw-and-delete's, see Theorem 3.4.1. 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 3.4.2 (2) and (3).

Proposition 3.5.2. Let $K>0$.
1 Pólya functions pl $[K]: \mathcal{N}_{*}(X) \rightarrow \mathcal{D}(\mathcal{N}[K](X))$ are natural in $X$.
2 Frequentist learning and Pólya satisfy the following equation:


3 The hypergeometric channel preserves Pólya distributions:


4 Doing a draw-and-add before Pólya has no effect: for $L>0$,


Proof. 1 Easy.
2 For a multiset/urn $v \in \mathcal{N}(X)$ with $\|v\|=L>0$, and for $x \in X$,

$$
\begin{aligned}
&(F l r n \odot p l[K])(v)(x)=\sum_{\varphi \in \mathcal{N}[K](\operatorname{supp}(v))} \operatorname{Flrn}(\varphi)(x) \cdot p l[K](v)(\varphi) \\
&=\frac{1}{K} \cdot \sum_{\varphi \in \mathcal{N}[K] \operatorname{supp}(v))} \varphi(x) \cdot p l[K](v)(\varphi) \\
&=1 / 28) \\
& \text { Flrn }(\nu)(x)
\end{aligned}
$$

3 The hypergeometric channel $h g[K]: \mathcal{N}[K+L](X) \rightarrow \mathcal{N}[K](X)$ is an iteration of $L$ draw-delete's, see Theorem 3.4.1. These draw-delete's preserve Pólya, see Proposition 3.2.9.
4 We use Exercises 1.8.11 and 1.8.12 in the marked equation $\stackrel{(*)}{=}$ below. For urn $v \in \mathcal{N}[L](X)$ and draw $\varphi \in \mathcal{N}[K](\operatorname{supp}(v))$,

$$
\begin{aligned}
&(p l[K] \odot D A)(v)(\varphi)=\sum_{x \in \operatorname{supp}(v)} \frac{v(x)}{L} \cdot p l[K](v+1|x\rangle)(\varphi) \\
&=\sum_{x \in \operatorname{supp}(v)} \frac{v(x)}{L} \cdot \frac{\left(\binom{v+1|x\rangle}{\varphi}\right)}{\left(\binom{L+1}{K}\right)} \\
& \stackrel{(\stackrel{(*)}{=}}{=} \sum_{x \in \operatorname{supp}(v)} \frac{v(x)+\varphi(x)}{L+K} \cdot \frac{\left(\binom{v}{\varphi}\right)}{\left(\binom{L}{K}\right)} \\
&=\frac{\left(\binom{v}{\varphi}\right)}{\left(\binom{L}{K}\right)} \\
&=p I[K](v)(\varphi) .
\end{aligned}
$$

In (3.24) 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: $\mathcal{N}[L](X) \hookrightarrow \mathcal{N}[K](X) \times \mathcal{N}[K+L](X)$, defined as:

$$
\begin{equation*}
p l s[K](v):=\sum_{\varphi \leq K^{v}} \frac{\binom{v}{\varphi}}{\left(\binom{K+L}{K}\right)}|\varphi, v+\varphi\rangle . \tag{3.29}
\end{equation*}
$$

In the hypergeometric case both the marginals of the 'store' channel are iterations of draw-delete's, see Proposition 3.4.3. In the Pólya case only the second marginal is an iteration, of draw-add's.

Proposition 3.5.3. The two marginals of the Pólya-store channel 3.29) are iterated draw-delete's in:


Proof. Commutation of the triangle on the left holds trivially. Commutation on the right follows from Theorem 3.2.7

### 3.5.1 Large urns and large draws

In the remainder of this section we look several aspects of both hypergeometric and Pólya distributions in 'limiting' situations, when either urns or draws are very large. This results are interesting in themselves, but will also be useful later on in this book.
When urns and draws differ considerably in size, both hypergeometric and Pólya distributions look like multinomials. We formulate these facts in a slightly informal style. It is intuitively clear that when an urn from which we draw in hypergeometric mode is very large and the draw inolves only a small number of balls, the withdrawals do not really affect the urn. Hence in this case the hypergeometric distribution behaves like a multinomial distribution, where the urn (as distribution) is obtained via frequentist learning.

## Proposition 3.5.4.

1 (See e.g. [100] §3.7]) Small hypergeometric draws from large urns look like multinomials: when the urn $v$ is very large, in comparison to the draw $\varphi \leq_{K}$ $v$ we get:

$$
h g[K](v)(\varphi) \approx m n[K](F \operatorname{lrn}(v))(\varphi)
$$

2 Large Pólya draws from small urns look like multinomials: when the draw $\varphi$ is very large, in comparison to the urn $v$,

$$
\begin{array}{r}
\operatorname{pl}[K](v)(\varphi) \approx \frac{1}{K^{N-1}} \cdot \operatorname{mn}[L-N](\operatorname{Flrn}(\varphi))(v-\mathbf{1}), \\
\text { where } \mathbf{1}=\sum_{x \in \operatorname{supp}(v)} 1|x\rangle \text { and } L=\|v\| \text { with } N=|\operatorname{supp}(v)|
\end{array}
$$

Proof. We use in both cases, essentially like in Lemma 1.2 .2 , that when the number $n$ is much bigger than $m$, then:

$$
\frac{n!}{(n-m)!} \approx \frac{(n-m)!\cdot n^{m}}{(n-m)!}=n^{m}
$$

1 We assume that the urn $v$ is very large, in comparison to the draw $\varphi \leq_{K} v$.
Then:

$$
\begin{aligned}
\operatorname{hg}[K](v)(\varphi)=\frac{\binom{v}{\varphi}}{\binom{L}{K}} & =\frac{K!}{\prod_{x} \varphi(x)!} \cdot \frac{(L-K)!}{L!} \cdot \prod_{x} \frac{v(x)!}{(v(x)-\varphi(x))!} \\
& \approx \mathbf{( \varphi )} \cdot \frac{1}{L^{K}} \cdot \prod_{x} v(x)^{\varphi(x)} \\
& =\mathbf{( \varphi )} \cdot \prod_{x}\left(\frac{v(x)}{L}\right)^{\varphi(x)} \\
& =\mathbf{( \varphi )} \cdot \prod_{x} F \operatorname{Flrn}(v)(x)^{\varphi(x)}=\operatorname{mn}[K](\operatorname{Fln}(v))(\varphi) .
\end{aligned}
$$

2 Let the support $\operatorname{supp}(v)$ of the urn $v$ have $N$ elements. In the Pólya case we assume that the draws $\varphi$ satisfy $\operatorname{supp}(\varphi) \subseteq \operatorname{supp}(v)$. When the urn $v$ of size $L$ is very small with respect to the draw $\varphi$ of size $K$, we get:

$$
\begin{aligned}
p l[K](v)(\varphi)=\frac{\left(\binom{v}{\varphi}\right)}{\left(\binom{L}{K}\right)} & =\frac{K!}{(K+L-1)!} \cdot \frac{(L-1)!}{(v-\mathbf{1}) \square} \cdot \frac{\prod_{x}(v(x)+\varphi(x)-1)!}{\prod_{x} \varphi(x)!} \\
& \approx \frac{1}{K^{L-1}} \cdot(v-\mathbf{1}) \cdot \prod_{x} \varphi(x)^{v(x)-1} \\
& =\frac{K^{L-N}}{K^{L-1}} \cdot(v-\mathbf{1}) \cdot \prod_{x}\left(\frac{\varphi(x)}{K}\right)^{v(x)-1} \\
& =\frac{1}{K^{N-1}} \cdot \operatorname{mn}[L-N](\operatorname{Flrn}(\varphi))(v-\mathbf{1}) .
\end{aligned}
$$

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 3.3.9. 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 resul ${ }^{2}$

Lemma 3.5.5. Let $0<N<M$ be given. Define for $n \in \mathbb{N}$,

$$
a_{n}:=\prod_{i<n} \frac{N+i}{M+i}
$$

Then: $\lim _{n \rightarrow \infty} a_{n}=0$.
Proof. We switch to the (natural) logarithm $\ln$ and prove the equivalent statement $\lim _{n \rightarrow \infty} \ln \left(a_{n}\right)=-\infty$. We use that the logarithm turns products into sums,

[^4]see Exercise 1.4.2, and that the derivative of $\ln (x)$ is $\frac{1}{x}$. Then:
\[

$$
\begin{aligned}
\ln \left(a_{n}\right)=\sum_{i<n} \ln \left(\frac{N+i}{M+i}\right) & =\sum_{i<n} \ln (N+i)-\ln (M+i) \\
& =-\sum_{i<n} \int_{N+1}^{M+1} \frac{1}{x} \mathrm{~d} x \\
& \stackrel{(*)}{\leq}-\sum_{i<n} \frac{(M+i)-(N+i)}{M+i} \\
& =(N-M) \cdot \sum_{i<n} \frac{1}{M+i}
\end{aligned}
$$
\]

It is well known that the harmonic series $\sum_{n>0} \frac{1}{n}$ is infinite. Since $M>N$ the above sequence $\ln \left(a_{n}\right)$ 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 3.5.6. Consider an urn $v \in \mathcal{N}[L](X)$ with a proper non-empty subset $S \subseteq \operatorname{supp}(v)$.

1 Write for $K \leq L$,

$$
H_{K}:=\sum_{\varphi \in \mathcal{N}[K](S), \varphi \leq v} h g[K](v)(\varphi) .
$$

Then $H_{K}>H_{K+1}$; this stops at $K=L$, when the urn is empty.
2 Now write, for arbitrary $K \in \mathbb{N}$,

$$
P_{K}:=\sum_{\varphi \in \mathcal{N}[K](S)} p l[K](v)(\varphi) .
$$

Then $P_{K}>P_{K+1}$ and $\lim _{K \rightarrow \infty} P_{K}=0$.
Proof. We write $L_{S}:=\sum_{x \in S} v(x)$ 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,

$$
H_{K}=\sum_{\varphi \in \mathcal{N}[K](S), \varphi \leq v} \frac{\left(\prod_{x \in S}\binom{v(x)}{\varphi(x)}\right) \cdot\left(\prod_{x \notin S}\binom{v(x)}{0}\right)}{\binom{L}{K}}=\frac{\binom{L_{S}}{K}}{\binom{L}{K}}=\frac{L_{S}!}{L!} \cdot \frac{(L-K)!}{\left(L_{S}-K\right)!} .
$$

Using a similar description for $H_{K+1}$ we get:

$$
\begin{aligned}
H_{K}>H_{K+1} & \Longleftrightarrow \frac{L_{S}!}{L!} \cdot \frac{(L-K)!}{\left(L_{S}-K\right)!}>\frac{L_{S}!}{L!} \cdot \frac{(L-(K+1))!}{\left(L_{S}-(K+1)\right)!} \\
& \Longleftrightarrow \frac{L-K}{L_{S}-K}>1 .
\end{aligned}
$$

The latter holds because $L>L_{S}$, since $S$ is a proper subset of $\operatorname{supp}(v)$.
2 In the Pólya case we get, as before, but now via the Vandermonde formula for multichoose, see Proposition 1.8.6

$$
P_{K}=\frac{\left(\binom{L_{S}}{K}\right)}{\left(\binom{L}{K}\right)}=\frac{(L-1)!}{\left(L_{S}-1\right)!} \cdot \frac{\left(L_{S}+K-1\right)!}{(K+L-1)!} .
$$

We define:

$$
\begin{aligned}
a_{K}:=\frac{P_{K+1}}{P_{K}} & =\frac{(L-1)!}{\left(L_{S}-1\right)!} \cdot \frac{\left(L_{S}+(K+1)-1\right)!}{(L+(K+1)-1)!} \cdot \frac{\left(L_{S}-1\right)!}{(L-1)!} \cdot \frac{(K+L-1)!}{\left(L_{S}+K-1\right)!} \\
& =\frac{L_{S}+K}{K+L}<1, \quad \text { since } L_{S}<L .
\end{aligned}
$$

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 $\lim _{K \rightarrow \infty} P_{K}=0$. This follows from $\lim _{K \rightarrow \infty} \prod_{i<K} a_{i}=0$, which we obtain from Lemma 3.5.5

## Exercises

3.5.1 Recall Exercises 3.3.8 and 3.4.4 and show for an urn $v \in \mathcal{N}[L](X)$ and elements $y, z \in X$, the following two points.

1 When $y \neq z$,

$$
\begin{aligned}
& \sum_{\varphi \in \mathcal{N}[K](X)} p l[K](v)(\varphi) \cdot \varphi(y) \cdot \varphi(z) \\
= & K \cdot(K-1) \cdot \operatorname{Flrn}(v)(y) \cdot \frac{v(z)}{L+1} .
\end{aligned}
$$

2 Prove also:

$$
\begin{aligned}
& \sum_{\varphi \in \mathcal{N}[K](X)} p l[K](v)(\varphi) \cdot \varphi(y)^{2} \\
= & K \cdot F \operatorname{lrn}(v)(y) \cdot \frac{(K-1) \cdot v(y)+(L+K)}{L+1} .
\end{aligned}
$$

3.5.2 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 $x \in \operatorname{supp}(\omega)$, use the Multinomial Theorem (1.40) to show that:

$$
\begin{aligned}
& \sum_{\varphi \in \mathcal{N}[K-k][X-x)} \operatorname{mn}[K](\omega)(k|x\rangle+\varphi) \\
= & m n[K]\left(\omega(x)|x\rangle+(1-\omega(x))\left|x^{\perp}\right\rangle\right)\left(k|x\rangle+(K-k)\left|x^{\perp}\right\rangle\right) \\
= & b n[K](\omega(x))(k) .
\end{aligned}
$$

2 Prove, via Vandermonde's formula, that for be an urn $v$ of size $L \geq K$ one has:

$$
\begin{aligned}
& \sum_{\substack{\varphi \leq K-k v-v(x)|x\rangle}} h g[K](v)(k|x\rangle+\varphi) \\
& =h g[K]\left(v(x)|x\rangle+(L-v(x))\left|x^{\perp}\right\rangle\right)\left(k|x\rangle+(K-k)\left|x^{\perp}\right\rangle\right) .
\end{aligned}
$$

3 Show again for an urn $v$,

$$
\begin{aligned}
& \sum_{\varphi \in N[K-k] \mid \text { supp(v)-x)}} p l[K](v)(k|x\rangle+\varphi) \\
& =p l[K]\left(v(x)|x\rangle+(L-v(x))\left|x^{\perp}\right\rangle\right)\left(k|x\rangle+(K-k)\left|x^{\perp}\right\rangle\right) .
\end{aligned}
$$

### 3.6 The parallel multinomial law: four definitions

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 can be described via a tensor product $\otimes$ of multinomials, and will therefore be called the parallel multinomial law, abbreviated as pml.

This law pml is an instance of what is called a distributive law in category theory. It has popped up, without an explicit description, in [111, 35] and also in [38], for continuous probability, to describe 'point processes' as distributions over multisets (and lists), see also Section 3.9. The explicit descriptions of the law that we use below come from [80]. This law satisfies several elementary properties that combine basic elements of 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:

$$
\begin{equation*}
\mathcal{N}[K](\mathcal{D}(X)) \xrightarrow{p m l[K]} \mathcal{D}(\mathcal{N}[K](X)) . \tag{3.30}
\end{equation*}
$$

The dependence of $p m l$ on $K$ (and $X$ ) is often left implicit. Notice that $p m l$ 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. This section is solely devoted to defining this law, in four different manners - yielding each time the same result. The subsequent section collects basic properties of pml . The fact that we have multiple equivalent formulations of the same law allows us to switch freely and use whichever formulation is most convenient in a particular situation.

## First definition

Since the law $\sqrt{3.30}$ is rather complicated, we start with an example.

Example 3.6.1. Let $X=\{a, b\}$ be a sample space with two distributions $\omega, \rho \in$ $\mathcal{D}(X)$, given by:

$$
\begin{equation*}
\omega=\frac{1}{3}|a\rangle+\frac{2}{3}|b\rangle \quad \text { and } \quad \rho=\frac{3}{4}|a\rangle+\frac{1}{4}|b\rangle . \tag{3.31}
\end{equation*}
$$

We will define $p m l$ 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:

$$
3|a\rangle \quad 2|a\rangle+1|b\rangle \quad 1|a\rangle+2|b\rangle \quad 3|b\rangle .
$$

The goal is to assign a probability to each of them. The map pml does this in
the following way.

```
\(\operatorname{pml}(2|\omega\rangle+1|\rho\rangle)\)
    \(=\omega(a) \cdot \omega(a) \cdot \rho(a)|3| a\rangle\rangle\)
        \(+(\omega(a) \cdot \omega(a) \cdot \rho(b)+\omega(a) \cdot \omega(b) \cdot \rho(a)+\omega(b) \cdot \omega(a) \cdot \rho(a))|2| a\rangle+1|b\rangle\rangle\)
    \(+(\omega(a) \cdot \omega(b) \cdot \rho(b)+\omega(b) \cdot \omega(a) \cdot \rho(b)+\omega(b) \cdot \omega(b) \cdot \rho(a))|1| a\rangle+2|b\rangle\rangle\)
    \(+\omega(b) \cdot \omega(b) \cdot \rho(b)|3| b\rangle\rangle\)
    \(\left.\left.\left.\left.=\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{3}{4}|3| a\right\rangle\right\rangle+\left(\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{4}+\frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{4}+\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{3}{4}\right)|2| a\right\rangle+1|b\rangle\right\rangle\)
    \(\left.\left.\left.\left.+\left(\frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{4}+\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{4}+\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{3}{4}\right)|1| a\right\rangle+2|b\rangle\right\rangle+\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{4}|3| b\right\rangle\right\rangle\)
    \(\left.\left.\left.\left.\left.\left.\left.\left.=\frac{1}{12}|3| a\right\rangle\right\rangle+\frac{13}{36}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{4}{9}|1| a\right\rangle+2|b\rangle\right\rangle+\frac{1}{9}|3| b\right\rangle\right\rangle\).
```

There is a pattern. Let's try to formulate the law pml from (3.30) in general, for arbitrary $K$ and $X$. It is defined on a natural multiset $\sum_{i} n_{i}\left|\omega_{i}\right\rangle$ with multiplicities $n_{i} \in \mathbb{N}$ satisfying $\sum_{i} n_{i}=K$, and with distributions $\omega_{i} \in \mathcal{D}(X)$. The number $\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right)(\varphi)$ describes the probability of the $K$-sized multiset $\varphi$ over $X$, by using for each element occurring in $\varphi$ the probability of that element in the corresponding distribution in $\sum_{i} n_{i}\left|\omega_{i}\right\rangle$.

In order to make this description precise we assume that the indices $i$ are somehow ordered, say as $i_{1}, \ldots, i_{m}$ and use this ordering to form a product state:

$$
\otimes_{i} \omega_{i}^{n_{i}}=\underbrace{\omega_{i_{1}} \otimes \cdots \otimes \omega_{i_{1}}}_{n_{i_{1}} \text { times }} \otimes \cdots \otimes \underbrace{\omega_{i_{m}} \otimes \cdots \otimes \omega_{i_{m}}}_{n_{i_{m}} \text { times }} \in \mathcal{D}\left(X^{K}\right) .
$$

Now we formulate the first definition:

$$
\begin{align*}
\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right) & :=\sum_{\vec{x} \in X^{K}}\left(\bigotimes_{i} \omega_{i}^{n_{i}}\right)(\vec{x})|\operatorname{acc}(\vec{x})\rangle \\
& =\sum_{\varphi \in \mathcal{N}[K](X)}\left(\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)}\left(\otimes_{i} \omega_{i}^{n_{i}}\right)(\vec{x})\right)|\varphi\rangle . \tag{3.32}
\end{align*}
$$

## Second definition

There is an alternative formulation of the parallel multinomial law, using convolution sums of parallel multinomial distributions via $\otimes$. This formulation is the basis for the phrase 'parallel multinomial'.

$$
\begin{align*}
\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right) & :=\mathcal{D}(+)\left(\bigotimes_{i} m n\left[n_{i}\right]\left(\omega_{i}\right)\right) \\
& =\sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)}\left(\Pi_{i} m n\left[n_{i}\right]\left(\omega_{i}\right)\left(\varphi_{i}\right)\right)\left|\sum_{i} \varphi_{i}\right\rangle . \tag{3.33}
\end{align*}
$$

The sum + that we use here as type:

$$
\mathcal{N}\left[n_{i_{1}}\right](X) \times \cdots \times \mathcal{N}\left[n_{i_{m}}\right](X) \xrightarrow{ } \mathcal{N}[\underbrace{n_{i_{1}}+\cdots+n_{i_{m}}}_{K}](X) .
$$

Thus, this sum has type $\prod_{i} \mathcal{N}\left[n_{i}\right](X) \rightarrow \mathcal{N}\left[\sum_{i} n_{i}\right](X)$.
This definition (3.33) may be seen as a convolution sum of multinomials, in the style of Definition 2.7.1. This is justified via inclusions $\mathcal{N}\left[n_{i}\right](X) \hookrightarrow \mathcal{N}(X)$ into the commutative monoid of multisets on $X$. This perspective is exploited further in the fourth definition below.

Proposition 3.6.2. The definitions of the law pml in (3.32) and (3.33) are equivalent.

Proof. Because:

$$
\begin{aligned}
& \operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right) \\
& \text { 3.32 } \sum_{\vec{x} \in X^{K}}\left(\otimes \omega_{i}^{n_{i}}\right)(\vec{x})|\operatorname{acc}(\vec{x})\rangle \\
& =\sum_{i, \vec{x}_{i} \in X^{n_{i}}}\left(\prod_{i} \omega_{i}^{n_{i}}\left(\vec{x}_{i}\right)\right)\left|\sum_{i} \operatorname{acc}\left(\vec{x}_{i}\right)\right\rangle \quad \text { see Exercise 1.7.12 } \\
& =\sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)}\left(\prod_{i} \sum_{\vec{x}_{i} \in \operatorname{acc}\left(\varphi_{i}\right)} \omega_{i}^{n_{i}}\left(\vec{x}_{i}\right)\right)\left|\sum_{i} \varphi_{i}\right\rangle \\
& =\sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)}\left(\Pi_{i} \mathcal{D}(\operatorname{acc})\left(\omega_{i}^{n_{i}}\right)\left(\varphi_{i}\right)\right)\left|\sum_{i} \varphi_{i}\right\rangle \\
& =\sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)}\left(\prod_{i}\left(\operatorname{acc} \odot \operatorname{iid}\left[n_{i}\right]\right)\left(\omega_{i}\right)\left(\varphi_{i}\right)\right)\left|\sum_{i} \varphi_{i}\right\rangle \\
& \stackrel{2.41}{=} \sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)}\left(\Pi_{i} m n\left[n_{i}\right]\left(\omega_{i}\right)\left(\varphi_{i}\right)\right)\left|\sum_{i} \varphi_{i}\right\rangle .
\end{aligned}
$$

Example 3.6.3. We continue Example 3.6.1 but now we describe the application of the parallel multinomial law pml in terms of multinomials, as in (3.33). We use the same multiset $2|\omega\rangle+1|\rho\rangle$ of distributions $\omega, \rho$ from 3.31]. The calculation of pml on this multiset, according to the second definition (3.33), is a bit more complicated than in Example 3.6.1 according to the first definition, since we have to evaluate the multinomial expressions. But of course the
outcome is the same.

$$
\begin{aligned}
& \operatorname{pml}(2|\omega\rangle+1|\rho\rangle) \\
& =\sum_{\varphi \in \mathcal{N}[2](X), \psi \in \mathcal{N}[1](X)} \operatorname{mn[2](\omega )(\varphi )\cdot mn[1](\rho )(\psi )|\varphi +\psi \rangle } \\
& =m n[2](\omega)(2|a\rangle) \cdot m n[1](\rho)(1|a\rangle)|3| a\rangle\rangle \\
& +(m n[2](\omega)(2|a\rangle) \cdot m n[1](\rho)(1|b\rangle) \\
& +m n[2](\omega)(1|a\rangle+1|b\rangle) \cdot m n[1](\rho)(1|a\rangle))|2| a\rangle+1|b\rangle\rangle \\
& +(m n[2](\omega)(1|a\rangle+1|b\rangle) \cdot m n[1](\rho)(1|b\rangle) \\
& +m n[2](\omega)(2|b\rangle) \cdot m n[1](\rho)(1|a\rangle))|1| a\rangle+2|b\rangle\rangle \\
& +m n[2](\omega)(2|b\rangle) \cdot m n[1](\rho)(1|b\rangle)|3| b\rangle\rangle \\
& \left.\left.=\binom{2}{2,0} \cdot \omega(a)^{2} \cdot\binom{1}{1,0} \cdot \rho(a)|3| a\right\rangle\right\rangle \\
& \left.\left.+\left(\binom{2}{2,0} \cdot \omega(a)^{2} \cdot\binom{1}{1,0} \cdot \rho(b)+\binom{2}{1,1} \cdot \omega(a) \cdot \omega(b) \cdot\binom{1}{1,0} \cdot \rho(a)\right)|2| a\right\rangle+1|b\rangle\right\rangle \\
& \left.\left.+\left(\binom{2}{1,1} \cdot \omega(a) \cdot \omega(b) \cdot\binom{1}{1,0} \cdot \rho(b)+\binom{2}{2,0} \cdot \omega(b)^{2} \cdot\binom{1}{1,0} \cdot \rho(a)\right)|1| a\right\rangle+2|b\rangle\right\rangle \\
& \left.\left.+\binom{2}{2,0} \cdot \omega(b)^{2} \cdot\binom{1}{1,0} \cdot \rho(b)|3| b\right\rangle\right\rangle \\
& \left.\left.\left.\left.=\left(\frac{1}{3}\right)^{2} \cdot \frac{3}{4}|3| a\right\rangle\right\rangle+\left(\left(\frac{1}{3}\right)^{2} \cdot \frac{1}{4}+2 \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{4}\right)|2| a\right\rangle+1|b\rangle\right\rangle \\
& \left.\left.\left.\left.+\left(2 \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{4}+\left(\frac{2}{3}\right)^{2} \cdot \frac{3}{4}\right)|1| a\right\rangle+2|b\rangle\right\rangle+\left(\frac{2}{3}\right)^{2} \cdot \frac{1}{4}|3| b\right\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.\left.\left.=\frac{1}{12}|3| a\right\rangle\right\rangle+\frac{13}{36}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{4}{9}|1| a\right\rangle+2|b\rangle\right\rangle+\frac{1}{9}|3| b\right\rangle\right\rangle .
\end{aligned}
$$

Indeed, this is what has been calculated in Example 3.6.1.

## Third definition

Our third definition of pml is more abstract than the previous ones. It uses the coequaliser property of accumulation, see Proposition 2.9.3 1]. It determines $p m l$ as the unique (dashed) map in:


We have already seen this situation in Exercise 2.9.4, where we did not introduce a name yet for the map that we now call pml.

Proposition 3.6.4. The definitions of pml in (3.32), (3.33) and 3.34) are all
equivalent. The latest definition (3.34) yields as new formulation pml $=$ acc $\odot$ $\otimes \odot a r r$. To be precise:

$$
\begin{equation*}
\operatorname{pml}[K]=\left(\mathcal{N}[K](\mathcal{D}(X)) \xrightarrow{\operatorname{arr}[K]} \mathcal{D}(X)^{K} \xrightarrow{\otimes[K]} X^{K} \xrightarrow{\text { acc }[K]} \mathcal{N}[K](X)\right) \tag{3.35}
\end{equation*}
$$

Proof. By the uniqueness property in the triangle (3.34) it suffices to prove that $p m l$ as described in the first (3.32) or second (3.33) formulation makes this triangle commute. For this we use the first version. We rely again on the fact that accumulation is stable under transposition. Assume we have $\vec{\omega}=$ $\left(\omega_{1}, \ldots, \omega_{K}\right) \in \mathcal{D}(X)^{K}$ with $\operatorname{acc}(\vec{\omega})=\sum_{i \in S} n_{i}\left|\omega_{i}\right\rangle$, for $S \subseteq\{1, \ldots, K\}$. Then:

$$
\begin{aligned}
(p m l \circ \operatorname{acc})(\vec{\omega}) & =\operatorname{pml}\left(\sum_{i \in S} n_{i}\left|\omega_{i}\right\rangle\right) \\
& \stackrel{3.32]}{=} \sum_{\vec{x} \in X^{K}}\left(\otimes_{i} \omega_{i}^{n_{i}}\right)(\vec{x})|\operatorname{acc}(\vec{x})\rangle \\
& =\sum_{\vec{x} \in X^{K}}\left(\omega_{1} \otimes \cdots \otimes \omega_{K}\right)(\vec{x})|\operatorname{acc}(\vec{x})\rangle \\
& =\mathcal{D}(\operatorname{acc})(\otimes(\vec{\omega}))
\end{aligned}
$$

Implicitly, for well-definedness of the first definition 3.32 of pml we already used that the precise ordering of states in the tensor $\otimes(\vec{\omega})$ is irrelevant in the formulation of pml .

The formulation in Equation (3.35) follows from Theorem 2.9.6 11.
This third diagrammatic formulation (3.34) of the parallel multinomial law is not very useful for actual calculations, like in Examples 3.6.1 and 3.6.3 But it is useful for proving properties about pml , via the uniqueness part of the third definition. This will be illustrated in Exercise 3.6 .3 below.
In contrast, formulation $\sqrt{3.35}$ is very useful. It follows a common pattern, also used for instance in multizip (3.1): first turn a multiset into sequences, apply a function (like zip, or tensor), and then accumulate the resulting sequences back into multisets.

In [118] 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 $\mathcal{N}[K](\mathcal{M}(X)) \rightarrow \mathcal{M}(\mathcal{N}[K](X))$, which are shown to (1) form a distributive law $\mathcal{N M} \Rightarrow \mathcal{M N}$ of monads, and (2) restrict to $\mathcal{N D} \Rightarrow \mathcal{D N}$. This is a variation to obtain $p m l$. We postpone its role as distributive law of monads to Section 3.8.

## Fourth definition

For our fourth and last definition we have to piece together some earlier observations.

1 Recall from Proposition 2.7.2 that if $M$ is a commutative monoid, then so is the set $\mathcal{D}(M)$ of distributions on $M$, with convolution sum:

$$
\omega+\rho=\mathcal{D}(+)(\omega \otimes \rho)=\sum_{x_{1}, x_{2} \in M}(\omega \otimes \rho)\left(x_{1}, x_{2}\right)\left|x_{1}+x_{2}\right\rangle .
$$

2 Recall from Proposition 1.6 .6 that such commutative monoid structure corresponds to an $\mathcal{N}$-algebra sum: $\mathcal{N}(\mathcal{D}(M)) \rightarrow \mathcal{D}(M)$, given by:

$$
\begin{align*}
\operatorname{sum}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right) & =\sum_{i} n_{i} \cdot \omega_{i} \\
& =\sum_{\vec{x} \in M^{K}}\left(\bigotimes_{i} \omega_{i}^{n_{i}}\right)(\vec{x})\left|\sum \vec{x}\right\rangle \quad \text { where } \quad K=\sum_{i} n_{i} . \tag{3.36}
\end{align*}
$$

3 For an arbitrary set $X$, the set $\mathcal{N}(X)$ of natural multisets on $X$ is a commutative monoid, see Lemma 1.6.3. Applying the previous two items with $M=\mathcal{N}(X)$ yields an $\mathcal{N}$-algebra:

$$
\begin{equation*}
\mathcal{N}(\mathcal{D N}(X)) \xrightarrow{\text { sum }} \mathcal{D N}(X) \tag{3.37}
\end{equation*}
$$

It interacts with $\mathcal{N}$ 's unit and flatten operations as described in Proposition 1.6 .6

We can now formulate the fourth definition:

$$
\begin{equation*}
p m l:=\left(\mathcal{N D}(X) \xrightarrow{\mathcal{N} \mathcal{D}\left(u n i t^{\mathcal{N}}\right)} \mathcal{N D \mathcal { D }}(X) \xrightarrow{\text { sum }} \mathcal{D N}(X)\right) . \tag{3.38}
\end{equation*}
$$

Proposition 3.6.5. The definition of pml in (3.38) restricts to $\mathcal{N}[K](\mathcal{D}(X)) \rightarrow$ $\mathcal{D}(\mathcal{N}[K](X))$, for each $K \in \mathbb{N}$. This restriction is the same pml as defined in (3.32), (3.33) and (3.34).

Proof. We elaborate formulation 3.38, on a $K$-sized multiset $\sum_{i} n_{i}\left|\omega_{i}\right\rangle$.

$$
\begin{aligned}
\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right) & \left.\stackrel{\sqrt{3.38}}{-} \operatorname{sum}\left(\sum_{i} n_{i} \mid \mathcal{D}(\text { unit })\left(\omega_{i}\right)\right\rangle\right) \\
& =\sum_{\vec{\varphi} \in \mathcal{N}(X)^{K}}\left(\bigotimes_{i} \mathcal{D}(\text { unit })\left(\omega_{i}\right)^{n_{i}}\right)(\vec{\varphi})\left|\sum \vec{\varphi}\right\rangle \\
& \left.\left.=\sum_{x_{1}, \ldots, x_{K} \in X}\left(\bigotimes_{i} \omega_{i}^{n_{i}}\right)\left(x_{1}, \ldots, x_{K}\right)|1| x_{1}\right\rangle+\cdots+1\left|x_{K}\right\rangle\right\rangle \\
& =\bigotimes_{\vec{x} \in X^{K}}\left(\omega_{i}^{n_{i}}\right)(\vec{x})|\operatorname{acc}(\vec{x})\rangle
\end{aligned}
$$

The last line coincides with the first formulation (3.32) of pml.

## Exercises

3.6.1 Check that the multinomial channel can be obtained via the parallel multinomial law, in two different ways.

1 Use the first or second formulation, 3.32) or 3.33), of pml to compute that for a distribution $\omega \in \mathcal{D}(X)$ and number $K \in \mathbb{N}$ one has:

$$
m n[K](\omega)=\operatorname{pml}(K|\omega\rangle),
$$

that is:


2 Prove the same thing via the third formulation (3.34p of pml, and via Theorem 2.6.7.
3.6.2 Apply the multiset flatten map flat: $\mathcal{M}(\mathcal{M}(X)) \rightarrow \mathcal{M}(X)$ in the setting of Example 3.6.1 to show that:

$$
\operatorname{flat}(2|\omega\rangle+1|\rho\rangle)=\frac{17}{12}|a\rangle+\frac{19}{12}|b\rangle=\operatorname{flat}(\operatorname{pml}(2|\omega\rangle+1|\rho\rangle)) .
$$

(The general formulation appears later on in Proposition 3.7.3.)
3.6.3 We claim pml is natural: for each function $f: X \rightarrow Y$ the following diagram commutes.


1 Prove this claim. The easiest way is via the formulation of pml in (3.35).
2 Give an alternative proof using the uniqueness part of the third formulation (3.34), as suggested in the diagram:

3.6.4 Generalise Exercise 3.3.9 from multinomials to parallel multinomials, as in the following diagram.


### 3.7 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 $p m l$ with a uniqueness property, in its third formulation (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 acculation map. This uses the coequaliser property of accumulation, see Proposition 2.9.3 1].
The next result enriches what we already now.
Proposition 3.7.1. The parallel multinomial law pml is the unique channel making both rectangles below commute.


Recall from Lemma 2.9.4 that we write transp $=$ arr $\odot$ acc for the transposition idempotent that occurs twice as the horizontal composite in this diagram.

Proof. The rectangle on the left is the third formulation of pml in 3.34) and thus provides uniqueness. Commutation of the rectangle of the right follows from a uniqueness argument, using that the outer rectangle commutes by Exercise 2.9.5

$$
\begin{aligned}
\otimes \odot \operatorname{arr} \odot \mathrm{acc} & =\operatorname{arr} \odot \operatorname{acc} \odot \otimes & & \text { by Exercise 2.9.5 } \\
& =\text { arr } \odot p m l \odot \text { acc } & & \text { by } 3.34 .
\end{aligned}
$$

This result shows that $p m l$ is squeezed between $\otimes$, both on the left and on the right. We have seen in Exercise 2.3 .11 that $\otimes$ is a distributive law. We shall prove the same about pml below.

But first we show how pml interacts with frequentist learning.

Theorem 3.7.2. The distributive law pml commutes with frequentist learning, in the sense that for $\Psi \in \mathcal{N}[K](\mathcal{D}(X))$,

$$
F l r n \gg \operatorname{pml}(\Psi)=\operatorname{flat}(F \operatorname{lrn}(\Psi)) .
$$

Equivalently, in diagrammatic form:


The channel sam: $\mathcal{D}(X) \rightsquigarrow X$ at the bottom is the identity function $\mathcal{D}(X) \rightarrow$ $\mathcal{D}(X)$, used as sample operation.

Proof. We use the formulation $\mathrm{pml}=\mathrm{acc} \odot \bigotimes \odot$ arr from (3.35) in the following diagram, using the uniform projection channel unpr.


The subdiagrams commute by Exercise 2.4.7

We include a result that generalises Exercise 3.6.2. It is a discrete version of [38, Lem. 13]. The proof below uses the 'mean' of multinomials, from Proposition 3.3.6

Proposition 3.7.3. 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: $\mathcal{M}(\mathcal{M}(X)) \rightarrow \mathcal{M}(X)$. Via these inclusions, one has, for $\Psi \in \mathcal{N}[K](\mathcal{D}(X))$,

$$
\operatorname{flat}(p m l(\Psi))=\operatorname{flat}(\Psi) .
$$

Proof. For $\Psi=n_{1}\left|\omega_{1}\right\rangle+\cdots+n_{k}\left|\omega_{k}\right\rangle \in \mathcal{N}[K](\mathcal{D}(X))$,

$$
\begin{aligned}
& \text { flat } \left.(\mathrm{pml}(\Psi))=\text { flat } \sum_{\varphi_{1} \in \mathcal{N}\left[n_{1}\right](X), \ldots, \varphi_{k} \in \mathcal{N}\left[n_{k}\right](X)}\left(\prod_{i} \operatorname{mn}\left[n_{i}\right](\omega)\left(\varphi_{i}\right)\right)\left|\sum_{i} \varphi_{i}\right\rangle\right) \\
& =\sum_{\varphi_{1} \in \mathcal{N}\left[n_{1}\right](X), \ldots, \varphi_{k} \in \mathcal{N}\left[n_{k}\right](X)}\left(\prod_{i} \operatorname{mn}\left[n_{i}\right](\omega)\left(\varphi_{i}\right)\right) \cdot\left(\sum_{i} \varphi_{i}\right) \\
& =\sum_{\varphi_{1} \in \mathcal{N}\left[n_{1}\right](X), \ldots, \varphi_{k} \in \mathcal{N}\left[n_{k}\right](X)}\left(\prod_{i} \operatorname{mn}\left[n_{i}\right](\omega)\left(\varphi_{i}\right)\right) \cdot \varphi_{1} \\
& +\cdots+\sum_{\varphi_{1} \in \mathcal{N}\left[n_{1}\right](X), \ldots, \varphi_{k} \in \mathcal{N}\left[n_{k}\right](X)}\left(\prod_{i} \operatorname{mn}\left[n_{i}\right](\omega)\left(\varphi_{i}\right)\right) \cdot \varphi_{k} \\
& =\sum_{\varphi_{1} \in \mathcal{N}\left[n_{1}\right](X)}^{\operatorname{mn}\left[n_{1}\right]\left(\omega_{1}\right)\left(\varphi_{1}\right) \cdot \varphi_{1}+\cdots+\sum_{\varphi_{k} \in \mathcal{N}\left[n_{k}\right](X)} \operatorname{mn}\left[n_{k}\right]\left(\omega_{k}\right)\left(\varphi_{k}\right) \cdot \varphi_{k}} \\
& =n_{1} \cdot \omega_{1}+\cdots+n_{k} \cdot \omega_{k} \quad \text { by Proposition } 3.3 .6 \\
& =\operatorname{flat}(\Psi) .
\end{aligned}
$$

The parallel multinomial law pml contains multinomial distributions. But it also interacts with the multinomial channel, as described next.

## Theorem 3.7.4.

1 The parallel multinomial law pml commutes with multinomials in the following manner.


2 There is a second form of exchange between pml and multinomials:


Proof. 1 We use that $m n[K]=\operatorname{acc} \odot \operatorname{iid}[K]$, see Theorem 2.6.7, in:


The rectangle on the left commutes by Exercise 2.4.8, and the one on the right by Proposition 3.7.1.

2 We show that precomposing both legs in the diagram with the accumulation map acc: $\mathcal{D}(X)^{K} \rightarrow \mathcal{M}[K] \mathcal{D}(X)$ yields an equality. This suffices by Proposition 2.9.3 (1).

```
\(\mathcal{D}(\) flat \() \circ m n[L] \circ p m l \circ\) acc
    \(\stackrel{3.38}{-} \mathcal{D}(\) flat \() \circ m n[L] \circ \mathcal{D}(\) acc \() \circ \bigotimes\)
    \(=\mathcal{D}(f l a t) \circ \mathcal{D N}(a c c) \circ m n[L] \circ \bigotimes \quad\) by naturality of \(m n[L]\)
    \(=\mathcal{D}(f l a t) \circ \mathcal{D N}(a c c) \circ\left(m z i p_{K} \odot(m n[L] \otimes \cdots \otimes m n[L])\right)\)
            by a generalisation of Corollary 3.3 .5
    \(=\mathcal{D}(\) flat \() \circ \mathcal{D N}(\) acc \() \circ\) flat \(\circ \mathcal{D}\left(\operatorname{mzip}_{K}\right) \circ \bigotimes \circ m n[L]^{K}\)
    \(=\) flat \(\circ \mathcal{D}^{2}(\) flat \() \circ \mathcal{D}^{2} \mathcal{N}(\mathrm{acc}) \circ \mathcal{D}\left(\mathrm{mzip}_{K}\right) \circ \bigotimes \circ m n[L]^{K}\)
    \(=\) flat \(\circ \mathcal{D}(\) unit \() \circ \mathcal{D}(+) \circ \otimes \circ m n[L]^{K} \quad\) by Proposition 3.1.10
    \(=\mathcal{D}(+) \circ \otimes \circ m n[L]^{K}\)
    \(=\mathcal{D}(f l a t) \circ \mathcal{D}(a c c) \circ \otimes \circ m n[L]^{K} \quad\) by Exercise 1.8 .3
    \(\stackrel{3.38}{-} \mathcal{D}(\) flat \() \circ \mathrm{pml} \circ\) acc \(\circ \mathrm{mn}[L]^{K}\)
    \(=\mathcal{D}(f l a t) \circ p m l \circ \mathcal{N}(m n[L]) \circ\) acc by naturality of acc.
```

We turn to pml and hypergeometric channels, which, as we have seen in Theorem 3.4.1, are composites of draw-and-delete maps. We know from Proposition 3.2 .9 that multinomial channels commute with draw-and-delete. The same holds for the parallel multinomial law.

Proposition 3.7.5. The following diagram commutes.


Proof. We use the probabilistic projection channel PD: $X^{K+1} \leadsto X^{K}$ from Definition 3.2.1 and its interaction with $\otimes$ in Exercise 3.2.10.

$$
\begin{aligned}
D D \odot p m l & \stackrel{\sqrt[3.35]{=}}{=} D D \odot a c c \odot \bigotimes \odot \text { arr } & & \\
& =\text { acc } \odot P D \odot \bigotimes \odot \text { arr } & & \text { by Lemma 3.2.2 } \\
& =\text { acc } \odot \otimes \odot P D \odot \text { arr } & & \text { by Exercise 3.2.10 } \\
& =p m l \odot \text { acc } \odot P D \odot \text { arr } & & \text { by 3.34. } \\
& =p m l \odot D D \odot \text { acc } \odot \text { arr } & & \text { by Lemma 3.2.2 } \\
& =p m l \odot D D & & \text { by } 2.28 .
\end{aligned}
$$

Corollary 3.7.6. The parallel multinomial law commutes with the hypergeo-
metric channel: for $L \geq K$ one has:


Proof. Theorem 3.4.1 shows that the hypergeometric distribution can be expressed via iterated draw-and-deletes. Hence the result follows from (iterated application of) Proposition 3.7.5.

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 [177, 178] (with full credits). 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 multisetcomputations and probability can be combined. Indeed, in Corollary 3.7.8 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 3.6.3 that it is natural.

Proposition 3.7.7. 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.



Proof. In Exercise 2.3.11 we have seen that the big tensor $\otimes: \mathcal{D}(X)^{K} \rightarrow$ $\mathcal{D}\left(X^{K}\right)$ is a distributive law. These properties will be used to show that pml is
a distributive law too. We exploit the uniqueness property of the third formulation 3.34

$$
\begin{aligned}
& \text { pml } \circ \mathcal{N}(\text { unit }) \circ \text { acc } \\
& =\text { pml } \circ \text { acc } \circ u n i t^{K} \\
& =\mathcal{D}(\text { acc }) \circ \bigotimes \circ u n i t^{K} \\
& =\mathcal{D}(\text { acc }) \circ \text { unit } \\
& =\text { unit } \circ \text { acc }
\end{aligned}
$$

$$
=p m l \circ \operatorname{acc} \circ u n i t^{K} \quad \text { by naturality of acc, see Exercise } 1.7 .12
$$

by 3.34
via the first diagram in Exercise 2.3 .11
by naturality of unit.

Simarly for the flatten-diagram:

This result says that $p m l$ 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 $=\operatorname{Chan}(\mathcal{D})$ of channels for the monad, see [75, 90] for details.

Corollary 3.7.8. The $K$-fold product $(-)^{K}$ : Sets $\rightarrow$ Sets and the $K$-sized multiset functor $\mathcal{N}[K]:$ Sets $\rightarrow$ Sets extend to functors $(-)^{K}:$ Chan $\rightarrow$ Chan and $\mathcal{N}[K]$ : Chan $\rightarrow$ Chan, in commuting diagrams:


Both versions of $(-)^{K}$ and of $\mathcal{N}[K]$ coincide on objects. On a channel / morphism $c: X \rightsquigarrow Y$ the extended functor $(-)^{K}:$ Chan $\rightsquigarrow$ Chan is defined as:

$$
\begin{equation*}
c^{K}:=\left(X^{K} \xrightarrow{c \times \cdots \times c} \mathcal{D}(Y) \times \cdots \times \mathcal{D}(Y) \xrightarrow{\otimes} \mathcal{D}\left(Y^{K}\right)\right) . \tag{3.40}
\end{equation*}
$$

The extension $\mathcal{N}[K]:$ Chan $\rightarrow$ Chan on a channel $c: X \mapsto Y$ be described in

$$
\begin{aligned}
& \text { flat } \circ \mathcal{D}(p m l) \circ p m l \circ \text { acc } \\
& =\text { flat } \circ \mathcal{D}(p m l) \circ \mathcal{D}(\text { acc }) \circ 囚 \quad \text { by 3.34) } \\
& =\text { flat } \circ \mathcal{D}(\mathcal{D}(\text { acc })) \circ \mathcal{D}(\otimes) \circ \bigotimes \quad \text { again by (3.34) } \\
& =\mathcal{D}(\mathrm{acc}) \circ \text { flat } \circ \mathcal{D}(\otimes) \circ \otimes \quad \text { by naturality of flat } \\
& =\mathcal{D}(\text { acc }) \circ \bigotimes \circ \text { flat }^{K} \quad \text { via Exercise 2.3.11 } \\
& =p m l \circ \mathrm{acc} \circ \mathrm{flat}^{K} \\
& =p m l \circ \mathcal{N}(f l a t) \circ a c c \\
& \text { once again by } 3.34 \\
& \text { by naturality of acc. }
\end{aligned}
$$

two (equivalent) ways as:


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 $\mathcal{N}:$ Sets $\rightarrow$ Sets, as in the above diagram (3.41). 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 2.26) and Exercise 2.4.8(3). Commutation of the above rectangle on the left in $\sqrt{3.39}$ is trivial on sets/objects. Now let $g$ be a function, that is, a morphism in Sets. The upgoing functors in 3.39) sends $g$ to the deterministic channel $\langle g\rangle=$ unit $\circ g$. We have:

$$
\begin{aligned}
& \langle g\rangle^{K} \stackrel{3.40}{=} \otimes \circ((\text { unit } \circ g) \times \cdots \times(\text { unit } \circ g)) \\
& =\otimes \circ(\text { unit } \times \cdots \times \text { unit }) \circ(g \times \cdots \times g) \\
& =(\text { unit } \otimes \cdots \otimes \text { unit }) \circ(g \times \cdots \times g) \\
& \stackrel{2.26}{=} \text { unit } \circ(g \times \cdots \times g) \\
& =\left\langle g^{K}\right\rangle .
\end{aligned}
$$

The functor $\mathcal{N}[K]$ : Chan $\rightarrow$ Chan is defined on objects/sets as $X \mapsto$ $\mathcal{N}[K](X)$. A morphism $c: X \leadsto Y$ in Chan is sent to the channel $\mathcal{N}[K](X) \mapsto$ $\mathcal{N}[K](Y)$, obtained in 3.41. We check that these two formulations coincide, via the formulation of $p \mathrm{ml}$ in Equation (3.35) and naturality of arrangement:

$$
\begin{aligned}
p m l \circ \mathcal{N}(c) & =(\operatorname{acc} \odot \otimes \odot \operatorname{arr}) \circ \mathcal{N}(c) \\
& =\mathcal{D}(\operatorname{acc}) \circ \text { flat } \circ \mathcal{D}(\otimes) \circ \operatorname{arr} \circ \mathcal{N}(c) \\
& =\mathcal{D}(\operatorname{acc}) \circ \text { flat } \circ \mathcal{D}(\otimes) \circ \mathcal{D}(c \times \cdots \times c) \circ \operatorname{arr} \\
& =\mathcal{D}(\operatorname{acc}) \circ \text { flat } \circ \mathcal{D}(c \otimes \cdots \otimes c) \circ \operatorname{arr} \\
& =\operatorname{acc} \odot c^{K} \odot \text { arr. }
\end{aligned}
$$

This extension $\mathcal{N}[K]$ preseves identities and composition by Proposition 3.7.7.
The rectangle on the right in 3.39) commutes too, via Proposition 3.7.7

$$
\begin{aligned}
\mathcal{N}[K](\langle g\rangle)=p m l \circ \mathcal{N}(\text { unit } \circ g) & =\text { pml } \circ \mathcal{N}(u n i t) \circ \mathcal{N}(g) \\
& =\text { unit } \circ \mathcal{N}(g)=\langle\mathcal{N}(g)\rangle .
\end{aligned}
$$

Next we show that parallel multinomials and multizip commute. This is a non-trivial technical result, which plays a crucial role in the subsequent result.

Lemma 3.7.9. The following diagram commutes.


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.


The upper rectangle commutes by Proposition 3.7.1 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 third definition of pml in (3.34).

Theorem 3.7.10. The extended functor $\mathcal{N}[K]:$ Chan $\rightarrow$ Chan commutes with multizip: for channels $f: X \leadsto U$ and $g: Y \leadsto V$ one has:

$$
\begin{equation*}
\operatorname{mzip} \odot(\mathcal{N}[K](f) \otimes \mathcal{N}[K](g))=\mathcal{N}[K](f \otimes g) \odot \text { mzip } \tag{3.42}
\end{equation*}
$$

Diagrammatically this amounts to:


In combination with the unit and associativity of Proposition 3.1.3) 22 and (5) this means that the extended functor $\mathcal{N}[K]$ : Chan $\rightarrow$ Chan is a monoidal functor, via mzip.

Proof. This result is rather subtle, since $f, g$ are used as channels. So when we write $\mathcal{N}[K](f)$ we mean application of the extended functor $\mathcal{N}[K]$ : Chan $\rightarrow$ Chan, as in 3.41 , producing another channel. We shall write the multiset functor $\mathcal{N}$ : Sets $\rightarrow$ Sets without parameter $K$.

The left-hand side of the equation (3.42) thus expands as in the first equation below.

$$
\begin{array}{ll}
\text { mzip } \odot(\mathcal{N}[K](f) \otimes \mathcal{N}[K](g)) & \\
=\text { mzip } \odot(p m l \otimes p m l) \circ(\mathcal{N}(f) \times \mathcal{N}(g)) & \\
=p m l \odot \mathcal{N}(\otimes) \odot \text { mzip } \circ(\mathcal{N}(f) \times \mathcal{N}(g)) & \text { by Lemma3.7.9 } \\
=p m l \odot \mathcal{N}(\otimes) \odot \mathcal{N}(f \times g) \odot \text { mzip } & \text { by Proposition3.1.3] } 1] \\
=p m l \odot \mathcal{N}(f \otimes g) \odot \text { mzip } & \\
=\mathcal{N}[K](f \otimes g) \odot \text { mzip. } &
\end{array}
$$

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 3.7.9 does not hold, see Exercise 3.7 .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 3.7.11. Arrangement and accumulation, and draw-delete are natural transformation in the situations:


Proposition 3.1 .3 and Lemma 3.2 .8 say that these arr and DD are monoidal, as natural transformations.

Proof. Let $c: X \rightsquigarrow Y$ be a channel. Then:

$$
\begin{aligned}
\operatorname{arr} \odot \mathcal{N}[K](c) & =\text { flat } \circ \mathcal{D}(\operatorname{arr}) \circ p m l \circ \mathcal{N}(c) & & \\
& =\text { flat } \circ \mathcal{D}(\bigotimes) \circ \operatorname{arr} \circ \mathcal{N}(c) & & \text { by Proposition 3.7.1 } \\
& =\text { flat } \circ \mathcal{D}(\bigotimes) \circ \mathcal{D}(c \times \cdots \times c) \circ \text { arr } & & \text { by naturality of arr } \\
& =\text { flat } \circ \mathcal{D}(c \otimes \cdots \otimes c) \circ \text { arr } & & \\
& =c^{K} \circ \text { arr. } & &
\end{aligned}
$$

For accumulation the required equality acc $\odot c^{K}=\mathcal{N}[K](c) \odot$ acc can be
obtained via a diagram chase:


The upper part is ordinary naturality of acc and the lower is the third formulation 3.34) of pml.

For naturality of draw-delete we use the equation $D D \odot p m l=p m l \odot D D$ from Proposition 3.7.5, together with naturality of $D D$ in:

$$
\begin{aligned}
D D \circ \mathcal{N}[K+1](c) & =\text { flat } \circ \mathcal{D}(D D) \circ p m l \circ \mathcal{N}(c) \\
& =\text { flat } \circ \mathcal{D}(p m l) \circ D D \circ \mathcal{N}(c) \\
& =\text { flat } \circ \mathcal{D}(p m l) \circ \mathcal{D}(\mathcal{N}(c)) \circ D D \\
& =\text { flat } \circ \mathcal{D}(\mathcal{N}[K](c)) \circ D D \\
& =\mathcal{N}[K](c) \circ D D .
\end{aligned}
$$

Not only acc, arr and $D D$ 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 $\mathcal{D}:$ Sets $\rightarrow$ Sets to a functor $\overline{\mathcal{D}}:$ Chan $\rightarrow$ Chan. We have already seen this (general) construction in Exercise 1.11.8. We recall that $\overline{\mathcal{D}}(X)=\mathcal{D}(X)$ and $\overline{\mathcal{D}}(c)=c »=(-): \mathcal{D}(X) \rightarrow \mathcal{D}(Y)$ for a channel $c: X \mapsto Y$. This extension $\overline{\mathcal{D}}$ is monoidal by a promotion of the usual tensor $\otimes: \mathcal{D}(X) \times \mathcal{D}(Y) \rightarrow \mathcal{D}(X \times Y)$ to a deterministic channel $\langle\otimes\rangle: \overline{\mathcal{D}}(X) \times \overline{\mathcal{D}}(Y) \rightsquigarrow \overline{\mathcal{D}}(X \times Y)$.

Theorem 3.7.12. The multinomial and hypergeometric channels are natural with respect to extended functors:

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 $c: X \leadsto Y$ one has:

$$
\begin{align*}
\mathcal{N}[K](c) \gg m n[K](\omega) & =m n[K](c \gg=\omega)  \tag{3.43}\\
\mathcal{N}[K](c) \gg h g[K](v) & =h g[K] \gg \mathcal{N}[L](c)(v),
\end{align*}
$$

for urns $\omega \in \mathcal{D}(X)$ and $v \in \mathcal{N}[L](X)$.
Proof. We can write the multinomial channel as $m n[K]=$ acc $\odot \operatorname{iid}[K]$, see Theorem 2.6.7 Lemma 3.7.11 tells that accumulation is a natural transformation acc: $(-)^{K} \Rightarrow \mathcal{N}[K]$ between extended functors. Similarly, iid $[K]$ forms a natural transformation $\overline{\mathcal{D}} \Rightarrow(-)^{K}$ by Exercise 2.4.8 (4). The composition $m n[K]$ is monoidal by Corollary 3.3.5.
The hypergeometric $h g[K]$ channel is an iteration of draw-delete's, see Theorem 3.4.1 It is thus natural with respect to channels, since draw-delete is, see Lemma 3.7.11 By the same argument this hypergeometric natural transformation is monoidal, see Lemma 3.2.8.

The Pólya channel from Section 3.5 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 3.5.2 (1), and indeed, functions do commute with copying, see Exercise 2.5.2.

### 3.7.1 Sampling of pushforward

We recall from Subsection 2.2.1 the programming language notation $x \leftarrow \omega$ for sampling an arbitrary element $x \in \operatorname{supp}(\omega)$ from the (support of the) distribution $\omega \in \mathcal{D}(X)$, in accordance with the probabilities of $\omega$. The sampling mechanism works 'correctly' if we accumulate a sufficiently long list of sampled elements to a multiset, then the resulting distribution obtained via frequentist learning approximates $\omega$. 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 3.3.3.


In this situation we sample $K$ elements from the distribution $\omega$, where the each multiset of $K$ elements comes with its own multinomial probability. In the above composition Flrn $\odot m n[K]$, these probabilities are taken into account. Recall that the sample channel sam: $\mathcal{D}(X) \leadsto X$ is the identity function id : $\mathcal{D}(X) \rightarrow \mathcal{D}(X)$. This works for any number $K$.

In 2.24) 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 $c: X \leadsto Y$. We suggested that the resulting
elements $y$ from samples of the pushforward distribution $c \gg \omega$ on $Y$. We now ask ourselves: is there a corresponding form of correctness?

We can use a multinomial distribution $m n[K](\omega)$ to sample $K$-sized multisets $\varphi \in \mathcal{N}[K](X)$ from $\omega$. Then for each element $x$ in $\varphi$ 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 $m n[L](c):=m n[L] \circ c: X \rightarrow \mathcal{D}(\mathcal{N}[L](Y))$, as already mentioned in Definition 2.6.6,
- the extension of the multiomial functor to channels, as in: $\mathcal{N}[K](c):=p m l \circ$ $\mathcal{N}(c): \mathcal{N}[K](X) \rightarrow \mathcal{D}(\mathcal{N}[K](Y))$, from Corollary 3.7.8

We combine these points in the following composite:

$$
\begin{equation*}
\mathcal{D}(X) \xrightarrow{m n[K]} \mathcal{N}[K](X) \xrightarrow{\mathcal{N}[K](m n[L](c))} \mathcal{N}[L](\mathcal{N}[K](Y)) \xrightarrow{f l a a^{\mathcal{N}}} \mathcal{N}[L \cdot K](Y) \tag{3.45}
\end{equation*}
$$

It first samples from a distribution $\omega \in \mathcal{D}(X)$, then applies $m n[L](c)$ 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 (3.45) is pushforward along the channel $c$. That is the context of the next result.

Proposition 3.7.13. Sequential sampling is correct, in the sense that the following diagram commutes.


Proof. We break the above rectangle up in three subdiagrams:


The triangle on the left is 3.44). The diamond at the top commutes by Theo-
rem 3.7.4

$$
\begin{aligned}
\text { flat }^{\mathcal{N}} \odot \mathcal{N}[K](m n[L](c)) & =\mathcal{D}\left(f l a t^{\mathcal{N}}\right) \circ p m l \circ \mathcal{N}(m n[L] \circ c) \\
& =\mathcal{D}\left(\text { flat }^{\mathcal{N}}\right) \circ m n[L] \circ p m l \circ \mathcal{N}(c) \\
& =\mathcal{D}\left(f l a t^{\mathcal{N}}\right) \circ m n[L](\mathcal{N}[K](c)) \\
& =f l a t^{\mathcal{N}} \odot m n[L](\mathcal{N}[K](c)) .
\end{aligned}
$$

Next, the rectangle at the bottom commutes:

$$
\begin{array}{ll}
\text { Flrn } \odot \text { flat }^{\mathcal{N}} \odot m n[L](\mathcal{N}[K](c)) & \\
=\text { flat } \circ \mathcal{D}(F l r n) \circ \mathcal{D}\left(f l a t^{\mathcal{N}}\right) \circ m n[L] \circ p m l \circ \mathcal{N}(c) \\
=\text { flat } \circ \mathcal{D}(F l r n) \circ p m l \circ \mathcal{N}(c) & \text { by Theorem 3.3.7 (2) } \\
=\text { flat } \circ \text { Flrn } \circ \mathcal{N}(c) & \text { by Theorem } 3.7 .2 \\
=\text { flat } \circ \mathcal{D}(c) \circ \text { Flrn } & \text { by naturality } \\
=c \odot \text { Flrn. } &
\end{array}
$$

There is also 'anchestral' sampling of joint states $\tau \in \mathcal{D}(X \times Y)$ for which we have a graph representation $\tau=\langle i d, c\rangle\rangle=\omega$. In general, such a graph form can be obtained via 'disintegration', see Section 7.2, but here we assume it given. This anchestral sampling works as follows.

$$
\begin{align*}
& \mathrm{x} \leftarrow \omega \\
& \mathrm{y} \leftarrow c(\mathrm{x})  \tag{3.46}\\
& \text { return }(\mathrm{x}, \mathrm{y})
\end{align*}
$$

Proposition 3.7.14. Ancestral sampling is correct, in the sense that the following diagram commutes.


Proof. This follows from Proposition 3.7.13 once we show that the following diagram commutes:


This requires an application of Corollary 3.3.5 in:


## Exercises

3.7.1 Consider the two distributions $\omega, \rho$ in 3.31 and check yourself the following equation, which is an instance of Theorem 3.7.2

$$
\begin{aligned}
\text { Flrn }>=\operatorname{pml}(2|\omega\rangle+1|\rho\rangle) & =\frac{17}{36}|a\rangle+\frac{19}{36}|b\rangle . \\
& =(\text { flat } \circ \text { Flrn })(2|\omega\rangle+1|\rho\rangle) .
\end{aligned}
$$

3.7.2 Check that the following diagram does not commute.


Consider for instance the multiset of multisets $2|1| a\rangle+2|b\rangle\rangle+1|3| a\rangle\rangle$.
3.7.3 Check that the construction of Corollary 3.7 .8 indeed yields a functor $\mathcal{N}[K]$ : Chan $\rightarrow$ Chan, i.e. that identities and composition are preserved.
3.7.4 Consider the channel $c:\{1,2,3\} \mapsto\{a, b\}$ given by:
$c(1)=\frac{1}{6}|a\rangle+\frac{5}{6}|b\rangle$
$c(2)=\frac{2}{3}|a\rangle+\frac{1}{3}|b\rangle$
$c(3)=\frac{1}{2}|a\rangle+\frac{1}{2}|b\rangle$.

Show that the channel $\mathcal{N}[K](c): \mathcal{N}[6](\{1,2,3\}) \rightarrow \mathcal{N}[6](\{a, b\})$ satisfies:

$$
\begin{aligned}
& \mathcal{N}[K](c)(1|1\rangle+2|2\rangle+3|3\rangle) \\
& \left.\left.\left.\left.\left.\left.\left.\left.=\frac{1}{108}|6| a\right\rangle\right\rangle+\frac{1}{12}|5| a\right\rangle+1|b\rangle\right\rangle+\frac{35}{144}|4| a\right\rangle+2|b\rangle\right\rangle+\frac{1}{3}|3| a\right\rangle+3|b\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.\quad+\frac{17}{72}|2| a\right\rangle+4|b\rangle\right\rangle+\frac{1}{12}|1| a\right\rangle+5|b\rangle\right\rangle+\frac{5}{432}|6| b\right\rangle\right\rangle .
\end{aligned}
$$

3.7.5 Consider a distribution $\omega \in \mathcal{D}(X)$ as a channel $1 \leadsto X$. Recall the extension of the functors $(-)^{K}$ and $\mathcal{N}[K]$ to the category of channels, see Corollary 3.7.8 and show that:
3.7.6 Show that frequentist learning is a natural transformation in:

3.7.7 Show that the extended functors $\mathcal{N}[K]$ : Chan $\rightarrow$ Chan commute with sums of multisets: for a channel $f: X \leadsto Y$,

3.7.8 The parallel multinomial law pml does not commute with tensors (of multisets and distributions), as in the following diagram.


Take for instance $X=\{a, b\}, Y=\{0,1\}$ with $K=2, L=1$ with distributions: $\left\{\begin{array}{l}\omega=\frac{3}{4}|a\rangle+\frac{1}{4}|b\rangle \\ \rho=\frac{2}{3}|0\rangle+\frac{1}{3}|1\rangle\end{array} \quad\right.$ and multisets: $\left\{\begin{array}{l}\varphi=2|\omega\rangle \\ \psi=1|\rho\rangle\end{array}\right.$

1 Calculate:

$$
\begin{aligned}
& (\otimes \odot(p m l \otimes p m l)(\varphi, \psi) \\
& \left.\left.\left.\left.\left.\left.=\frac{3}{8}|2| a, 0\right\rangle\right\rangle+\frac{1}{4}|1| a, 0\right\rangle+1|b, 0\rangle\right\rangle+\frac{1}{24}|2| b, 0\right\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.\quad \quad+\frac{3}{16}|2| a, 1\right\rangle\right\rangle+\frac{1}{8}|1| a, 1\right\rangle+1|b, 1\rangle\right\rangle+\frac{1}{48}|2| b, 1\right\rangle\right\rangle .
\end{aligned}
$$

2 And also:

$$
\begin{aligned}
& (p m l \odot \mathcal{M}[K \cdot L](\otimes) \odot \otimes)(\varphi, \psi) \\
& \left.\left.\left.\left.\left.\left.\left.\left.=\frac{1}{4}|2| a, 0\right\rangle\right\rangle+\frac{1}{16}|2| a, 1\right\rangle\right\rangle+\frac{1}{36}|2| b, 0\right\rangle\right\rangle+\frac{1}{144}|2| b, 1\right\rangle\right\rangle \\
& \left.\left.\left.\left.\quad+\frac{1}{4}|1| a, 0\right\rangle+1|a, 1\rangle\right\rangle+\frac{1}{6}|1| a, 0\right\rangle+1|b, 0\rangle\right\rangle \\
& \left.\left.\left.\left.\quad+\frac{1}{12}|1| a, 0\right\rangle+1|b, 1\rangle\right\rangle+\frac{1}{12}|1| a, 1\right\rangle+1|b, 0\rangle\right\rangle \\
& \left.\left.\left.\left.\quad+\frac{1}{24}|1| a, 1\right\rangle+1|b, 1\rangle\right\rangle+\frac{1}{36}|1| b, 0\right\rangle+1|b, 1\rangle\right\rangle .
\end{aligned}
$$

3.7.9 Let $\Omega \in \mathcal{D}^{2}(X)$ be a distribution of distributions. Argue that the following two program fragments yield the same outcome.
$\bar{x} \leftarrow \operatorname{flat}(\Omega)$
return $x$

| $\omega \leftarrow \Omega$ |
| :--- |
| $\mathrm{x} \leftarrow \omega$ |
| return x |

Hint: Recall from Exercise 2.4.4 that flat $(\Omega)=s a m \gg \Omega$.

### 3.8 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 $p m l[K]: \mathcal{N}[K](\mathcal{D}(X)) \rightarrow \mathcal{D}(\mathcal{N}[K](X))$ in a restricted manner, namely restricted to multisets of size $K$. The same holds for the tensor map $\otimes[K]: \mathcal{D}(X)^{K} \rightarrow \mathcal{D}\left(X^{K}\right)$. What if we drop this size restriction?
The fourth formulation (3.38) already describes pml as a map $\mathcal{N D}(X) \rightarrow$ $\mathcal{D N}(X)$, without size restrictions. It forms a natural transformation $\mathcal{N D} \Rightarrow$ $\mathcal{D N}$ by Exercise 3.6.3. There is also such a formulation for the tensor $\otimes[K]$.

Definition 3.8.1. For a set $X$ we define the parallel states map pst : $\mathcal{L}(\mathcal{D}(X)) \rightarrow$ $\mathcal{D}(\mathcal{L}(X))$ as:

$$
\begin{align*}
\operatorname{pst}\left(\left[\omega_{1}, \ldots, \omega_{K}\right]\right) & :=\sum_{\vec{x} \in \mathcal{L}(X), \| \vec{x} \mid=K} \prod_{1 \leq i \leq K} \omega_{i}\left(x_{i}\right)|\vec{x}\rangle \\
& =\sum_{\vec{x} \in X^{K}} \bigotimes[K](\vec{\omega})(\vec{x})|\vec{x}\rangle  \tag{3.47}\\
& =\bigotimes[K](\vec{\omega}) . \\
& =\omega_{1} \otimes \cdots \otimes \omega_{K} .
\end{align*}
$$

By Lemma 2.3.5 1) this yields a natural transformation pst: $\mathcal{L D} \Rightarrow \mathcal{D} \mathcal{L}$.
We collect some basic results. As usual with multiple functor applications $F(G(X))$ we often spare on parentheses and simply write $F G(X)$ for $F(G(X))$.

Proposition 3.8.2. Let $X$ be an arbitrary set.
1 Parallel multinomial and parallel states commute with accumulation in:


2 The parallel states law pst commutes in the following way with the unit and flatten operations of the list monad $\mathcal{L}$.



3 The parallel multinomial law pml: $\mathcal{N D}(X) \rightarrow \mathcal{D N}(X)$ commutes in the following way with the unit and flatten operations of the (natural) multiset monad $\mathcal{N}$.


4 The parallel multinomial law pml: $\mathcal{N D}(X) \rightarrow \mathcal{D N}(X)$ is a map of monoids.
The analogous property for pst : $\mathcal{L D}(X) \rightarrow \mathcal{D} \mathcal{L}(X)$ 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:

$$
\operatorname{pst}\left(\ell_{1}+\ell_{2}\right)=\mathcal{D}(+)\left(p s t\left(\ell_{1}\right) \otimes \operatorname{pst}\left(\ell_{2}\right)\right)
$$

Proof. 1 For a list $\vec{\omega} \in \mathcal{L D}(X)$, say of length $K \in \mathbb{N}$, with $\operatorname{acc}(\vec{\omega})=\sum_{i} n_{i}\left|\omega_{i}\right\rangle$, we have:

$$
\begin{aligned}
(\mathcal{D}(\mathrm{acc}) \circ p s t)(\vec{\omega}) & =\sum_{\vec{x} \in X^{K}} \prod_{1 \leq i \leq K} \omega_{i}\left(x_{i}\right)|\operatorname{acc}(\vec{x})\rangle \\
& =\sum_{\vec{x} \in X^{K}}\left(\otimes_{i} \omega_{i}^{n_{i}}\right)(\vec{x})|\operatorname{acc}(\vec{x})\rangle \\
& \stackrel{\text { 3.32] }}{=} \operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right)=(p m l \circ \operatorname{acc})(\vec{x}) .
\end{aligned}
$$

2 The equation for the units is easy. For $\omega \in \mathcal{D}(X)$ we have:

$$
(\text { pst } \circ \text { unit })(\omega)=p s t([\omega])=\sum_{x \in X} \omega(x)|[x]\rangle=\mathcal{D}(\text { unit })(\omega) .
$$

Let lists of distributions $L_{1}, \ldots, L_{K} \in \mathcal{L} \mathcal{D}(X)$ be given with $\left\|L_{i}\right\|=N_{i}$. We write $L:=\operatorname{flat}\left(\left[L_{1}, \ldots, L_{K}\right]\right)=L_{1}+\cdots+L_{K}$, see Exercise 1.4.5, so that

$$
\begin{aligned}
& \|L\|=\sum_{i} N_{i} . \text { Then: } \\
& (\mathcal{D}(\text { flat }) \circ p s t \circ \mathcal{L}(p s t))(\vec{L}) \\
& =(\mathcal{D}(\text { flat }) \circ p s t)\left(\left[p s t\left(L_{1}\right), \ldots, p s t\left(L_{K}\right)\right]\right) \\
& \left.=\sum_{\ell_{1} \in X^{N_{1}}, \ldots, \ell_{K} \in X^{N_{K}}}\left(p s t\left(L_{1}\right) \otimes \cdots \otimes p s t\left(L_{K}\right)\right)\left(\ell_{1}, \ldots, \ell_{K}\right) \mid \text { flat }\left(\left[\ell_{1}, \ldots, \ell_{K}\right]\right)\right\rangle \\
& =\sum_{\ell_{1} \in X^{N_{1}}, \ldots, \ell_{K} \in X^{N_{K}}} \prod_{1 \leq i \leq K} p s t\left(L_{i}\right)\left(\ell_{i}\right)\left|\ell_{1}+\cdots+\ell_{K}\right\rangle \\
& =\sum_{\ell_{1} \in X^{N_{1}}, \ldots, \ell_{K} \in X^{N_{K}}} \prod_{1 \leq i \leq K} \prod_{1 \leq j \leq N_{i}} L_{i j}\left(\ell_{i j}\right)\left|\ell_{1}+\cdots+\ell_{K}\right\rangle \\
& =\sum_{\ell \in X^{N_{1}+\cdots+N_{K}}} \prod_{1 \leq k \leq \Sigma_{i} N_{i}} L_{k}\left(\ell_{k}\right)|\ell\rangle \\
& =p s t(L)=p s t\left(f l a t\left(\left[L_{1}, \ldots, L_{K}\right]\right)\right)=(p s t \circ \text { flat })(\vec{L}) .
\end{aligned}
$$

3 First, for $\omega \in \mathcal{D}(X)$ we have, by Exercise 3.6.1.

$$
\begin{aligned}
(p m l \circ u n i t)(\omega)=\operatorname{pml}(1|\omega\rangle)=m n[1](\omega) & \left.\left.=\sum_{x \in X} \omega(x)|1| x\right\rangle\right\rangle \\
& =\mathcal{D}(\text { unit })(\omega) .
\end{aligned}
$$

For flatten we have to do a bit more work. We recall from 3.36, for a commutative monoid $M$, the $\mathcal{N}$-algebra sum: $\mathcal{N}(\mathcal{D}(M)) \rightarrow \mathcal{D}(M)$ given by convolution. We also recall that by Proposition 1.6.6 the following two diagrams commute, where $f: M_{1} \rightarrow M_{2}$ is a map of (commutative) monoids.


We use the fourth formulation (3.38) namely pml $=$ sum $\circ \mathcal{N D}$ (unit). Then:

$$
\begin{aligned}
& \mathcal{D}(f l a t) \circ p m l \circ \mathcal{N}(p m l) \\
& =\mathcal{D}(\text { flat }) \circ \text { sum } \circ \mathcal{N D}(\text { unit }) \circ \mathcal{N}(\mathrm{pml}) \quad \text { by 3.38) } \\
& =\operatorname{sum} \circ \mathcal{N D}(\text { flat }) \circ \mathcal{N D}(\text { unit }) \circ \mathcal{N}(p m l) \text { by (3.48), on the right } \\
& =\operatorname{sum} \circ \mathcal{N}(p m l) \\
& =\operatorname{sum} \circ \mathcal{N}(\text { sum } \circ \mathcal{N} \mathcal{D}(\text { unit })) \\
& =\text { sum } \circ \text { flat } \circ \mathcal{N} \mathcal{N D} \text { (unit) } \\
& =\operatorname{sum} \circ \mathcal{N} \mathcal{D}(\text { unit }) \circ \text { flat } \\
& =p m l \circ \text { flat } \\
& \text { by a flatten-unit law } \\
& \text { by (3.38) again } \\
& \text { by (3.48, on the left } \\
& \text { by naturality } \\
& \text { once again by 3.38. }
\end{aligned}
$$

4 In order to show that the parallel multinomial law pml: $\mathcal{N D}(X) \rightarrow \mathcal{D N}(X)$ is a map of monoids it suffices by Proposition 1.6 .6 to show that the diagram (1.31) commutes. This is the outer rectangle in:


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 (3.48).

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:

$$
\begin{aligned}
& \mathcal{D}(++)(p s t(\vec{\omega}) \otimes p s t(\vec{\rho})) \\
& =\mathcal{D}(++)(\bigotimes[K](\vec{\omega}) \otimes \bigotimes[L](\vec{\rho})) \\
& =\sum_{\vec{x} \in X^{K}} \sum_{\vec{y} \in X^{L}} \bigotimes[K](\vec{\omega})(\vec{x}) \cdot \bigotimes[L](\vec{\rho})(\vec{y})|\vec{x}+\vec{y}\rangle \\
& =\sum_{\vec{z} \in X^{K+L}} \bigotimes[K+L](\vec{\omega}+\vec{\rho})(\vec{z})|\vec{z}\rangle \\
& =\bigotimes[K+L](\vec{\omega}+\vec{\rho})=\operatorname{pst}(\vec{\omega}+\vec{\rho}) .
\end{aligned}
$$

Theorem 3.8.3. The two composite functors

are both monads. Their unit and flatten maps are given as follows, via pst and pml.

- For $\mathcal{D} \mathcal{L}$,

$$
\begin{aligned}
& \text { unit }^{\mathcal{D} \mathcal{L}}:=(X \xrightarrow[\text { unit }]{\longrightarrow} \mathcal{L}(X) \xrightarrow[\mathcal{L}(\text { unit })]{\text { unit }} \mathcal{D}(X) \xrightarrow{\mathcal{D}(\text { unit })} \mathcal{L}(X)) \\
& \text { flat }^{\mathcal{D} \mathcal{L}}:=\left(\mathcal{D} \mathcal{L} \mathcal{D} \mathcal{L}(X) \xrightarrow{\mathcal{D}(p s t)} \mathcal{D}^{2} \mathcal{L}^{2}(X) \xrightarrow[\mathcal{D}^{2}(f l a t)]{\text { flat }} \mathcal{D} \mathcal{L}^{2}(X) \xrightarrow{\mathcal{D}(\text { flat })} \mathcal{L}(X) \xrightarrow[\text { flat }]{ } \mathcal{D} \mathcal{L}(X)\right)
\end{aligned}
$$

- For $\operatorname{DN}$,

$$
\begin{aligned}
& \text { unit }^{\mathcal{D N}}:=(X \underset{\text { unit }}{\text { unit }} \mathcal{D}(X) \xrightarrow[N]{ } \xrightarrow[\mathcal{N}(X)]{\substack{\text { (unit) }}} \mathcal{D N}(X)) \\
& f l a t^{\mathcal{D N}}:=\left(\mathcal{D N} \mathcal{D N}(X) \xrightarrow{\mathcal{D}(p m l)} \mathcal{D}^{2} \mathcal{N}^{2}(X) \xrightarrow[\mathcal{D}^{2}(\text { flat })]{\text { flat }} \mathcal{D N}^{2}(X) \xrightarrow{\mathcal{D}(\text { flat })} \mathcal{D}(X) \xrightarrow[\text { flat }]{\sim} \mathcal{N}(X)\right)
\end{aligned}
$$

Proof. This is a standard result in category theory, originally by Beck, see e.g. [13, 10, 75]. The two diamonds in the above descriptions of unit and flatten commute by naturality. The monad laws for $\mathcal{D} \mathcal{L}$ and $\mathcal{D N}$ hold by items (2) and (3) in Proposition 3.8.2.

In the remainder of this section we concentrate on the monad $\mathcal{D N}$ : Sets $\rightarrow$ Sets. The aim is to show that there is a map of monads $\mathcal{D N} \Rightarrow \mathcal{M}$, where $\mathcal{M}$ : Sets $\rightarrow$ 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 [38], under the name 'intensity'. Therefor we shall write it as intsy: $\mathcal{D N} \Rightarrow \mathcal{M}$.
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:

$$
\mathcal{D} \stackrel{\sigma}{\Longrightarrow} \mathcal{M} \quad \mathcal{N} \xlongequal{\tau} \mathcal{M} .
$$

These $\sigma$ and $\tau$ are maps of monads. They are used implicitly for instance in the equation flat $(p m l(\Psi))=$ flat $(\Psi)$ in Proposition 3.7.3 As a commuting diagram, it looks as follows.


Theorem 3.8.4 (From [38]). The intensity natural transformation intsy: $\mathcal{D N} \Rightarrow$ $\mathcal{M}$ has components:

$$
\begin{equation*}
\text { intsy }:=(\underset{\mathcal{D N}(X) \xrightarrow{\sigma} \mathcal{M} \mathcal{N}(X) \xrightarrow{\sim} \mathcal{D M}(X) \xrightarrow[\sigma]{\sim} \mathcal{M}(\tau)}{\longrightarrow} \mathcal{M}(X) \xrightarrow{\text { flat }} \mathcal{M}(X)) \tag{3.50}
\end{equation*}
$$

```
intsy \(\circ\) flat \({ }^{\mathcal{D N}}\)
    \(\stackrel{3.50}{-}\) flat \(^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(\tau) \circ \mathcal{D}\left(f l a t^{\mathcal{N}}\right) \circ\) flat \(^{\mathcal{D}} \circ \mathcal{D}(p m l)\)
    \(=\) flat \(^{\mathcal{M}} \circ \sigma \circ \mathcal{D}\left(\right.\) flat \(\left.^{\mathcal{M}}\right) \circ \mathcal{D \mathcal { M } ( \tau ) \circ \mathcal { D } ( \tau ) \circ \text { flat } ^ { \mathcal { D } } \circ \mathcal { D } ( p m l ) , ~ ( p )}\)
    \(=f l a t^{\mathcal{M}} \circ \mathcal{M}\left(f l a t^{\mathcal{M}}\right) \circ \mathcal{M}^{2}(\tau) \circ \mathcal{M}(\tau) \circ \sigma \circ f l a t^{\mathcal{D}} \circ \mathcal{D}(p m l)\)
    \(=f f a t^{\mathcal{M}} \circ f l a t^{\mathcal{M}} \circ \mathcal{M}^{2}(\tau) \circ \mathcal{M}(\tau) \circ f l a t^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(\sigma) \circ \mathcal{D}(p m l)\)
    \(=\) flat \(^{\mathcal{M}} \circ\) fla \(^{\mathcal{M}} \circ\) flat \(^{\mathcal{M}} \circ \mathcal{M}^{3}(\tau) \circ \mathcal{M}^{2}(\tau) \circ \sigma \circ \mathcal{D}(\sigma) \circ \mathcal{D}(p m l)\)
    \(=\) flat \(^{\mathcal{M}} \circ\) flat \(^{\mathcal{M}} \circ \mathcal{M}\left(\right.\) flat \(\left.^{\mathcal{M}}\right) \circ \mathcal{M}^{3}(\tau) \circ \sigma \circ \mathcal{D} \mathcal{M}(\tau) \circ \mathcal{D}(\sigma) \circ \mathcal{D}(p m l)\)
    \(=f l a t^{\mathcal{M}} \circ\) flat \(^{\mathcal{M}} \circ \mathcal{M}^{2}(\tau) \circ \mathcal{M}\left(f l a t^{\mathcal{M}}\right) \circ \sigma \circ \mathcal{D} \mathcal{M}(\tau) \circ \mathcal{D}(\sigma) \circ \mathcal{D}(p m l)\)
    \(=\) flat \(^{\mathcal{M}} \circ\) flat \(^{\mathcal{M}} \circ \sigma \circ \mathcal{D} \mathcal{M}(\tau) \circ \mathcal{D}\left(\right.\) flat \(\left.^{\mathcal{M}}\right) \circ \mathcal{D} \mathcal{M}(\tau) \circ \mathcal{D}(\sigma) \circ \mathcal{D}(p m l)\)
    \({ }^{3.49}\) flat \(^{\mathcal{M}} \circ \mathcal{M}\left(\right.\) flat \(\left.^{\mathcal{M}}\right) \circ \sigma \circ \mathcal{D} \mathcal{M}(\tau) \circ \mathcal{D}\left(\right.\) flat \(\left.^{\mathcal{M}}\right) \circ \mathcal{D}(\tau) \circ \mathcal{D N}(\sigma)\)
    \(=f l a t^{\mathcal{M}} \circ \sigma \circ \mathcal{D}\left(\right.\) flat \(\left.^{\mathcal{M}}\right) \circ \mathcal{D}\left(\right.\) flat \(\left.^{\mathcal{M}}\right) \circ \mathcal{D} \mathcal{M}^{2}(\tau) \circ \mathcal{D}(\tau) \circ \mathcal{D N}(\sigma)\)
    \(=f l a t^{\mathcal{M}} \circ \sigma \circ \mathcal{D}\left(f l a t^{\mathcal{M}}\right) \circ \mathcal{D} \mathcal{M}\left(\right.\) flat \(\left.^{\mathcal{M}}\right) \circ \mathcal{D} \mathcal{M}^{2}(\tau) \circ \mathcal{D}(\tau) \circ \mathcal{D N}(\sigma)\)
    \(=\) flat \(^{\mathcal{M}} \circ \mathcal{M}\left(\right.\) flat \(\left.^{\mathcal{M}}\right) \circ \sigma \circ \mathcal{D} \mathcal{M}\left(\right.\) flat \(\left.^{\mathcal{M}^{\mathcal{M}}}\right) \circ \mathcal{D} \mathcal{M}^{2}(\tau) \circ \mathcal{D}(\tau) \circ \mathcal{D N}(\sigma)\)
        flat \({ }^{\mathcal{M}} \circ\) flat \(^{\mathcal{M}} \circ \sigma \circ \mathcal{D}(\tau) \circ \operatorname{DN}\left(f l a t^{\mathcal{M}}\right) \circ \operatorname{DN} \mathcal{M}(\tau) \circ \mathcal{D N}(\sigma)\)
        flat \({ }^{\mathcal{M}} \circ\) intsy \(\circ \mathcal{D N}(\) intsy \()\).
```

Figure 3.1 Equational proof that the intensity natural transformation intsy from 3.50, commutes with flattens, as part of the proof of Theorem 3.8.4

This intensity intsy is a map of monads. It thus induces a functor $\mathbf{C h a n}(\mathcal{D N}) \rightarrow$ Chan $(\mathcal{M})$ between the associated categories of channels.

This definition 3.50 looks impressive, but for practical purposes we can just write intsy $(\Psi)=$ flat $(\Psi)$, leaving inclusions $\sigma, \tau$ 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 ${ }^{\mathcal{D N}}$ for the unit of $\mathcal{D N}$, see Theorem 3.8.3 Then:

```
intsy \(\circ\) unit \({ }^{\mathcal{D N}}\)
    \(\stackrel{3.50}{=}\) flat \({ }^{\mathcal{M}} \circ \mathcal{M}(\tau) \circ \sigma \circ \mathcal{D}\left(\right.\) unit \(\left.^{\mathcal{N}}\right) \circ u n i t^{\mathcal{D}}\)
    \(=f l a t^{\mathcal{M}} \circ \mathcal{M}(\tau) \circ \mathcal{M}\left(u_{n i t}{ }^{\mathcal{N}}\right) \circ \sigma \circ u n i t^{\mathcal{D}}\)
    \(=\) flat \(^{\mathcal{M}} \circ \mathcal{M}\left(u_{n i t}{ }^{\mathcal{M}}\right) \circ u n i t^{\mathcal{M}} \quad\) since \(\tau, \sigma\) are maps of monads
    \(=u n i t^{\mathcal{M}}\).
```

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 3.8.5. For each set $X$, the intensity map is a homomorphism of
monoids, so that we can rephrase Diagram (3.49) as a triangle of monoid homomorphisms:


Proof. We use the convolution monoid structure on $\mathcal{D N}(X)$, defined in Proposition 2.7.2, in the form of a map sum: $\mathcal{N D} \mathcal{N}(X) \rightarrow \mathcal{D N}(X)$, as above. The zero element in $\mathcal{D N}(X)$ is $1|0\rangle$, where $\mathbf{0} \in \mathcal{N}(X)$ is the empty multiset. It satisfies, for $x \in X$,

$$
\operatorname{intsy}(1|\mathbf{0}\rangle)(x)=\operatorname{flat}(1|\mathbf{0}\rangle)(x)=1 \cdot \mathbf{0}(x)=0 .
$$

Hence intsy $(1|\mathbf{0}\rangle)=\mathbf{0}$.
For distributions $\omega, \rho \in \mathcal{D N}(X)$ we have intsy $(\omega+\rho)=\operatorname{intsy}(\omega)+\operatorname{intsy}(\rho)$ since for each $x \in X$,

$$
\begin{aligned}
& \text { intsy }(\omega+\rho)(x)=\operatorname{flat}(\omega+\rho)(x)=\sum_{\varphi \in \mathcal{N}(X)}(\omega+\rho)(\varphi) \cdot \varphi(x) \\
& =\sum_{\varphi \in \mathcal{N}(X)} \mathcal{D}(+)(\omega \otimes \rho)(\varphi) \cdot \varphi(x) \\
& =\sum_{\psi, \chi \in \mathcal{N}(X)} \omega(\psi) \cdot \rho(\chi) \cdot(\psi+\chi)(x) \\
& =\sum_{\psi, \chi \in \mathcal{N}(X)} \omega(\psi) \cdot \rho(\chi) \cdot(\psi(x)+\chi(x)) \\
& =\sum_{\psi \in \mathcal{N}(X)} \omega(\psi) \cdot\left(\sum_{\chi \in \mathcal{N}(X)} \rho(\chi)\right) \cdot \psi(x)+\sum_{\chi \in \mathcal{N}(X)}\left(\sum_{\psi \in \mathcal{N}(X)} \omega(\psi)\right) \cdot \rho(\chi) \cdot \chi(x) \\
& =\operatorname{flat}(\omega)(x)+\operatorname{flat}(\rho)(x) \\
& =(\operatorname{intsy}(\omega)+\operatorname{intsy}(\rho))(x) .
\end{aligned}
$$

## Exercises

3.8.1 Check that the intensity natural transformation intsy: $\mathcal{D N} \Rightarrow \mathcal{M}$ defined in (3.50) restricts to $\mathcal{D N}[K] \Rightarrow \mathcal{M}[K]$, for each $K \in \mathbb{N}$.
3.8.2 Check that the 'mean' results for multinomial, hypergeometric and Pólya distributions from Proposition 3.3.6 and from Lemmas 3.4.5 (2)
and 3.5.1 (2) can be reformulated in terms of intensity as:

$$
\begin{aligned}
\operatorname{intsy}(\operatorname{mn}[K](\omega)) & =K \cdot \omega \\
\operatorname{intsy}(\operatorname{hg}[K](\psi)) & =K \cdot \operatorname{Flrn}(\psi) \\
\operatorname{intsy}(\operatorname{pl}[K](\psi)) & =K \cdot \operatorname{Flrn}(\psi)
\end{aligned}
$$

3.8.3 In Corollary 3.7.8 we have seen the extended functor $\mathcal{N}[K]:$ Chan $\rightarrow$ Chan. The flatten operation flat: $\mathcal{N} \mathcal{N} \Rightarrow \mathcal{N}$ for (natural) multisets, from Subsection 1.6.2, restricts to $\mathcal{N}[K] \mathcal{N}[L] \Rightarrow \mathcal{N}[K \cdot L]$, making $\mathcal{N}[K]$ : Sets $\rightarrow$ Sets into what is called a graded monad, see e.g. [134, 59]. Also the extended functor $\mathcal{N}[K]$ : Chan $\rightarrow$ Chan is such a graded monad, essentially by Proposition 3.8.2

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.


Elaborating a counterexample is quite initimidating, so we proceed step by step. We take as spaces:

$$
X=\{a, b\} \quad \text { and } \quad Y=\{0,1\} .
$$

We use $K=2$ and $L=3$ for the multisets of multisets:

$$
\begin{aligned}
& \Phi=1|2| a\rangle+1|b\rangle\rangle+1|1| a\rangle+2|b\rangle\rangle \in \mathcal{N}[2] \mathcal{N}[3](X) \\
& \Psi=1|2| 0\rangle+1|1\rangle\rangle+1|3| 1\rangle\rangle \in \mathcal{N}[2] \mathcal{N}[3](Y) .
\end{aligned}
$$

1 Check that going east-south in the above diagram yields:

$$
\begin{aligned}
\operatorname{mzip}(\text { flat }(\Phi), \text { flat }(\Psi))= & \operatorname{mzip}(3|a\rangle+3|b\rangle, 2|0\rangle+4|1\rangle) \\
= & \left.\left.\frac{1}{5}|3| a, 1\right\rangle+2|b, 0\rangle+1|b, 1\rangle\right\rangle \\
& \left.\left.+\frac{3}{5}|1| a, 0\right\rangle+2|a, 1\rangle+1|b, 0\rangle+2|b, 1\rangle\right\rangle \\
& \left.\left.+\frac{1}{5}|2| a, 0\right\rangle+1|a, 1\rangle+3|b, 1\rangle\right\rangle
\end{aligned}
$$

2 The other path, south-east, will be done in several steps. Write $\Phi=$ $1\left|\varphi_{1}\right\rangle+1\left|\varphi_{2}\right\rangle$ and $\Psi=1\left|\psi_{1}\right\rangle+1\left|\psi_{2}\right\rangle$ where:

$$
\left\{\begin{array} { l } 
{ \varphi _ { 1 } = 2 | a \rangle + 1 | b \rangle } \\
{ \varphi _ { 2 } = 1 | a \rangle + 2 | b \rangle }
\end{array} \quad \left\{\begin{array}{l}
\psi_{1}=2|0\rangle+1|1\rangle \\
\psi_{2}=3|1\rangle .
\end{array}\right.\right.
$$

Show then that:

$$
\left.\left.\left.\left.\operatorname{mzip}(\Phi, \Psi)=\frac{1}{2}|1| \varphi_{1}, \psi_{1}\right\rangle+1\left|\varphi_{2}, \psi_{2}\right\rangle\right\rangle+\frac{1}{2}|1| \varphi_{1}, \psi_{2}\right\rangle+1\left|\varphi_{2}, \psi_{1}\right\rangle\right\rangle .
$$

3 Show next:

$$
\begin{aligned}
& \left.\left.\left.\left.\operatorname{mzip}\left(\varphi_{1}, \psi_{1}\right)=\frac{2}{3}|1| a, 0\right\rangle+1|a, 1\rangle+1|b, 0\rangle\right\rangle+\frac{1}{3}|2| a, 0\right\rangle+1|b, 1\rangle\right\rangle \\
& \left.\left.\operatorname{mzip}\left(\varphi_{1}, \psi_{2}\right)=1|2| a, 1\right\rangle+1|b, 1\rangle\right\rangle \\
& \left.\left.\left.\left.\operatorname{mzip}\left(\varphi_{2}, \psi_{1}\right)=\frac{1}{3}|1| a, 1\right\rangle+2|b, 0\rangle\right\rangle+\frac{2}{3}|1| a, 0\right\rangle+1|b, 0\rangle+1|b, 1\rangle\right\rangle \\
& \left.\left.\operatorname{mzip}\left(\varphi_{2}, \psi_{2}\right)=1|1| a, 1\right\rangle+2|b, 1\rangle\right\rangle .
\end{aligned}
$$

4 Show now that:

$$
\begin{aligned}
& \operatorname{pml}\left(1\left|\operatorname{mzip}\left(\varphi_{1}, \psi_{1}\right)\right\rangle+1\left|\operatorname{mzip}\left(\varphi_{2}, \psi_{2}\right)\right\rangle\right) \\
& \left.\left.\left.\left.=\frac{2}{3}|1| 1|a, 0\rangle+1|a, 1\rangle+1|b, 0\rangle\right\rangle+1|1| a, 1\right\rangle+2|b, 1\rangle\right\rangle\right\rangle \\
& \left.\left.\left.\left.\quad \quad+\frac{1}{3}|1| 2|a, 0\rangle+1|b, 1\rangle\right\rangle+1|1| a, 1\right\rangle+2|b, 1\rangle\right\rangle\right\rangle \\
& \operatorname{pml}\left(1\left|\operatorname{mzip}\left(\varphi_{1}, \psi_{2}\right)\right\rangle+1\left|\operatorname{mzip}\left(\varphi_{2}, \psi_{1}\right)\right\rangle\right) \\
& \left.\left.\left.\left.=\frac{1}{3}|1| 1|a, 1\rangle+2|b, 0\rangle\right\rangle+1|2| a, 1\right\rangle+1|b, 1\rangle\right\rangle\right\rangle+ \\
& \left.\left.\left.\left.\quad+\frac{2}{3}|1| 1|a, 0\rangle+1|b, 0\rangle+1|b, 1\rangle\right\rangle+1|2| a, 1\right\rangle+1|b, 1\rangle\right\rangle\right\rangle .
\end{aligned}
$$

5 Finally, check that the south-east past yields:

$$
\begin{aligned}
& (\text { flat } \odot \mathcal{N}[2](\text { mzip }) \odot \text { mzip })(\Phi, \Psi) \\
& \left.\left.\quad=\frac{2}{3}|1| a, 0\right\rangle+2|a, 1\rangle+1|b, 0\rangle+2|b, 1\rangle\right\rangle \\
& \left.\left.\quad+\frac{1}{6}|2| a, 0\right\rangle+3|b, 1\rangle+1|a, 1\rangle\right\rangle \\
& \left.\left.\quad+\frac{1}{6}|3| a, 1\right\rangle+2|b, 0\rangle+1|b, 1\rangle\right\rangle .
\end{aligned}
$$

This differs from what we get in the first item, via the east-south route.

### 3.9 Discrete Poisson point processes

In the previous section we have seen the composite monads $\mathcal{D} \mathcal{L}$ and $\mathcal{D N}$. 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 a composite distribution on (natural) multisets of arbitrary size. We actually stretch things a bit, because we are not using $\mathcal{D} \mathcal{L}$ and $\mathcal{D N}$ but the infinite version $\mathcal{D}_{\infty} \mathcal{L}$ and $\mathcal{D}_{\infty} \mathcal{N}$. This does not change things fundamentally.

Definition 3.9.1. For an arbitrary set $X$ and an intensity parameter $\lambda>0$ we define the Poisson-iid and Poisson-multinomial maps:

$$
\mathcal{D}(X) \xrightarrow{\text { Piid }[\lambda]} \mathcal{D}_{\infty}(\mathcal{L}(X)) . \quad \mathcal{D}(X) \xrightarrow{\text { Pmn }[\lambda]} \mathcal{D}_{\infty}(\mathcal{N}(X)) .
$$

On a distribution $\omega \in \mathcal{D}(X)$ they are given as follows.
1 The Poisson-idd is:

$$
\begin{align*}
\operatorname{Piid}[\lambda](\omega) & :=\sum_{\ell \in \mathcal{L}(\operatorname{supp}(\omega))} e^{-\lambda} \cdot \frac{\lambda^{\|\ell\|}}{\|\ell\|!} \cdot \prod_{1 \leq i \leq\|\ell \ell\|} \omega\left(\ell_{i}\right)|\ell\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\vec{x} \in X^{K}} e^{-\lambda} \cdot \frac{\lambda^{K}}{K!} \cdot \prod_{1 \leq i \leq K} \omega\left(x_{i}\right)|\vec{x}\rangle  \tag{3.51}\\
& =\sum_{K \in \mathbb{N}} \sum_{\vec{x} \in X^{K}} \operatorname{pois}[\lambda](K) \cdot \operatorname{iid}[K](\omega)(\vec{x})|\vec{x}\rangle \\
& =\operatorname{iid}[-](\omega) \gg=\operatorname{pois}[\lambda] .
\end{align*}
$$

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:

$$
\begin{align*}
\operatorname{Pmn}[\lambda](\omega) & :=\sum_{\varphi \in \mathcal{N}(\operatorname{supp}(\omega))} e^{-\lambda} \cdot \frac{\lambda^{\|\varphi\|}}{\varphi!} \cdot \prod_{x \in X} \omega(x)^{\varphi(x)}|\varphi\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} e^{-\lambda} \cdot \frac{\lambda^{K}}{K!} \cdot(\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)}|\varphi\rangle  \tag{3.52}\\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{pois}[\lambda](K) \cdot \operatorname{mn}[K](\omega)(\varphi)|\varphi\rangle \\
& =\operatorname{mn}[-](\omega) \gg=\operatorname{pois}[\lambda] .
\end{align*}
$$

Example 3.9.2. 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:

$$
\omega=\frac{1}{6}|a\rangle+\frac{1}{8}|b\rangle+\frac{3}{8}|c\rangle+\frac{1}{3}|d\rangle .
$$

The Poisson point process distribution $\operatorname{Pmn}[5](\omega) \in \mathcal{D}_{\infty}(\mathcal{N}(\{a, b, c, d\}))$ now gives for each multiset of diseases the arrival probability. We elaborate this
distribution for the first few sizes.

$$
\begin{aligned}
& \operatorname{Pmn}[5](\omega) \\
& =\operatorname{pois}[5](0) \cdot \operatorname{mn}[0](\omega)+\text { pois[5](1) } \operatorname{mn}[1](\omega)+\text { pois[5](2) } \cdot \operatorname{mn}[2](\omega)+\cdots \\
& =e^{-5}\left(\frac{5^{0}}{0!} \cdot \operatorname{mn}[0](\omega)+\frac{5^{1}}{1!} \cdot \operatorname{mn}[1](\omega)+\frac{5^{2}}{2!} \cdot \operatorname{mn}[2](\omega)+\cdots\right) \\
& \left.\left.\left.\left.\left.\left.\left.=e^{-5}\left(1|\mathbf{0}\rangle+\frac{5}{6}|1| a\right\rangle\right\rangle+\frac{5}{8}|1| b\right\rangle\right\rangle+\frac{15}{8}|1| c\right\rangle\right\rangle+\frac{5}{3}|1| d\right\rangle\right\rangle+ \\
& \left.\left.\left.\left.\left.\left.\left.\left.\quad \frac{25}{72}|2| a\right\rangle\right\rangle+\frac{25}{48}|1| a\right\rangle+1|b\rangle\right\rangle+\frac{25}{128}|2| b\right\rangle\right\rangle+\frac{25}{16}|1| a\right\rangle+1|c\rangle\right\rangle+ \\
& \left.\left.\left.\left.\left.\left.\left.\left.\frac{75}{64}|1| b\right\rangle+1|c\rangle\right\rangle+\frac{225}{128}|2| c\right\rangle\right\rangle+\frac{25}{18}|1| a\right\rangle+1|d\rangle\right\rangle+\frac{25}{24}|1| b\right\rangle+1|d\rangle\right\rangle+ \\
& \left.\left.\left.\left.\left.\quad \frac{25}{8}|1| c\right\rangle+1|d\rangle\right\rangle+\frac{25}{18}|2| d\right\rangle\right\rangle+\cdots\right) \\
& =0.00674|\mathbf{0}\rangle+0.00561|1| a\rangle\rangle+0.00421|1| b\rangle\rangle+0.0126|1| c\rangle\rangle+ \\
& \quad 0.0112|1| d\rangle\rangle+0.00234|2| a\rangle\rangle+0.00351|1| a\rangle+1|b\rangle\rangle+0.00132|2| b\rangle\rangle+ \\
& \quad 0.0105|1| a\rangle+1|c\rangle\rangle+0.0079|1| b\rangle+1|c\rangle\rangle+0.0118|2| c\rangle\rangle+ \\
& \quad 0.00936|1| a\rangle+1|d\rangle\rangle+0.00702|1| b\rangle+1|d\rangle\rangle+0.0211|1| c\rangle+1|d\rangle\rangle+ \\
& \\
& \quad 0.00936|2| d\rangle\rangle+\cdots
\end{aligned}
$$

The next result collects some basic properties.
Lemma 3.9.3. Consider the Poisson-iid and Poisson-multinomial maps Piid $[\lambda]: \mathcal{D}(X) \rightarrow$ $\mathcal{D}_{\infty}(\mathcal{L}(X))$ and $\operatorname{Pmn}[\lambda]: \mathcal{D}(X) \rightarrow \mathcal{D}_{\infty}(\mathcal{N}(X))$ introduced in Definition 3.9.1 satisfy the following properties
1 The are both natural in $X$;
2 Using the two size functions size : $\mathcal{L}(X) \rightarrow \mathbb{N}$ and size: $\mathcal{N}(X) \rightarrow \mathbb{N}$ satisfy:

$$
\mathcal{D}(\operatorname{size})(\operatorname{Piid}[\lambda](\omega))=\operatorname{pois}[\lambda]=\mathcal{D}(\operatorname{size})(\operatorname{Pmn}[\lambda](\omega))
$$

Equivalently, as string diagrams:


3 Pushforward with uniform projection and frequentist learning yields the original distribution:

$$
\text { unpr } \gg=\operatorname{Piid}[\lambda](\omega)=\omega=\operatorname{Flrn} \gg=\operatorname{Pmn}[\lambda](\omega) .
$$

$4 \operatorname{intsy}(\operatorname{Pmn}[\lambda](\omega))=\lambda \cdot \omega$, where intsy is the intensity natural transformation from 3.50.

Proof. 1 We do the multinomial case and use that multinomials $m n[K]: \mathcal{D}(X) \rightarrow$ $\mathcal{D}(\mathcal{N}[K](X))$ are natural in $X$. Thus, for a function $f: X \rightarrow Y$,

$$
\begin{aligned}
& (\operatorname{Pmn}[\lambda] \circ \mathcal{D}(f))(\omega) \\
& =\sum_{K \in \mathbb{N}} \sum_{\psi \in \mathcal{N}[K](Y)} \operatorname{pois}[\lambda](K) \cdot m n[K](\mathcal{D}(f)(\omega))(\psi)|\psi\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\psi \in \mathcal{N}[K](Y)} \operatorname{pois}[\lambda](K) \cdot \mathcal{D N}(f)(\operatorname{mn}[K](\omega))(\psi)|\psi\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\psi \in \mathcal{N}[K](Y)} \operatorname{pois}[\lambda](K) \cdot\left(\sum_{\varphi \in \mathcal{N}(f)^{-1}(\psi)} \operatorname{mn}[K](\omega)(\varphi)\right)|\psi\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{pois}[\lambda](K) \cdot \operatorname{mn}[K](\omega)(\varphi)|\mathcal{N}(f)(\varphi)\rangle \\
& =\left(\mathcal{D}_{\infty} \mathcal{N}(f) \circ \operatorname{Pmn}[\lambda]\right)(\omega) .
\end{aligned}
$$

2 We now do the iid case:

$$
\begin{aligned}
\mathcal{D}(\operatorname{size})(\operatorname{Piid}[\lambda](\omega)) & =\sum_{K \in \mathbb{N}} \sum_{\vec{x} \in X^{K}} \operatorname{pois}[\lambda](K) \cdot \operatorname{iid}[K](\omega)(\vec{x})|\operatorname{size}(\vec{x})\rangle \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot\left(\sum_{\vec{x} \in X^{K}} \operatorname{iid}[K](\omega)(\vec{x})\right)|K\rangle \\
& =\operatorname{pois}[\lambda] .
\end{aligned}
$$

3 For the multinomial case we use the equation Flrn $\gg m n[K](\omega)=\omega$ from Theorem 3.3.3 For $x \in X$,

$$
\begin{aligned}
& (\text { Flrn 》= Pmn }[\lambda](\omega))(x) \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{pois}[\lambda](K) \cdot \operatorname{mn}[K](\omega)(\varphi) \cdot \operatorname{Flrn}(\varphi)(x) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \operatorname{Flrn}(\varphi)(x) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot(\operatorname{Flrn} \gg \operatorname{mn}[K](\omega))(x) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot \omega(x)=\omega(x) .
\end{aligned}
$$

For the iid case we similarly use unpr $\gg \operatorname{iid}[K](\omega)=\omega$, see Exercise 2.4.7

4 For $x \in X$ we get:

$$
\begin{aligned}
\operatorname{intsy}(\operatorname{Pmn}[\lambda](\omega))(x) & =\operatorname{flat}(\operatorname{Pmn}[\lambda](\omega))(x) \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{pois}[\lambda](K) \cdot \operatorname{mn}[K](\omega)(\varphi) \cdot \varphi(x) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \varphi(x) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot K \cdot \omega(x) \quad \text { by Lemma3.3.2 } \\
& =\omega(x) \cdot \sum_{K \geq 1} e^{-\lambda} \cdot \frac{\lambda^{K}}{(K-1)!} \\
& =\lambda \cdot \omega(x) \cdot \sum_{K \geq 0} e^{-\lambda} \cdot \frac{\lambda^{K}}{K!} \\
& =\lambda \cdot \omega(x) .
\end{aligned}
$$

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 2.7.5 and Exercise 3.3.9 (or Theorem 3.4.4, using marginalisation).

Proposition 3.9.4. Poisson-iid and Poisson-multinomial distributions are closed under convolution, as expressed by the following two diagrams.


Proof. We do the multinomial case and use closure of multinomials under
convolution and reason as in the proof of Proposition 2.7.6.

$$
\begin{aligned}
& \mathcal{D}(+)\left(\operatorname{Pmn}\left[\lambda_{1}\right](\omega) \otimes \operatorname{Pmn}\left[\lambda_{2}\right](\omega)\right) \\
&= \sum_{K_{1}, K_{2} \in \mathbb{N}} \sum_{\varphi_{1} \in \mathcal{N}\left[K_{1}\right](X)} \sum_{\varphi_{2} \in \mathcal{N}\left[K_{2}\right](X)} e^{-\lambda_{1}} \cdot \frac{\lambda_{1}^{K_{1}}}{K_{1}!} \cdot e^{-\lambda_{2}} \cdot \frac{\lambda_{2}^{K_{2}}}{K_{2}!} \\
&= \sum_{L \in \mathbb{N}} \sum_{K \leq L} e^{\left.-\left(\lambda_{1}+\lambda_{2}\right)\right](\omega)\left(\varphi_{1}\right)} \cdot \operatorname{mn} \cdot \frac{\lambda_{1}^{K}}{K!} \cdot \frac{\left.\lambda_{2}^{L-K}\right](\omega)\left(\varphi_{2}\right)\left|\varphi_{1}+\varphi_{2}\right\rangle}{(L-K)!} \\
& \quad \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[L-K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \operatorname{mn}[L-K](\omega)(\psi)|\varphi+\psi\rangle \\
&= \sum_{L \in \mathbb{N}} \frac{e^{-\left(\lambda_{1}+\lambda_{2}\right)}}{L!} \cdot\left(\sum_{K \leq L}\binom{L}{K} \cdot \lambda_{1}^{K} \cdot \lambda_{2}^{L-K}\right) \cdot \mathcal{D}(+)(\operatorname{mn}[K](\omega) \otimes \operatorname{mn}[L-K](\omega)) \\
&= \sum_{L \in \mathbb{N}} \frac{e^{-\left(\lambda_{1}+\lambda_{2}\right)}}{L!} \cdot\left(\lambda_{1}+\lambda_{2}\right)^{L} \cdot \operatorname{mn}[L](\omega) \\
&= P m n\left[\lambda_{1}+\lambda_{2}\right](\omega) .
\end{aligned}
$$

For the next result we shall now use the Poisson-iid / multinomial and the iid / multinomial maps as binary functions:

$$
\begin{array}{ll}
\mathbb{R}_{>0} \times \mathcal{D}(X) \xrightarrow{\text { Piid }[-](-)} \mathcal{D}_{\infty}(\mathcal{L}(X)) & \mathbb{N} \times \mathcal{D}(X) \xrightarrow{\text { iid[-](-) }} \mathcal{D}_{\infty}(\mathcal{L}(X)) \\
\mathbb{R}_{>0} \times \mathcal{D}(X) \xrightarrow{\text { Pmn[-](-) }} \mathcal{D}_{\infty}(\mathcal{N}(X)) & \mathbb{N} \times \mathcal{D}(X) \xrightarrow{\text { mn[-](-) }} \mathcal{D}_{\infty}(\mathcal{N}(X))
\end{array}
$$

Theorem 3.9.5. 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 3.9.3 (2).

Proof. For $\lambda \in \mathbb{R}_{>0}$ and $\omega \in \mathcal{D}(X)$,

$$
\begin{aligned}
\langle\text { size }, \text { id }\rangle \gg \operatorname{Piid}[\lambda](\omega) & =\sum_{K \in \mathbb{N}} \sum_{\vec{x} \in X^{K}} \operatorname{pois}[\lambda](K) \cdot \operatorname{iid}[K](\omega)(\vec{x})|K, \vec{x}\rangle \\
& =\langle\operatorname{id}, \operatorname{iid}[-](\omega)\rangle \gg=\operatorname{pois}[\lambda] . \\
\langle\text { size }, \text { id }\rangle \gg \operatorname{Pmn}[\lambda](\omega) & =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{pois}[\lambda](K) \cdot \operatorname{mn}[K](\omega)(\varphi)|K, \varphi\rangle \\
& =\langle i d, \operatorname{mn}[-](\omega)\rangle \gg=\operatorname{pois}[\lambda] .
\end{aligned}
$$

Recall from Theorem 3.3.1 (2) the diagrammatic relationship between iid and multinomial via accumulation and arrangement. This also works for the Poisson versions.

Theorem 3.9.6. Poisson-iid and Poisson-multinomial channels are related as follows.


Proof. We reason essentially as in the proof of Theorem 3.3.1

$$
\begin{aligned}
& \langle\text { id, arr〉>>Pmn[ Pm]( } \omega \text { ) } \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in X^{K}} \operatorname{arr}(\varphi)(\vec{x}) \cdot \operatorname{Pmn}[\lambda](\omega)(\vec{x})|\varphi, \vec{x}\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{1}{(\varphi)} \cdot \operatorname{pois}[\lambda](K) \cdot \operatorname{mn}[K](\omega)(\varphi)|\varphi, \vec{x}\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in \operatorname{accc}^{-1}(\varphi)} \frac{1}{(\varphi)} \cdot \operatorname{pois}[\lambda](K) \cdot(\varphi) \cdot \prod_{y \in X} \omega(y)^{\varphi(y)}|\varphi, \vec{x}\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in \operatorname{acc} c^{-1}(\varphi)} \operatorname{pois}[\lambda](K) \cdot \prod_{1 \leq i \leq K} \omega\left(x_{i}\right)|\varphi, \vec{x}\rangle \\
& =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\vec{x} \in X^{K}}^{\langle\operatorname{acc}\rangle(\vec{x})(\varphi) \cdot \operatorname{pois}[\lambda](K) \cdot \operatorname{iid}[K](\omega)(\vec{x})|\varphi, \vec{x}\rangle} \\
& =\langle\operatorname{acc}, i d\rangle \gg \operatorname{Piid}[\lambda](\omega) .
\end{aligned}
$$

We finish 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 3.9.7. Let $X$ be a finite set with a distribution $\Omega \in \mathcal{D}_{\infty}(\mathcal{N}(X))$. We
call this $\Omega$ 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) $\Omega(n|x\rangle)=\frac{\Omega(1|x\rangle)^{n}}{n!\cdot \Omega(\mathbf{0})^{n-1}}$, for each $x \in X$ and $n>1$;
(d) $\Omega(\varphi+\psi)=\frac{\Omega(\varphi) \cdot \Omega(\psi)}{\Omega(\mathbf{0})}$, for multisets $\varphi, \psi \in \mathcal{N}(X)$ with disjoint support: $\operatorname{supp}(\varphi) \cap \operatorname{supp}(\psi)=\emptyset$.

Theorem 3.9.8. Let $X$ be a finite set.
1 For a distribution $\omega \in \mathcal{D}(X)$ and a rate $\lambda>0$, the Poisson-multinomial $\operatorname{Pmn}[\lambda](\omega) \in \mathcal{D}_{\infty}(\mathcal{N}(X))$ is a counting distribution of multisets: it satisfies requirements (a) - (d) in Definition 3.9.7
2 If $\Omega \in \mathcal{D}_{\infty}(\mathcal{N}(X))$ is a counting distribution of multisets, then it is of the form $\Omega=\operatorname{Pmn}[\lambda](\omega)$ 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 $\Omega:=\operatorname{Pmn}[\lambda](\omega) \in$ $\mathcal{D}_{\infty}(\mathcal{N}(X))$. We check the four requirements in Definition 3.9.7
(a) The empty multiset $\mathbf{0} \in \mathcal{N}(X)$ satisfies:

$$
\Omega(\mathbf{0})=\operatorname{Pmn}[\lambda](\omega)(\mathbf{0}) \stackrel{\sqrt{3.52}}{=} e^{\lambda} \cdot \frac{\lambda^{0}}{0!} \cdot \prod_{x \in X} \omega(x)^{0}=e^{-\lambda}
$$

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:

$$
\Omega(1|x\rangle)=\operatorname{Pmn}[\lambda](\omega)(1|x\rangle) \stackrel{\boxed{3.52}}{=} e^{\lambda} \cdot \frac{\lambda^{1}}{1!} \cdot \omega(x)^{1}=e^{-\lambda} \cdot \lambda \cdot \omega(x) .
$$

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 $n>1$,

$$
\begin{aligned}
\Omega(n|x\rangle)=\operatorname{Pmn}[\lambda](\omega)(n|x\rangle) & \stackrel{\sqrt{3.52}}{-} e^{-\lambda} \cdot \frac{\lambda^{n}}{n!} \cdot \omega(x)^{n} \\
& =e^{-\lambda} \cdot \frac{\Omega(1|x\rangle)^{n}}{n!\cdot\left(e^{-\lambda}\right)^{n}}=\frac{\Omega(1|x\rangle)^{n}}{n!\cdot \Omega(\mathbf{0})^{n-1}} .
\end{aligned}
$$

(d) Let $\varphi \in \mathcal{N}[K](X)$ and $\psi \in \mathcal{N}[L](X)$ have disjoint support. By Exercise 3.3.3.

$$
\begin{aligned}
\Omega(\varphi+\psi) & =\operatorname{pois}[\lambda](K+L) \cdot \operatorname{mn}[K+L](\varphi+\psi) \\
& =e^{-\lambda} \cdot \frac{\lambda^{K+L}}{(K+L)!} \cdot\binom{K+L}{K} \cdot \operatorname{mn}[K](\omega)(\varphi) \cdot \operatorname{mn}[L](\omega)(\psi) \\
& =\frac{1}{e^{-\lambda}} \cdot e^{-\lambda} \cdot \frac{\lambda^{K}}{K!} \cdot \operatorname{mn}[K](\omega)(\varphi) \cdot e^{-\lambda} \cdot \frac{\lambda^{L}}{L!} \cdot \operatorname{mn}[L](\omega)(\psi) \\
& =\frac{\operatorname{Pmn}[\lambda](\omega)(\varphi) \cdot \operatorname{Pmn}[\lambda](\omega)(\psi)}{e^{-\lambda}}=\frac{\Omega(\varphi) \cdot \Omega(\psi)}{\Omega(\mathbf{0})}
\end{aligned}
$$

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):

$$
\sum_{x \in X} \omega(x)=\sum_{x \in X} \frac{\Omega(1|x\rangle)}{-\Omega(\mathbf{0}) \cdot \ln (\Omega(\mathbf{0}))}=\frac{\sum_{x \in X} \Omega(1|x\rangle)}{-\Omega(\mathbf{0}) \cdot \ln (\Omega(\mathbf{0}))} \stackrel{\text { b }}{=} 1 .
$$

Now, for a multiset $\varphi \in \mathcal{N}(X)$,

$$
\begin{aligned}
\Omega(\varphi)=\Omega\left(\sum_{x \in X} \varphi(x)|x\rangle\right) & \stackrel{\text { d }}{=} \Omega(\mathbf{0}) \cdot \prod_{x \in X} \frac{\Omega(\varphi(x)|x\rangle)}{\Omega(\mathbf{0})} \\
& \stackrel{\text { d }}{=} e^{-\lambda} \cdot \prod_{x \in X} \frac{\Omega(1|x\rangle)^{\varphi(x)}}{\varphi(x)!\cdot \Omega(\mathbf{0})^{\varphi(x)}} \\
& =e^{-\lambda} \cdot \frac{1}{\prod_{x \in X} \varphi(x)!} \cdot \prod_{x \in X} \lambda^{\varphi(x)} \cdot \omega(x)^{\varphi(x)} \\
& =e^{-\lambda} \cdot \frac{1}{\varphi!} \cdot \lambda^{\sum_{x \in X} \varphi(x)} \cdot \prod_{x \in X} \omega(x)^{\varphi(x)} \\
& \stackrel{\text { B.52] }}{=} \operatorname{Pmn}[\lambda](\omega)(\varphi) .
\end{aligned}
$$

## Exercises

3.9.1 Check that the point process map $\operatorname{Pmn}[\lambda]: \mathcal{D} \Rightarrow \mathcal{D}_{\infty} \mathcal{N}$ is not a map of monads, e.g. because units are not preserved.
3.9.2 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) $\sum_{x \in X} \Omega([x])=-\Omega([]) \cdot \ln (\Omega([]))$;
(c) $\Omega\left(\ell_{1}+\ell_{2}\right)=\frac{\Omega\left(\ell_{1}\right) \cdot \Omega\left(\ell_{2}\right)}{\Omega([]) \cdot\binom{\left\|\ell_{1}\right\| l+\left\|\ell_{2}\right\|}{\left\|1_{1}\right\|}}$.

Show that these three points characterise Poisson-iid distributions.

## Observables and validity

So far we have seen (discrete probability) distributions as formal convex sums $\sum_{i} r_{i}\left|x_{i}\right\rangle$ in $\mathcal{D}(X)$ and (probabilistic) channels $X \rightsquigarrow Y$, as functions $X \rightarrow \mathcal{D}(Y)$, 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 | $X \rightarrow \mathbb{R}$ |
| factor / potential function | $X \rightarrow \mathbb{R}_{\geq 0}$ |
| (fuzzy) predicate / (soft / uncertain) evidence | $X \rightarrow[0,1]$ |
| sharp predicate / event | $X \rightarrow\{0,1\}$ |

We shall use the term of 'observable' as generic expression for all the entries in this table. A function $X \rightarrow \mathbb{R}$ is thus the most general type of observable, and a sharp predicate $X \rightarrow\{0,1\}$ is the most specific one. Predicates are the most appropriate observable for probabilistic logical reasoning. Often attention is restricted to subsets $U \subseteq X$ as predicates (or events [158]), but here, in this book, the fuzzy versions $X \rightarrow[0,1]$ 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 $=\ll$, 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 $\omega \vDash p$, namely the validity of an observable $p$ in a state $\omega$. It is the expected value of $p$ in $\omega$. 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 4.2

In the previous chapter we have seen that a state $\omega$ on the domain $X$ of a channel $c: X \leadsto Y$ can be transformed into a state $c \gg=\omega$ on the channel's codomain $Y$. Analogous to such state transformation $\gg$ there is also observable transformation $=\ll$, acting in the opposite direction: for an observable $q$ on the codomain $Y$ of a channel $c: X \leadsto Y$, there is a transformed observable $c=\ll q$ on the domain $X$. When $=\ll$ 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 [78]. At the end of this chapter, in Section 4.5, 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 'rational' distributions of the form $\operatorname{Flrn}(\varphi)$, for natural multisets $\varphi$, 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:

- $\operatorname{Obs}(X):=\mathbb{R}^{X}$ for the set of observables on a set $X$;
- $\operatorname{Fact}(X):=\left(\mathbb{R}_{\geq 0}\right)^{X}$ for the set of factors on $X$;
- $\operatorname{Pred}(X):=[0,1]^{X}$ for the set of predicates on $X$;
- $\operatorname{SPred}(X):=\{0,1\}^{X}$ for the set of sharp predicates (events) on $X$.

There are inclusions:

$$
\operatorname{SPred}(X) \subseteq \operatorname{Pred}(X) \subseteq \operatorname{Fact}(X) \subseteq \operatorname{Obs}(X) .
$$

The first set $\operatorname{SPred}(X)=\{0,1\}^{X}$ 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 $\mathbf{1}_{U}: X \rightarrow\{0,1\}$ 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 function $\mathbf{1}_{U}: X \rightarrow[0,1]$ is the (sharp) predicate associated with the subset $U \subseteq X$. The mapping $U \mapsto \mathbf{1}_{U}$ forms the isomorphism $\mathcal{P}(X) \xlongequal{\cong}$ $\{0,1\}^{X}$ that we just mentioned.

2 We use special notation for two extreme cases $U=X$ and $U=\emptyset$, giving the truth predicate 1: $X \rightarrow[0,1]$ and the falsity predicate $\mathbf{0}: X \rightarrow[0,1]$ on $X$. Explicitly:

$$
\mathbf{1}:=\mathbf{1}_{X} \quad \text { and } \mathbf{0}:=\mathbf{1}_{0} \quad \text { so that } \quad \mathbf{1}(x)=1 \text { and } \mathbf{0}(x)=0,
$$

for all $x \in X$.
3 For a singleton subset $\{x\}$ we simply write $\mathbf{1}_{x}$ for $\mathbf{1}_{\{x\}}$. Such functions $\mathbf{1}_{x}: X \rightarrow$ $[0,1]$ are also called point predicates, where the element $x \in X$ is seen as a point.

Recall that a state of the form $\operatorname{unit}(x)=1|x\rangle$ is called a point state or a point distribution.

4 There is a (sharp) equality predicate $E q: X \times X \rightarrow[0,1]$ defined in the obvious way as:

$$
E q\left(x, x^{\prime}\right):= \begin{cases}1 & \text { if } x=x^{\prime} \\ 0 & \text { otherwise }\end{cases}
$$

5 If the set $X$ can be mapped into $\mathbb{R}$ in an obvious manner, we write this inclusion function as incl ${ }_{X}: X \hookrightarrow \mathbb{R}$ 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 $p: X \rightarrow \mathbb{R}$ on $X$.
1 For a distribution / state $\omega=\sum_{i} r_{i}\left|x_{i}\right\rangle$ on $X$ we define its validity $\omega \vDash p$ as:

$$
\begin{equation*}
\omega \vDash p:=\sum_{i} r_{i} \cdot p\left(x_{i}\right)=\sum_{x \in X} \omega(x) \cdot p(x) \tag{4.2}
\end{equation*}
$$

2 If $X$ is a finite set, with size $|X| \in \mathbb{N}$, one can define the average $\operatorname{avg}(p)$ of the observable $p$ as its validity in the uniform state unif ${ }_{X}$ on $X$, i.e.,

$$
\operatorname{avg}(p):=u n i f_{X} \vDash p=\sum_{x \in X} \frac{p(x)}{|X|}
$$

Notice that validity $\omega \vDash p$ 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 as $E[p]$, where the state $\omega$ is left implicit. The notation $\omega \vDash p$ makes the state explicit, which is important in situations where we change the state, for instance via state transformation.
The validity $\omega \vDash p$ 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 $\omega=\sum_{i} r_{i}\left|x_{i}\right\rangle$ add up to one means that the validity $\omega \vDash \mathbf{1}$ of truth is one. Notice that for a point predicate one has $\omega \vDash \mathbf{1}_{x}=\omega(x)$ and similarly, for a point state, $1|x\rangle \vDash p=p(x)$.

For a fixed state $\omega \in \mathcal{D}(X)$ we can view $\omega \vDash(-)$ as a function $\operatorname{Pred}(X) \rightarrow$ $[0,1]$ that assigns a likelihood to a belief function (predicate). This is at the heart of the Bayesian interpretation of probability, see Section 6.8 for more details.

As an aside, we typically do not write brackets in equations like $(\omega \vDash p)=$ $a$, but use the convention that $\vDash$ has higher precedence than $=$, so that the equation can simply be written as $\omega \vDash p=a$. Similarly, one can have validity $c »>\omega \vDash p$ in a transformed state, which should be read as $(c \gg \omega) \vDash p$.

Definition 4.1.3. In the presence of a map incl $_{X}: X \hookrightarrow \mathbb{R}$, one can define the mean mean $(\omega)$, also known as average, of a distribution $\omega$ on $X$ as the validity of incl $_{X}$, considered as an observable:

$$
\operatorname{mean}(\omega):=\omega \vDash \operatorname{incl}_{X}=\sum_{x \in X} \omega(x) \cdot x
$$

We will use the same definition of mean when $\omega$ is a multiset instead of a distribution.

The definition of mean may be used in situations when the set $X$ is a cone, equipped with scalar multiplication and addition. We shall do so occasionally.

## Example 4.1.4.

1 Let $\operatorname{flip}\left(\frac{3}{10}\right)=\frac{3}{10}|1\rangle+\frac{7}{10}|0\rangle$ 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 abservable $v:\{0,1\} \rightarrow \mathbb{R}$ with $v(0)=-50$ and $v(1)=100$. We get an anwer to the above question by computing the validity:

$$
\begin{aligned}
\operatorname{flip}\left(\frac{3}{10}\right) \vDash v & =\sum_{x \in\{0,1\}} \operatorname{flip}\left(\frac{3}{10}\right)(x) \cdot v(x) \\
& =\operatorname{flip}\left(\frac{3}{10}\right)(0) \cdot v(0)+\operatorname{flip}\left(\frac{3}{10}\right)(1) \cdot v(1) \\
& =\frac{7}{10} \cdot-50+\frac{3}{10} \cdot 100=-35+30=-5 .
\end{aligned}
$$

Hence it is wiser not to play.
2 We write pips for the set $\{1,2,3,4,5,6\}$, considered as a subset of $\mathbb{R}$, via the map incl : pips $\hookrightarrow \mathbb{R}$, which is used as an observable. As state we use the (uniform) fair dice dice $=$ unif $_{\text {pips }}=\frac{1}{6}|1\rangle+\frac{1}{6}|2\rangle+\frac{1}{6}|3\rangle+\frac{1}{6}|4\rangle+\frac{1}{6}|5\rangle+\frac{1}{6}|6\rangle$. The average is the validity dice $\vDash$ incl is $\frac{21}{6}=\frac{7}{2}$. It is the expected outcome for throwing a dice.
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 $e:$ pips $\rightarrow$ $[0,1]$ with $e(1)=e(3)=e(5)=0$ and $e(2)=e(4)=e(6)=1$. Then, as expected:

$$
\begin{aligned}
\operatorname{dice} \models e & =\sum_{x \in \operatorname{pips}} \operatorname{dice}(x) \cdot e(x) \\
& =\frac{1}{6} \cdot 0+\frac{1}{6} \cdot 1+\frac{1}{6} \cdot 0+\frac{1}{6} \cdot 1+\frac{1}{6} \cdot 0+\frac{1}{6} \cdot 1=\frac{1}{2} .
\end{aligned}
$$

We now consider a non-sharp 'evenish' predicate $p:$ pips $\rightarrow[0,1]$. It ex-
presses that the even 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:

$$
\text { dice } \begin{aligned}
\models p & =\frac{1}{6} \cdot \frac{1}{10}+\frac{1}{6} \cdot \frac{9}{10}+\frac{1}{6} \cdot \frac{8}{10}+\frac{1}{6} \cdot \frac{8}{10}+\frac{1}{6} \cdot \frac{2}{10}+\frac{1}{6} \cdot \frac{7}{10} \\
& =\frac{1+9+3+8+2+7}{60}=\frac{30}{60}=\frac{1}{2} .
\end{aligned}
$$

4 Recall the binomial distribution $b n[K](r)$ 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 3.3.2. But one can also compute directly:

$$
\begin{aligned}
\operatorname{mean}(b n[K](r)) & =\sum_{0 \leq i \leq K} b n[K](r)(i) \cdot i \\
& =\sum_{0 \leq i \leq K}\binom{K}{i} \cdot r^{i} \cdot(1-r)^{K-i} \cdot i \\
& =\sum_{1 \leq i \leq K} \frac{K!}{(i-1)!\cdot(K-i)!} \cdot r^{i} \cdot(1-r)^{K-i} \\
& =K \cdot r \cdot \sum_{0 \leq j \leq K-1} \frac{(K-1)!}{j!\cdot((K-1)-j)!} \cdot r^{j} \cdot(1-r)^{(K-1)-j} \\
& =K \cdot r \cdot \sum_{0 \leq j \leq K-1} b n[K-1](r)(j) \\
& =K \cdot r .
\end{aligned}
$$

In the two plots of binomial distributions bn[10]((%5Cfrac%7B1%7D%7B3%7D)) and bn[10] $\left(\frac{3}{4}\right)$ in Figure 2.2 (on page 93 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 4.4
We include another, extended example with factors defined on joint distributions.

Example 4.1.5. 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 $H_{H}=\sum_{i \in H} \frac{1}{60}|i\rangle$ on $H$.





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 .5 for details.

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 $\operatorname{unif}_{H} \otimes 1|45\rangle$ 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 $p: H \times H \rightarrow \mathbb{N}$, 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 com-
pute the expected arrival time, via the sum formula of Proposition 1.2.6 (1).

$$
\begin{aligned}
\operatorname{unif}_{H} \otimes 1|45\rangle \vDash p & =\sum_{(i, b) \in H \times H} \frac{p(i, b)}{60} \\
& =\sum_{1 \leq i \leq 45} \frac{45-i}{60} \\
& =\sum_{1 \leq i \leq 44} \frac{i}{60}=\frac{44 \cdot 45}{2 \cdot 60}=\frac{33}{2}=16.5 .
\end{aligned}
$$

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:

$$
v:=\frac{1}{12}|43\rangle+\frac{1}{6}|44\rangle+\frac{1}{2}|45\rangle+\frac{1}{6}|46\rangle+\frac{1}{12}|47\rangle .
$$

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 unif $f_{H} \otimes v \in \mathcal{D}(H \times H)$. What is the expected waiting time now, as given by the validity of $p$ now? It turns out be slightly higher than for the punctual bus. It is computed as:

$$
\begin{aligned}
& \text { unif }_{H} \otimes v \vDash p \\
& =\frac{1}{12}\left(\sum_{1 \leq i \leq 43} \frac{43-i}{60}\right)+\frac{1}{6}\left(\sum_{1 \leq i \leq 44} \frac{44-i}{60}\right)+\frac{1}{2}\left(\sum_{1 \leq i \leq 45} \frac{45-i}{60}\right) \\
& \quad \quad+\frac{1}{6}\left(\sum_{1 \leq i \leq 46} \frac{46-i}{60}\right)+\frac{1}{12}\left(\sum_{1 \leq i \leq 47} \frac{47-i}{60}\right) \\
& =\frac{42 \cdot 43}{12 \cdot 2 \cdot 60}+\frac{43 \cdot 44}{6 \cdot 2 \cdot 60}+\frac{44 \cdot 45}{2 \cdot 2 \cdot 60}+\frac{45 \cdot 46}{6 \cdot 2 \cdot 60}+\frac{46 \cdot 47}{12 \cdot 2 \cdot 60} \\
& =\frac{1981}{120}=16.508333 \cdots
\end{aligned}
$$

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 $u n i f_{H} \otimes 1|15\rangle \otimes 1|45\rangle$ on $H \times H \times H$.

The arrival time factor $q: H \times H \times H \rightarrow \mathbb{N}$ now is:

$$
q\left(i, b_{1}, b_{2}\right):= \begin{cases}b_{1}-i & \text { if } i \leq b_{1} \text { and }\left(b_{2}<i \text { or } b_{1} \leq b_{2}\right) \\ b_{2}-i & \text { if } i \leq b_{2} \text { and }\left(b_{1}<i \text { or } b_{2} \leq b_{1}\right) \\ 0 & \text { otherwise } .\end{cases}
$$

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 $\left(b_{2}<i\right)$ or arriving later than bus $1\left(b_{1} \leq b_{2}\right)$. The second clause is symmetric. We now get as expected waiting time:

$$
\begin{aligned}
\operatorname{unif}_{H} \otimes 1|15\rangle \otimes 1|45\rangle \vDash q & =\sum_{1 \leq i \leq 15} \frac{15-i}{60}+\sum_{16 \leq i \leq 45} \frac{45-i}{60} \\
& =\frac{14 \cdot 15+29 \cdot 30}{2 \cdot 60}=9 .
\end{aligned}
$$

Clearly, with two buses the waiting is less.
4 What if the two buses are both not punctual, with distributions $v_{1}$ and $v_{2}$ around 15 and 45, see Figure 4.1 at the bottom, and:

$$
\begin{aligned}
& v_{1}:=\frac{1}{12}|13\rangle+\frac{1}{6}|14\rangle+\frac{1}{2}|15\rangle+\frac{1}{6}|16\rangle+\frac{1}{12}|17\rangle \\
& v_{2}:=\frac{1}{12}|43\rangle+\frac{1}{6}|44\rangle+\frac{1}{2}|45\rangle+\frac{1}{6}|46\rangle+\frac{1}{12}|47\rangle .
\end{aligned}
$$

Again there is a (slightly) longer expected waiting time than with two punctual buses:

$$
u n i f_{H} \otimes v_{1} \otimes v_{2} \vDash q=\frac{361}{40}=9.025 .
$$

This concludes the bus illustration. Exercise 4.1.4 contains two follow-up questions.

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 $M$. This is used below where $M$ is the additive monoid $\mathbb{N}$ of natural numbers. The result can be generalised to any additive submonoid of the reals.

Lemma 4.1.6. The mean function, for distributions on natural numbers, is a homomorphism of monoids of the form:

$$
(\mathcal{D}(\mathbb{N}),+, 0) \xrightarrow{\text { mean }}\left(\mathbb{R}_{\geq 0},+, 0\right) .
$$

Proof. Preservation of the zero element is easy, since:

$$
\operatorname{mean}(1|0\rangle)=1 \cdot 0=0 .
$$

Next, for $\omega, \rho \in \mathcal{D}(\mathbb{N})$,

$$
\begin{aligned}
\operatorname{mean}(\omega+\rho) & \stackrel{[2.44}{=} \sum_{n \in \mathbb{N}} \mathcal{D}(+)(\omega \otimes \rho)(n) \cdot n \\
& =\sum_{m, k \in \mathbb{N}}(\omega \otimes \rho)(m, k) \cdot(m+k) \\
& =\sum_{m, k \in \mathbb{N}} \omega(m) \cdot \rho(k) \cdot(m+k) \\
& =\left(\sum_{m, k \in \mathbb{N}} \omega(m) \cdot \rho(k) \cdot m\right)+\left(\sum_{m, k \in \mathbb{N}} \omega(m) \cdot \rho(k) \cdot k\right) \\
& =\left(\sum_{m \in \mathbb{N}} \omega(m) \cdot\left(\sum_{k \in \mathbb{N}} \rho(k)\right) \cdot m\right)+\left(\sum_{k \in \mathbb{N}}\left(\sum_{m \in \mathbb{N}} \omega(m)\right) \cdot \rho(k) \cdot k\right) \\
& =\left(\sum_{m \in \mathbb{N}} \omega(m) \cdot m\right)+\left(\sum_{k \in \mathbb{N}} \rho(k) \cdot k\right) \\
& =\operatorname{mean}(\omega)+\operatorname{mean}(\rho) .
\end{aligned}
$$

Remark 4.1.7. What is the difference between a multiset and a factor, and between a distribution and a predicate? A multiset on $X$ is a function $X \rightarrow \mathbb{R}_{\geq 0}$ with finite support, and thus a factor. In the same way a distribution on $X$ is a function $X \rightarrow[0,1]$ with finite support, and thus a predicate on $X$. Hence, there are inclusions:

$$
\mathcal{M}(X) \subseteq \operatorname{Fact}(X)=\left(\mathbb{R}_{\geq 0}\right)^{X} \quad \text { and } \quad \mathcal{D}(X) \subseteq \operatorname{Pred}(X)=[0,1]^{X}
$$

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, distributions are convex sets, whereas predicates are effect modules carrying a commutative monoid, see the next section. Keeping states and predicates apart is a matter of mathematicaly hygienf ${ }^{11}$. We have already seen state transformation $\gg=$ along a channel; it preserves the convex structure. Later on in this chapter we shall also see predicate transformation $=\ll$ along a channel, in opposite direction; this operation preserves the effect module structure on predicates.
Apart from these mathematical differences, states and predicates play entirely different roles and are understood in different ways: states play an ono-

[^5]tologoical role and describe 'states of affairs'; predicates are epistemological in nature, and describe evidence (what we know).

Whenever we do make use of the above inclusions, we shall make this usage explicit.

### 4.1.1 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 $E[Y]$. 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. For such a pair it makes sense to talk about expectation, namely as their validity.

Consider a space Pet $=\{d, c, r, h\}$ for $d=\operatorname{dog}, c=\mathrm{cat}, r=$ rabbit and $h=$ hamster. We look at the cost (e.g. of food) of a pet per month, via an observable $q:$ Pet $\rightarrow \mathbb{R}$ 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 $\omega$ is on the left and the costs per pet is in the middle. The plot on the right describes $\frac{7}{20}|10\rangle+\frac{13}{20}|50\rangle$, which is the distribution of pet costs. It is obtained via the functoriality of $\mathcal{D}$, as image distribution $\mathcal{D}(q)(\omega) \in \mathcal{D}(\mathbb{R})$. In more traditional notation it is described as:

$$
P[q=10]=\frac{7}{20} \quad P[q=50]=\frac{13}{20} .
$$

Sometimes, a random variable is described as such an (image) distribution on the reals, with the underlying distribution $\omega$ and observable $q$ left implicit. As this example illustrates, there may be much more structure around. We prefer to 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 $q \sim \operatorname{pois}[\lambda]$ with $\ldots$ ". This means that $q: \mathbb{N} \rightarrow$ $\mathbb{R}$ is an observable, on the underlying space $\mathbb{N}$ of the Poisson distribution. It thus describes a random variable as a pair of a state and an observable, with the same underlying space. In the literature one should be aware of what is
called in [113, §16] the "notational confusion between a random variable and its distribution".

## Definition 4.1.8.

1 A random variable on a sample space (set) $X$ consists of two parts:

- a distribution/state $\omega \in \mathcal{D}(X)$;
- an observable $R: X \rightarrow \mathbb{R}$.

2 The probability mass function $P[R=(-)]: \mathbb{R} \rightarrow[0,1]$ associated with the random variable $(\omega, R)$ is the image distribution on $\mathbb{R}$ given by:

$$
\begin{equation*}
P[R=(-)]:=\mathcal{D}(R)(\omega)=R \gg=\omega, \tag{4.3}
\end{equation*}
$$

where $R$ is understood as a deterministic channel in the last expression $R \gg=$ $\omega$, see Lemma 1.10.3 (4).
3 The expected value $E[R]$ of a random variable $(\omega, R)$ is the validity of $R$ in $\omega$, which can be expressed as mean of the image distribution:

$$
\begin{equation*}
\omega \vDash R=\operatorname{mean}(R \gg=\omega)=\operatorname{mean}(\mathcal{D}(R)(\omega)) . \tag{4.4}
\end{equation*}
$$

The image distribution in (4.3) can be described in several (equivalent) ways:

$$
\begin{aligned}
P[R=r]=\mathcal{D}(R)(\omega)(r)=(R \gg=\omega)(r) & =\omega \gg=R \vDash \mathbf{1}_{r} \\
& =\sum_{x, R(x)=r} \omega(x) \\
& =\sum_{x \in R^{-1}(r)} \omega(x) \\
& =\omega \vDash \mathbf{1}_{R^{-1}(r) .} .
\end{aligned}
$$

In the second item of the above definition, Equation (4.4) holds since:

$$
\begin{aligned}
\operatorname{mean}(R \gg=\omega) & =\sum_{r \in \mathbb{R}}(R \gg=\omega)(r) \cdot r \quad \text { see Definition 4.1.3 } \\
& =\sum_{r \in \mathbb{R}} \mathcal{D}(R)(\omega)(r) \cdot r \\
& =\sum_{r \in \mathbb{R}}\left(\sum_{x \in R^{-1}(r)} \omega(x)\right) \cdot r \\
& =\sum_{x \in X} \omega(x) \cdot R(x) \\
& =\omega \models R .
\end{aligned}
$$

Example 4.1.9. We consider the expected value for the sum of two dices. In this situation we have an observable $S:$ pips $\times$ pips $\rightarrow \mathbb{R}$, 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 state dice $\otimes$ dice $\in \mathcal{D}($ pips $\times$ pips $)$. Recall, dice $=$ unif $=\sum_{i \in p i p s} \frac{1}{6}|i\rangle$ is the uniform distribution unif on pips. First, the distribution for the sum of the pips is the image distribution:

$$
\begin{align*}
S \gg \text { dice } \otimes \text { dice }= & \mathcal{D}(+)(\text { dice } \otimes \text { dice }) \\
= & \sum_{2 \leq n \leq 12}\left(\sum_{i, j, i+j=n}(\text { dice } \otimes \text { dice })(i, j)\right)|n\rangle  \tag{4.5}\\
= & \frac{1}{36}|2\rangle+\frac{1}{18}|3\rangle+\frac{1}{12}|4\rangle+\frac{1}{9}|5\rangle+\frac{5}{36}|6\rangle+\frac{1}{6}|7\rangle \\
& \quad+\frac{5}{36}|8\rangle+\frac{1}{9}|9\rangle+\frac{1}{12}|10\rangle+\frac{1}{18}|11\rangle+\frac{1}{36}|12\rangle .
\end{align*}
$$

The expected value of the random variable (dice $\otimes$ dice, $S$ ) is, according to Definition 4.1.8 (3),

$$
\begin{aligned}
\operatorname{mean}(S \gg \text { dice } \otimes \text { dice }) & =\text { dice } \otimes \text { dice } \vDash S \\
& =\sum_{i, j \in p i p s}(\text { dice } \otimes \text { dice })(i, j) \cdot S(i, j) \\
& =\sum_{i, j \in p i p s}^{6} \cdot \frac{1}{6} \cdot(i+j) \\
& =\frac{\sum_{i, j \in p i p s} i+j}{36}=\frac{252}{36}=7 .
\end{aligned}
$$

There is a more abstract way to look at this example, using Lemma 4.1.6 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 4.1.6, we also get:

$$
\operatorname{mean}(\text { dice }+ \text { dice })=\operatorname{mean}(\text { dice })+\operatorname{mean}(\text { dice })=\frac{7}{2}+\frac{7}{2}=7
$$

We conclude with a result for which it is relevant to know in which state we are evaluating an observable.

Lemma 4.1.10. Let $M=(M,+, 0)$ be commutative monoid, so that the set of distributions $\mathcal{D}(M)$ is also a commutative monoid by Proposition 2.7.2. Let observable $q: M \rightarrow \mathbb{R}$ be a map of (additive) monoids. The function "validity of $q$ " is then also a map of monoids in:

$$
\mathcal{D}(M) \xrightarrow{(-) \equiv q} \mathbb{R}
$$

Explicitly, this means, for $\omega, \rho \in \mathcal{D}(M)$,

$$
(\omega+\rho) \vDash q=(\omega \vDash q)+(\rho \vDash q) \quad \text { and } \quad 1|0\rangle \vDash q=0 .
$$

This result involves three different random variables, namely:

$$
(\omega+\rho, q) \quad(\omega, q) \quad(\rho, q)
$$

Proof. By unravelling the definitions:

$$
\begin{aligned}
(\omega+\rho) \vDash q & =\sum_{x \in M} \mathcal{D}(+)(\omega \otimes \rho)(x) \cdot q(x) \\
& =\sum_{x \in M} \sum_{y, z \in M, y+z=x} \omega(y) \cdot \rho(z) \cdot q(x) \\
& =\sum_{y, z \in M} \omega(y) \cdot \rho(z) \cdot q(y+z) \\
& =\sum_{y, z \in M} \omega(y) \cdot \rho(z) \cdot(q(y)+q(z)) \\
& =\sum_{y \in M} \omega(y) \cdot\left(\sum_{z \in M} \rho(z)\right) \cdot q(y)+\sum_{z \in M}\left(\sum_{y \in M} \omega(y)\right) \cdot \rho(z) \cdot q(z) \\
& =\sum_{y \in M} \omega(y) \cdot q(y)+\sum_{z \in M} \rho(z) \cdot q(z) \\
& =(\omega \models q)+(\rho \vDash q) .
\end{aligned}
$$

Similarly, $1|0\rangle \vDash q=q(0)=0$.

## Exercises

4.1.1 Check that the average of the set $\boldsymbol{n} \boldsymbol{+ 1}=\{0,1, \ldots, n\}$, considered as random variable, is $\frac{n}{2}$.
4.1.2 In Example 4.1.9 we have seen that dice $\otimes$ dice $\vDash S=7$, for the observable $S:$ pips $\times$ pips $\rightarrow \mathbb{R}$ given by $S(x, y)=x+y$.
1 Now define $T:$ pips $^{3} \rightarrow \mathbb{R}$ by $T(x, y, z)=x+y+z$. Prove that dice $\otimes$ dice $\otimes$ dice $\vDash T=\frac{21}{2}$.
2 Can you generalise and show that summing on pips ${ }^{n}$ yields validity $\frac{7 n}{2}$ in state dice ${ }^{n}$ ?
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 days $:=\{1,2,3,4,5,6,7\}$. The set $\mathcal{N}[3]$ (days) of multisets of size three contains the possible combinations of the three birthdays. Define the sharp predicate $p: \mathcal{N}[3]($ days $) \rightarrow\{0,1\}$ by $p(\varphi)=1 \mathrm{iff}(\varphi) \leq 3$. Check that $p$ holds in those cases where at least two birthdays coincide - see also Exercise 1.7.9.
4 Show that the probability $\frac{19}{49}$ of at least two coinciding birthdays can also be obtained via validity, namely as:

$$
m n[3]\left(\text { unif }_{\text {days }}\right) \vDash p=\frac{19}{49} .
$$

4.1.4 Consider the situation with one bus (per hour) in Example 4.1.5.

1 Compute the expected waiting time when the bus arrives at minute 60 , that is, compute the validity $u n i f_{H} \otimes 1|60\rangle \vDash p$, for the waiting time factor $p$ from Example 4.1.5.
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 $_{H} \vDash p$ equals $\frac{3599}{360}$, which is almost 10 minutues.
Hint: Use Proposition 1.2.6 1) and (2).
4.1.5 Consider an arbitrary distribution $\omega \in \mathcal{D}(X)$.

1 Check that for a function $f: X \rightarrow Y$ and an observable $q \in O b s(Y)$,

$$
f \gg=\omega \vDash q=\omega \vDash q \circ f
$$

2 Observe that we get as special case, for an observable $p: X \rightarrow \mathbb{R}$,

$$
\operatorname{mean}(p \gg \omega)=p \gg=\omega \vDash \text { id }=\omega \vDash p,
$$

where the identity function id: $\mathbb{R} \rightarrow \mathbb{R}$ is used as observable.
4.1.6 Let $\omega, \rho \in \mathcal{D}(X)$ be given distributions. Prove the following two equations, in the style of Lemma4.1.6.
1 For observables $p, q: X \rightarrow \mathbb{R}$, and thus for random variables, $(\omega, p)$ and $(\rho, q)$, one has:

$$
\operatorname{mean}(\mathcal{D}(p)(\omega)+\mathcal{D}(q)(\rho))=(\omega \vDash p)+(\rho \vDash q)
$$

In the notation of Definition 2 we can also write the left-hand side as: $\operatorname{mean}(P[p=(-)]+P[q=(-)])$.
2 For channels $c, d: X \mapsto \mathbb{R}$,

$$
\operatorname{mean}((c \gg=\omega)+(d \gg \rho))=(\omega \vDash \text { mean } \circ c)+(\rho \vDash \text { mean } \circ d) .
$$

4.1.7 Let $\Omega \in \mathcal{D}(\mathcal{D}(X))$ be a distribution of distributions, with an observable $p: X \rightarrow \mathbb{R}$. We turn $p$ into a 'validity of $p$ ' observable $(-) \vDash p: \mathcal{D}(X) \rightarrow \mathbb{R}$ on $\mathcal{D}(X)$. Show that:

$$
\operatorname{flat}(\Omega) \vDash p=\Omega \vDash((-) \vDash p) .
$$

4.1.8 We have introduced the mean in Definition 4.1.3 for distributions whose space is a subset of the reals. By definition, these distributions have finite support. But the definition carries over to distributions with infinite support, as described in Definition 2.1.5 Apply this to the Poisson distribution pois $[\lambda] \in \mathcal{D}_{\infty}(\mathbb{N})$ and show that

$$
\operatorname{mean}(\operatorname{pois}[\lambda])=\lambda
$$

4.1.9 Consider the 'mean' operation from Definition 4.1.3 as a function mean: $\mathcal{D}(\mathbb{R}) \rightarrow \mathbb{R}$. 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 $p: X \rightarrow \mathbb{R}$ be an observable, giving a function $(-) \vDash p: \mathcal{D}(X) \rightarrow$ $\mathbb{R}$, sending $\omega \in \mathcal{D}(X)$ to $\omega \vDash p$ in $\mathbb{R}$. Show that the following diagram commutes.


4 Prove that the following diagram commutes.

where the $\operatorname{sum} \mathcal{D}(\mathbb{R}) \times \mathcal{D}(\mathbb{R}) \rightarrow \mathcal{D}(\mathbb{R})$ at the top is convolution sum of distributions from Definition 2.7.1.

From the first two items we can conclude that mean is an (EilenbergMoore) algebra of the distribution monad, see Subsection 1.11.4. The third item says that $(-) \vDash p: \mathcal{D}(X) \rightarrow \mathbb{R}$ is a homomorphism of algebras. And the fourth point tells that mean is a map of (commutative) monoids.
4.1.10 The following result is a discrete formulation of what is called Fubini's theorem. For the more familiar continuous version, see Theorem ??.

Let $\omega_{1} \in \mathcal{D}\left(X_{1}\right)$ and $\omega_{2} \in \mathcal{D}\left(X_{2}\right)$ be two distributions, with an observable (relation) $r: X_{1} \times X_{2} \rightarrow \mathbb{R}$. Prove that:

$$
\omega_{1} \vDash r_{1}=\omega_{1} \otimes \omega_{2} \vDash r=\omega_{2} \vDash r_{2},
$$

where $r_{1}: X_{1} \rightarrow \mathbb{R}$ and $r_{2}: X_{2} \rightarrow \mathbb{R}$ are the observables defined by:

$$
\begin{aligned}
& r_{1}\left(x_{1}\right):=\omega_{2} \vDash r\left(x_{1},-\right)=\sum_{x_{2}} \omega_{2}\left(x_{2}\right) \cdot r\left(x_{1}, x_{2}\right) \\
& r_{2}\left(x_{2}\right):=\omega_{1} \vDash r\left(-, x_{2}\right)=\sum_{x_{1}} \omega_{1}\left(x_{1}\right) \cdot r\left(x_{1}, x_{2}\right) .
\end{aligned}
$$

4.1.11 For a distribution $\omega \in \mathcal{D}(X)$ write $I(\omega): X \rightarrow \mathbb{R}_{\geq 0}$ for the information content or surprise of the distribution $\omega$, defined as factor:

$$
I(\omega)(x):= \begin{cases}-\ln (\omega(x)) & \text { if } \omega(x) \neq 0, \text { i.e. if } x \in \operatorname{supp}(\omega) \\ 0 & \text { otherwise }\end{cases}
$$

1 Check that Kullback-Leibler divergence, see Definition 2.8.1 can be described as validity: for $\omega, \rho \in \mathcal{D}(X)$ with $\operatorname{supp}(\omega) \subseteq \sup (\rho)$,

$$
D_{K L}(\omega, \rho)=\omega \vDash I(\rho)-I(\omega)
$$

2 The Shannon entropy $H(\omega)$ of $\omega$ and the cross entropy $H(\omega, \rho)$ of $\omega, \rho$, are defined as: validities:

$$
\begin{aligned}
H(\omega) & :=\omega \vDash I(\omega)=-\sum_{x} \omega(x) \cdot \ln (\omega(x)) \\
H(\omega, \rho) & :=\omega \vDash I(\rho)=-\sum_{x} \omega(x) \cdot \ln (\rho(x))
\end{aligned}
$$

Thus, Shannon entropy is 'expected surprise'. Show that:

$$
H(\omega, \rho)=H(\omega)+D_{K L}(\omega, \rho)
$$

4.1.12 Consider the entropy function $H$ from the previous exercise.

1 Show that $H(\omega)=0$ implies that $\omega \in \mathcal{D}(X)$ is a point state $\omega=$ $1|x\rangle$, for a unique element $x \in X$.
2 Show that for $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$,

$$
H(\sigma \otimes \tau)=H(\sigma)+H(\tau)
$$

3 Strengthen the previous item to non-entwinedness: for $\omega \in \mathcal{D}(X \times$ $Y$ ),

$$
H(\omega)=H(\omega[1,0])+H(\omega[0,1]) \Longleftrightarrow \omega=\omega[1,0] \otimes \omega[0,1]
$$

Hint: Use for finitely many numbers $r_{i}, s_{i} \in[0,1]$, where $\sum_{i} r_{i}=$ $1=\sum_{i} s_{i}$, that $\sum_{i} r_{i} \cdot \ln \left(r_{i}\right)=\sum_{i} r_{i} \cdot \ln \left(s_{i}\right)$ implies $r_{i}=s_{i}$ for each $i$.
4.1.13 Let $\omega \in \mathcal{D}(X)$ be an arbitrary distribution on a finite set $X$. Use Proposition 2.8.4 (1) to prove that uniform distributions have maximal Shannon entropy, that is, $H(\omega) \leq H$ (unif), where unif $\in \mathcal{D}(X)$ is the uniform distribution.

### 4.2 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 | $X \rightarrow \mathbb{R}$ | ordered vector spaces |
| factor | $X \rightarrow \mathbb{R}_{\geq 0}$ | ordered cones |
| predicate | $X \rightarrow[0,1]$ | effect modules |
| sharp predicate | $X \rightarrow\{0,1\}$ | join semilattices |

The least well-known structures in this table are effect modules. They will thus be described in greatest detail, in Subsection 4.2.3.

### 4.2.1 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 $X \rightarrow \mathbb{R}^{n}$ to the $n$ dimensional space of real numbers. Whenever needed, we shall use such maps as $n$-ary tuples $\left\langle p_{1}, \ldots, p_{n}\right\rangle: X \rightarrow \mathbb{R}^{n}$ of observables $p_{i}: X \rightarrow \mathbb{R}$, see also Section 1.3 .

Let us fix a set $X$, and consider the collection $\operatorname{Obs}(X)=\mathbb{R}^{X}$ of observables on $X$. What structure does it have?

- Given two observables $p, q \in \operatorname{Obs}(X)$, we can add them pointwise, giving $p+q \in \operatorname{Obs}(X)$ via $(p+q)(x)=p(x)+q(x)$.
- Given an observable $p \in \operatorname{Obs}(X)$ and a scalar $s \in \mathbb{R}$ we can form a 'rescaled' observable $s \cdot p \in \operatorname{Obs}(X)$ 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 $\mathbf{0}=\mathbf{1}_{\emptyset} \in \operatorname{Obs}(X)$ 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 $p: X \rightarrow \mathbb{R}$ 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 [4, 89, 138, 144].

The set of observables $\operatorname{Obs}(X)=\mathbb{R}^{X}$ 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).

### 4.2.2 Factors

We recall that a factor is a non-negative observables and that we write $\operatorname{Fact}(X)=$ $\left(\mathbb{R}_{\geq 0}\right)^{X}$ 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 $\operatorname{Fact}(X)$ looks like a vector space, except that there are no negatives. Using the order on observables introduced in the previous subsection, we can write:

$$
\operatorname{Fact}(X)=\{p \in \operatorname{Obs}(X) \mid p \geq \mathbf{0}\}
$$

Factors can be added pointwise, with identity element $\mathbf{0} \in \operatorname{Fact}(X)$, like random variables. But a factor $p \in \operatorname{Fact}(X)$ cannot be re-scaled with an arbitrary real number, but only with a non-negative number $s \in \mathbb{R}_{\geq 0}$, giving $s \cdot p \in \operatorname{Fact}(X)$. These structures are often called cones. The cone $\operatorname{Fact}(X)$ 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 $\operatorname{Obs}(X)$ restricts to $\operatorname{Fact}(X)$, since $\mathbf{1} \geq \mathbf{0}$ and if $p, q \geq \mathbf{0}$ then also $p \& q \geq \mathbf{0}$.

### 4.2.3 Predicates

We first note that the set $\operatorname{Pred}(X)=[0,1]^{X}$ 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 $p \otimes q$. Thus: $p \otimes q$ is defined if $p(x)+q(x) \leq 1$ for all $x \in X$, and in that case $(p \boxtimes q)(x)=p(x)+q(x)$.

This operation $\otimes$ is commutative and associative in a suitably partial sense. Moreover, it has $\mathbf{0}$ as identity element: $p \boxtimes \mathbf{0}=p=\mathbf{0} \otimes p$. This is structure $(\operatorname{Pred}(X), \mathbf{0}, \otimes)$ is called a partial commutative monoid, see below for details.

- There is a 'negation' of predicates, written as $p^{\perp}$, and called orthosupple$m e n t$. It is defined as $p^{\perp}=\mathbf{1}-p$, that is, as $p^{\perp}(x)=1-p(x)$. Then: $p \boxtimes p^{\perp}=\mathbf{1}$ 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 $\boxtimes$, in the sense that $s \cdot(p \boxtimes q)=(s \cdot p) \boxtimes(s \cdot q)$.

The combination of these items means that the set $\operatorname{Pred}(X)$ carries the structure of an effect module [73], see also [48]. These structures arose in mathematical physics [55] in order to axiomatise the structure of quantum predicates on Hilbert spaces.

The effect module $\operatorname{Pred}(X)$ also carries a commutative monoid structure for conjunction, namely $(\mathbf{1}, \&)$. Indeed, when $p, q \in \operatorname{Pred}(X)$, then also $p \& q \in$ $\operatorname{Pred}(X)$. We have $p \& \mathbf{0}=\mathbf{0}$ and $p \&\left(q_{1} \otimes q_{2}\right)=\left(p \& q_{1}\right) \otimes\left(p \& q_{2}\right)$.

Since these effect structures are not so familiar, we include more formal descriptions.

## Definition 4.2.1.

1 A partial commutative monoid (PCM) consists of a set $M$ with a zero element $0 \in M$ and a partial binary operation $\boxtimes: M \times M \rightarrow M$ satisfying
the three requirements below. They involve the notation $x \perp y$ for: $x \boxtimes y$ is defined; in that case $x, y$ are called orthogonal.

- Commutativity: $x \perp y$ implies $y \perp x$ and $x \boxtimes y=y \boxtimes x$;
- Associativity: $y \perp z$ and $x \perp(y \boxtimes z)$ implies $x \perp y$ and $(x \boxtimes y) \perp z$ and also $x \otimes(y \boxtimes z)=(x \otimes y) \otimes z$;
- Zero: $0 \perp x$ and $0 \otimes x=x$;

2 An effect algebra is a $\operatorname{PCM}(E, 0, \otimes)$ with an orthosupplement. The latter is a total unary 'negation' operation $(-)^{\perp}: E \rightarrow E$ satisfying:

- $x^{\perp} \in E$ is the unique element in $E$ with $x \otimes x^{\perp}=1$, where $1=0^{\perp}$;
- $x \perp 1 \Rightarrow x=0$.

A homomorphism $E \rightarrow D$ of effect algebras is given by a function $f: E \rightarrow$ $D$ between the underlying sets satisfying $f(1)=1$, and if $x \perp x^{\prime}$ in $E$ then both $f(x) \perp f\left(x^{\prime}\right)$ in $D$ and $f\left(x \otimes x^{\prime}\right)=f(x) \otimes f\left(x^{\prime}\right)$. 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:

$$
1 \cdot x=x \quad(r \cdot s) \cdot x=r \cdot(s \cdot x)
$$

and preserving sums (that exist) in both arguments:

$$
\begin{array}{ll}
0 \cdot x=0 & (r+s) \cdot x=r \cdot x \boxtimes s \cdot x \\
s \cdot 0=0 & s \cdot(x \boxtimes y)=s \cdot x \otimes s \cdot y .
\end{array}
$$

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. [40, 141]. It will be used occasionally later on, for instance in Exercise 6.1.6.

Definition 4.2.2. A test or, more explicitly, an $n$-test on a set $X$ is an $n$-tuple of predicates $p_{1}, \ldots, p_{n}: X \rightarrow[0,1]$ satisfying $p_{1} \otimes \cdots \otimes p_{n}=\mathbf{1}$.

This notion of test can be formulated in an arbitrary effect algebra. Here we do it in $\operatorname{Pred}(X)$ only.

Each predicate $p$ forms a 2-test $p, p^{\perp}$. Exercise 4.2 .13 asks to show that an $n$-test of predicates on $X$ corresponds to a channel $X \leadsto \boldsymbol{n}$. In particular, each predicate on $X$ corresponds to a channel $X \rightarrow \mathbf{2}$, see Exercise 4.3.6. The probabilities $\omega\left(x_{i}\right) \in[0,1]$ 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 4.2.3. Consider a finite set $X=\left\{x_{1}, \ldots, x_{n}\right\}$.
1 The point predicates $\mathbf{1}_{x_{1}}, \ldots, \mathbf{1}_{x_{n}}$ form a test on $X$.
2 Each predicate $p: X \rightarrow[0,1]$ has a normal form $p=\bigotimes_{i} r_{i} \cdot \mathbf{1}_{x_{i}}$, with scalars $r_{i}=p\left(x_{i}\right) \in[0,1]$.

On a set $X=\{a, b, c\}$ we can write a distribution and a predicate as:

$$
\frac{1}{3}|a\rangle+\frac{1}{6}|b\rangle+\frac{1}{2}|c\rangle \quad \frac{2}{3} \cdot \mathbf{1}_{a} \otimes \frac{5}{6} \cdot \mathbf{1}_{b} \otimes 1 \cdot \mathbf{1}_{c} .
$$

Writing $\otimes$ looks a bit pedantic, so we often simply write + instead. Recall that in a predicate the probabilities need not add up to one, see Remark4.1.7

A set of predicates $p_{1}, \ldots, p_{n}$ on the same space, can be turned into a test via (pointwise) normalisation. Below we describe an alternative construction 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 4.2.4. Let $p_{1}, \ldots, p_{n}$ be arbitrary predicates, all on the same set, for $n \geq 1$. We can turn them via "stick breaking" into an $n+1$-test $q_{1}, \ldots, q_{n+1}$ via definitions:

$$
\begin{array}{rlr}
q_{1} & :=p_{1} \\
q_{i+1} & :=p_{1}^{\perp} \& \cdots \& p_{i}^{\perp} \& p_{i+1} \quad \text { for } 0<i<n \\
q_{n+1} & :=p_{1}^{\perp} \& \cdots \& p_{n}^{\perp} .
\end{array}
$$

### 4.2.4 Sharp predicates

The structure of the set $\operatorname{SPred}(X)=\{0,1\}^{X}=\mathbf{2}^{X}$ is most familiar to logicians: it is a Boolean algebra. The join $p \vee q$ of $p, q \in \operatorname{SPred}(X)$ 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 $(\mathbf{0}, \mathrm{V})$ as forming an additive structure with negation $(-)^{\perp}$. The partial sum $\otimes$ restricts from $\operatorname{Pred}(X)$ to $\operatorname{SPred}(X)$. For sharp predicates $p, q$ the sum $p \boxtimes q$ is defined iff $p, q$ are disjoint, as subsets. Conjunction $(\mathbf{1}, \&)$ forms an additional commutative monoid structure on $\operatorname{SPred}(X)$.

Formally one can say that $\operatorname{SPred}(X)$ 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 4.2.12 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 [89, 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 4.2.5.

1 For an arbitrary set $X$, the indicator function

$$
\operatorname{SPred}(X) \cong \mathcal{P}(X) \xrightarrow[\mathbf{1}_{(-)}]{ } \operatorname{Pred}(X)
$$

is a homomorphism of effect algebras.
2 Let $X$ now be finite. This indicator map $\mathbf{1}_{(-)}$makes $\operatorname{Pred}(X)$ into the free effect module on $\mathcal{P}(X)$, as described below: for each effect module $E$ with a map of effect algebras $f: \mathcal{P}(X) \rightarrow E$, there is a unique map of effect modules $\bar{f}: \operatorname{Pred}(X) \rightarrow E$ in:


## Proof. 1 See Exercise 4.2.2.

2 We use that each predicate $p \in \operatorname{Pred}(X)$ can be written in normal form as $p=\mathbb{Q}_{x \in X} p(x) \cdot \mathbf{1}_{x}$, see Lemma 4.2.3 (2). So we define $\bar{f}$ as:

$$
\bar{f}(p):=\bigotimes_{x \in X} p(x) \cdot f(\{x\}) .
$$

Obviously, $\bar{f}(\mathbf{0})=\mathbf{0}$. Also, since $f$ preserves $\otimes$ and top,

$$
\bar{f}(\mathbf{1})=\bigotimes_{x \in X} 1 \cdot f(\{x\})=f\left(\bigcup_{x \in X}\{x\}\right)=f(X)=1
$$

If $p \perp q$, then $p(x)+q(x) \leq 1$ for each $x \in X$. Hence:

$$
\begin{aligned}
\bar{f}(p \boxtimes q)=\bigotimes_{x \in X}(p \boxtimes q)(x) \cdot f(\{x\}) & =\bigotimes_{x \in X}(p(x)+q(x)) \cdot f(\{x\}) \\
& =\bigotimes_{x \in X} p(x) \cdot f(\{x\}) \boxtimes q(x) \cdot f(\{x\}) \\
& =\bigotimes_{x \in X} p(x) \cdot f(\{x\}) \boxtimes \bigotimes_{x \in X} q(x) \cdot f(\{x\}) \\
& =\bar{f}(p) \boxtimes \bar{f}(q) .
\end{aligned}
$$

Scalar multiplication is also preserved:

$$
\begin{aligned}
\bar{f}(r \cdot p)=\bigotimes_{x \in X}(r \cdot p)(x) \cdot f(\{x\}) & =\bigotimes_{x \in X} r \cdot(p(x) \cdot f(\{x\})) \\
& =r \cdot \bigotimes_{x \in X} p(x) \cdot f(\{x\}) \\
& =r \cdot \bar{f}(p) .
\end{aligned}
$$

Finally, for uniqueness, let $g: \operatorname{Pred}(X) \rightarrow E$ be a map of effect modules with $g\left(\mathbf{1}_{U}\right)=f(U)$ for each $U \in \mathcal{P}(X)$. Then $g=\bar{f}$ since:

$$
\begin{aligned}
\bar{f}(p)=\bigotimes_{x \in X} p(x) \cdot f(\{x\}) & =\emptyset_{x \in X} p(x) \cdot g\left(\mathbf{1}_{x}\right) \\
& =g\left(\emptyset_{x \in X} p(x) \cdot \mathbf{1}_{x}\right)=g(p) .
\end{aligned}
$$

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 $(1, \&)$.

Lemma 4.2.6. Let $\omega \in \mathcal{D}(X)$ be a distribution on a set $X$. Operations on observables $p, q$ on $X$ satisfy, whenever defined,

$$
\begin{aligned}
& 1 \omega \vDash \mathbf{0}=0 ; \\
& 2 \omega \vDash(p+q)=(\omega \vDash p)+(\omega \vDash q) ; \\
& 3 \omega \vDash p^{\perp}=1-(\omega \vDash p) ; \\
& 4 \omega \vDash(s \cdot p)=s \cdot(\omega \vDash p) .
\end{aligned}
$$

### 4.2.5 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 $=\lll$ of observables, see Section 4.3 - in particular Exercise 4.3.8.

## Definition 4.2.7.

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 $(p \otimes \mathbf{1})(x, y)=$ $p(x)$.

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:

$$
\underbrace{1 \otimes \cdots \otimes \mathbf{1}}_{i-1 \text { times }} \otimes p \otimes \underbrace{\mathbf{1} \otimes \cdots \otimes \mathbf{1}}_{n-i \text { times }} .
$$

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: for a state $\omega \in \mathcal{D}\left(X_{1} \times \cdots \times X_{n}\right)$ and an observable $p$ on $X_{i}$ we have:

$$
\begin{align*}
\omega \vDash \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes p \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} & =\mathcal{D}\left(\pi_{i}\right)(\omega) \vDash p  \tag{4.7}\\
& =\omega[0, \ldots, 0,1,0, \ldots, 0] \vDash p,
\end{align*}
$$

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 4.3.3

We illustrate how to solve a famous riddle via validity.
Example 4.2.8. The so-called Monty Hall problem is a famous riddle in probability theory, due to [164], see also e.g. [66, 174]:

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?

The first choice of you as participant can be described as the uniform distribution $\omega=\frac{1}{3}|1\rangle+\frac{1}{3}|2\rangle+\frac{1}{3}|3\rangle$ over the set of doors $D=\{1,2,3\}$. Let us assume, without loss of generality, that the car is behind door 2 . The win predicate $p: D \rightarrow[0,1]$ is then the point predicate $p=\mathbf{1}_{2}$. Obviously, at the start you have one third chance $\omega \models p=\frac{1}{3}$ of winning.

We describe the act of opening the door by the host via the channel op: D $\rightarrow$ $D$ given by:

$$
o p(1)=1|3\rangle \quad o p(2)=\frac{1}{2}|1\rangle+\frac{1}{2}|2\rangle \quad o p(3)=1|1\rangle
$$

The input to this channel is your choice of door. Since the host opens a door, different from the one you chose, with a goat behind it, there is only one option if your choice is door 1 or door 3 . When you chose door 2 , the host can open either door 1 or door 2 . We use a uniform distribution to cover both options.

At this stage we can form the joint 'graph' state $\tau:=\langle i d, o p\rangle\rangle=\omega \in \mathcal{D}(D \times$ $D)$. It is:

$$
\tau=\frac{1}{3}|1,3\rangle+\frac{1}{6}|2,1\rangle+\frac{1}{6}|2,3\rangle+\frac{1}{3}|3,1\rangle .
$$

The first number inside the ket is your choice of door, and the second number is for the door that the host opens. In order to now compute your chance of winning, we have to weaken the predicate $p$ on $D$ to $p \otimes \mathbf{1}$ on $D \times D$. At this stage your chance is still one third:

$$
\tau \vDash p \otimes \mathbf{1}=\frac{1}{6}+\frac{1}{6}=\frac{1}{3} .
$$

Now suppose you change your choice. Such a change is deterministic, since in each situation there is one other door that you can open. This result of your decision to change transforms $\tau$ into $\tau^{\prime}$ below.

$$
\tau^{\prime}=\frac{1}{3}|2,3\rangle+\frac{1}{6}|3,1\rangle+\frac{1}{6}|1,3\rangle+\frac{1}{3}|2,1\rangle .
$$

Your chance of winning is now twice as big:

$$
\tau^{\prime} \vDash p \otimes \mathbf{1}=\frac{1}{3}+\frac{1}{3}=\frac{2}{3}
$$

This shows that it makes sense to change.
We have given a formal account of the situation. More informally, the host knows where the car is, so his choice is not arbitrary. By opening a door with a goat behind it, the host is giving you information that you can exploit to improve your choice: two of your possible choices are wrong, but in those two out three cases the host gives you information how to correct your choice.

We mention some basic results about parallel products of observables. More such 'logical' results can be found in the exercises.

Lemma 4.2.9. For states $\sigma \in \mathcal{D}(X), \tau \in \mathcal{D}(Y)$ and observables $p \in \operatorname{Obs}(X)$, $q \in \operatorname{Obs}(Y)$ one has:

$$
(\sigma \otimes \tau \vDash p \otimes q)=(\sigma \vDash p) \cdot(\tau \vDash q) .
$$

Proof. Easy:

$$
\begin{aligned}
(\sigma \otimes \tau \vDash p \otimes q) & =\sum_{z \in X \times Y}(\sigma \otimes \tau)(z) \cdot(p \otimes q)(z) \\
& =\sum_{x \in X, y \in Y}(\sigma \otimes \tau)(x, y) \cdot(p \otimes q)(x, y) \\
& =\sum_{x \in X, y \in Y} \sigma(x) \cdot \tau(y) \cdot p(x) \cdot q(y) \\
& =\left(\sum_{x \in X} \sigma(x) \cdot p(x)\right) \cdot\left(\sum_{y \in Y} \tau(y) \cdot q(y)\right) \\
& =(\sigma \vDash p) \cdot(\tau \vDash q) .
\end{aligned}
$$

## Lemma 4.2.10.

1 For observables $p_{i}, q_{i} \in \operatorname{Obs}\left(X_{i}\right)$ one has:

$$
\left(p_{1} \otimes \cdots \otimes p_{n}\right) \&\left(q_{1} \otimes \cdots \otimes q_{n}\right)=\left(p_{1} \& q_{1}\right) \otimes \cdots \otimes\left(p_{n} \& q_{n}\right) .
$$

2 Parallel composition $p \otimes q$ of observables $p \in \operatorname{Obs}(X)$ and $q \in \operatorname{Obs}(Y)$ can be defined in terms of weakening and conjunction, namely as:

$$
p \otimes q=(p \otimes \mathbf{1}) \&(\mathbf{1} \otimes q) .
$$

Proof. 1 For elements $x_{i} \in X_{i}$,

$$
\begin{aligned}
& \left(\left(p_{1} \otimes \cdots \otimes p_{n}\right) \&\left(q_{1} \otimes \cdots \otimes q_{n}\right)\right)\left(x_{1}, \ldots, x_{n}\right) \\
& =\left(p_{1} \otimes \cdots \otimes p_{n}\right)\left(x_{1}, \ldots, x_{n}\right) \cdot\left(q_{1} \otimes \cdots \otimes q_{n}\right)\left(x_{1}, \ldots, x_{n}\right) \\
& =\left(p_{1}\left(x_{1}\right) \cdot \ldots \cdot p_{n}\left(x_{n}\right)\right) \cdot\left(q_{1}\left(x_{1}\right) \cdot \ldots \cdot q_{n}\left(x_{n}\right)\right) \\
& =\left(p_{1}\left(x_{1}\right) \cdot q_{1}\left(x_{1}\right)\right) \cdot \ldots \cdot\left(p_{n}\left(x_{n}\right) \cdot q_{n}\left(x_{n}\right)\right) \\
& =\left(p_{1} \& q_{1}\right)\left(x_{1}\right) \cdot \cdots \cdot\left(p_{n} \& q_{n}\right)\left(x_{n}\right) \\
& =\left(\left(p_{1} \& q_{1}\right) \otimes \cdots \otimes\left(p_{n} \& q_{n}\right)\right)\left(x_{1}, \ldots, x_{n}\right) .
\end{aligned}
$$

2 Directly from the previous item:

$$
(p \otimes \mathbf{1}) \&(\mathbf{1} \otimes q)=(p \& \mathbf{1}) \otimes(\mathbf{1} \& q)=p \otimes q .
$$

We conclude with some observations about proving equalities of states or predicates via validity.

Remark 4.2.11. For two states $\omega, \omega^{\prime} \in \mathcal{D}\left(X_{1} \times \cdots \times X_{n}\right)$ we have:

$$
\begin{gather*}
\omega=\omega^{\prime} \Longleftrightarrow \omega \vDash p_{1} \otimes \cdots \otimes p_{n}=\omega^{\prime} \vDash p_{1} \otimes \cdots \otimes p_{n} \\
\text { for all } p_{i} \in \operatorname{Pred}\left(X_{i}\right) . \tag{4.8}
\end{gather*}
$$

The direction $(\Rightarrow)$ is obvious. For $(\Leftarrow)$ we use for an arbitrary $x_{i} \in X_{i}$ the corresponding tuple point predicate satisfies $\mathbf{1}_{\left(x_{1}, \ldots, x_{n}\right)}=\mathbf{1}_{x_{1}} \otimes \cdots \otimes \mathbf{1}_{x_{n}}$. Hence:

$$
\begin{aligned}
\omega\left(x_{1}, \ldots, x_{n}\right)=\omega \vDash \mathbf{1}_{\left(x_{1}, \ldots, x_{n}\right)} & =\omega \vDash \mathbf{1}_{x_{1}} \otimes \cdots \otimes \mathbf{1}_{x_{n}} \\
& =\omega^{\prime} \vDash \mathbf{1}_{x_{1}} \otimes \cdots \otimes \mathbf{1}_{x_{n}} \\
& =\omega^{\prime} \vDash \mathbf{1}_{\left(x_{1}, \ldots, x_{n}\right)}=\omega^{\prime}\left(x_{1}, \ldots, x_{n}\right) .
\end{aligned}
$$

Similarly, for predicates $p, p^{\prime} \in \operatorname{Pred}\left(X_{1} \times \cdots \times X_{n}\right)$ one has:

$$
\begin{gather*}
p=p^{\prime} \Longleftrightarrow \omega_{1} \otimes \cdots \otimes \omega_{n} \vDash p=\omega_{1} \otimes \cdots \otimes \omega_{n} \vDash p^{\prime},  \tag{4.9}\\
\text { for all } \omega_{i} \in \mathcal{D}\left(X_{i}\right) .
\end{gather*}
$$

Again, $(\Rightarrow)$ is obvious and for $(\Leftarrow)$ we use point states:

$$
\begin{aligned}
p\left(x_{1}, \ldots, x_{n}\right)=1\left|x_{1}, \ldots, x_{n}\right\rangle \vDash p & =1\left|x_{1}\right\rangle \otimes \cdots \otimes 1\left|x_{n}\right\rangle \vDash p \\
& =1\left|x_{1}\right\rangle \otimes \cdots \otimes 1\left|x_{n}\right\rangle \vDash p^{\prime} \\
& =1\left|x_{1}, \ldots, x_{n}\right\rangle \vDash p^{\prime}=p^{\prime}\left(x_{1}, \ldots, x_{n}\right) .
\end{aligned}
$$

Thus, proofs of equality of states, or of predicates, can be performed via validity $\mid=$. This may be convenient in situations where certain properties of validity can be used.

Here is a frequently occurring situation where these observations apply.

## Lemma 4.2.12.



Algebraically (or categorically) oriented readers may recognise an adjointness in the formulation on the right-hand-side.

Proof. This follows from the equivalence (4.8) in Remark 4.2.11 using that,
by Lemma 4.3.2 (7) and Proposition 4.3.3

$$
\begin{aligned}
c(x) \vDash(f=\ll p) \& q & =c(x) \vDash\langle f, \text { id }\rangle=\ll(p \otimes q) \\
& =\langle f, \text { id }\rangle \gg c(x) \vDash p \otimes q \\
d(x) \vDash p \&(g=\ll q) & =d(x) \vDash\langle\mathrm{id}, g\rangle=\ll(p \otimes q) \\
& =\langle\mathrm{id}, g\rangle \gg d(x) \vDash p \otimes q .
\end{aligned}
$$

### 4.2.6 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 4.2.1;
- BA, with Boolean algebras as objects and homomorphisms of Boolean algebras (preserving finite joins and negations, and then also finite meets).

Recall from Subsection 1.11 .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 4.2.13. Taking particular observables on a set is functorial: there are functors:

1 Obs: Sets $\rightarrow$ Vect $^{\text {op }}$;
2 Fact: Sets $\rightarrow$ Cone $^{\text {op }}$;
3 Pred: Sets $\rightarrow$ EMod $^{\text {op }}$;
4 SPred: Sets $\rightarrow \mathbf{B A}^{\mathrm{op}}$.
On maps $f: X \rightarrow Y$ 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 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 $\operatorname{Obs}(X)=\mathbb{R}^{X}$ is a vector space, and thus an object of the category Vect. Each function $f: X \rightarrow Y$ in Sets gives rise to a function $\operatorname{Obs}(f): \operatorname{Obs}(Y) \rightarrow \operatorname{Obs}(X)$ in the opposite direction. It
maps an observable $q: Y \rightarrow \mathbb{R}$ on $Y$ to the observable $q \circ f: X \rightarrow \mathbb{R}$ on $X$. It is not hard to see that this function $\operatorname{Obs}(f)=(-) \circ f$ 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 $\operatorname{Obs}(f)(\mathbf{0})$ is the function that map $x \in X$ to:

$$
\operatorname{Obs}(f)(\mathbf{0})(x)=(\mathbf{0} \circ f)(x)=\mathbf{0}(f(x))=0
$$

Hence $\operatorname{Obs}(f)(\mathbf{0})$ maps everything to 0 and is thus equal to the zero function itself: $\operatorname{Obs}(f)(\mathbf{0})=\mathbf{0}$. Next, addition + is preserved since:

$$
\begin{aligned}
\operatorname{Obs}(f)(p+q) & =(p+q) \circ f \\
& =[x \mapsto(p+q)(f(x))] \\
& =[x \mapsto p(f(x))+q(f(x))] \\
& =[x \mapsto O b s(f)(p)(x)+\operatorname{Obs}(f)(q)(x)] \\
& =\operatorname{Obs}(f)(p)+\operatorname{Obs}(f)(q) .
\end{aligned}
$$

We leave preservation of scalar multiplication to the reader and conclude that $\operatorname{Obs}(f)$ is a linear function, and thus a morphism $\operatorname{Obs}(f): \operatorname{Obs}(Y) \rightarrow \operatorname{Obs}(X)$ in Vect. Hence $\operatorname{Obs}(f)$ is a morphism $\operatorname{Obs}(X) \rightarrow \operatorname{Obs}(Y)$ in the opposite category Vect ${ }^{\text {op }}$. We still need to check that identity maps and composition are preserved. We do the latter. For $f: X \rightarrow Y$ and $g: Y \rightarrow Z$ in Sets we have, for $r \in O b s(Z)$,

$$
\begin{aligned}
\operatorname{Obs}(g \circ f)(r)=r \circ(g \circ f) & =(r \circ g) \circ f \\
& =\operatorname{Obs}(g)(r) \circ f \\
& =O b s(f)(\operatorname{Obs}(g)(r)) \\
& =(\operatorname{Obs}(f) \circ \operatorname{Obs}(g))(r) \\
& =(\operatorname{Obs}(g) \circ \text { op } \operatorname{Obs}(f))(r) .
\end{aligned}
$$

This yields $\operatorname{Obs}(g \circ f)=\operatorname{Obs}(g) \circ^{\text {op }} \operatorname{Obs}(f)$, so that we get a functor of the form Obs: Sets $\rightarrow$ Vect $^{\text {op }}$.

Notice that saying that we have a functor like Pred : Sets $\rightarrow$ EMod $^{\text {op }}$ contains remarkably much information, about the mathematical structure on objects $\operatorname{Pred}(X)$, about preservation of this structure by maps $\operatorname{Pred}(f)$, and about preservation of identity maps and composition by Pred( - ) on morphisms (see also Exercise 4.3.10. This makes the language of category theory both powerful and efficient.

## Exercises

4.2.1 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:

$$
\frac{11}{25}=F \operatorname{lrn}(4|R\rangle+6|B\rangle) \otimes \operatorname{Flrn}(16|R\rangle+x|B\rangle) \vDash\left(\mathbf{1}_{R} \otimes \mathbf{1}_{R}\right) \otimes\left(\mathbf{1}_{B} \otimes \mathbf{1}_{B}\right)
$$

2 Prove, by solving the above equation, that there are 4 blue balls in the second urn.
4.2.2 1 Check that (sharp) indicator predicates $\mathbf{1}_{E}: X \rightarrow[0,1]$, for subsets $E \subseteq X$, satisfy:

- $\mathbf{1}_{E \cap D}=\mathbf{1}_{E} \& \mathbf{1}_{D}$, and thus $\mathbf{1} \otimes \mathbf{1}=\mathbf{1}, \mathbf{1} \otimes \mathbf{0}=\mathbf{0}=\mathbf{0} \otimes \mathbf{1}$;
- $\mathbf{1}_{E \cup D}=\mathbf{1}_{E} \otimes \mathbf{1}_{D}$, if $E, D$ are disjoint;
- $\left(\mathbf{1}_{E}\right)^{\perp}=\mathbf{1}_{\neg E}$, where $\neg E=X \backslash E=\{x \in X \mid x \notin E\}$ is the complement of $E$.
Formally, the function $\mathbf{1}_{(-)}: \mathcal{P}(X) \rightarrow \operatorname{Pred}(X)=[0,1]^{X}$ is a homomorphism of effect algebras, see Theorem 4.2.5 (1).
2 Show that this indicator function is natural, in the sense that for $f: X \rightarrow Y$ the following diagram commutes.


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 $\mathbf{1}_{(x, y)}=\mathbf{1}_{x} \otimes \mathbf{1}_{y}$.
4.2.3 Find examples of predicates $p, q$ on a set $X$ and a distribution $\omega$ on $X$ such that $\omega \vDash p \& q$ and $(\omega \vDash p) \cdot(\omega \vDash q)$ are different.
4.2.4 One may expect the following implication between inequalities of validities:

$$
(\sigma \vDash p) \leq(\tau \vDash p) \Longrightarrow(\sigma \vDash p \& q) \leq(\tau \vDash p \& q) .
$$

However, it fails. This exercise elaborates a counterexample. Take $X=\{a, b, c\}$ with states:

$$
\sigma=\frac{19}{100}|a\rangle+\frac{47}{100}|b\rangle+\frac{17}{50}|c\rangle \quad \tau=\frac{1}{5}|a\rangle+\frac{9}{20}|b\rangle+\frac{7}{20}|c\rangle
$$

with predicates:

$$
p=1 \cdot \mathbf{1}_{a}+\frac{7}{10} \cdot \mathbf{1}_{b}+\frac{1}{2} \cdot \mathbf{1}_{c} \quad q=\frac{1}{10} \cdot \mathbf{1}_{a}+\frac{1}{2} \cdot \mathbf{1}_{b}+\frac{1}{5} \cdot \mathbf{1}_{c} .
$$

Now check consecutively that:
$\sigma \vDash p=\frac{689}{1000}<\frac{69}{100}=\tau \vDash p$.
$p \& q=\frac{1}{10} \cdot \mathbf{1}_{a}+\frac{7}{20} \cdot \mathbf{1}_{b}+\frac{1}{10} \cdot \mathbf{1}_{c}$.
$\sigma \vDash p \& q=\frac{87}{400}>\frac{85}{400}=\tau \vDash p \& q$.
Find a counterexample yourself in which the predicate $q$ is sharp.
4.2.5 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 $\sigma$. Show that:

$$
\left(\sigma \vDash \frac{p}{q}\right) \geq 1 \Longrightarrow(\sigma \vDash p) \geq(\sigma \vDash q)
$$

where $\frac{p}{q}(x)=\frac{p(x)}{q(x)}$.
4.2.6 Prove that the following items are equivalent, for a state $\omega \in \mathcal{D}(X)$ and event $E \subseteq X$.

```
supp (\omega)\subseteqE;
\omega}=\mp@subsup{\mathbf{1}}{E}{}=1
\omega\vDashp& 1}\mp@subsup{\mathbf{1}}{E}{}=\omega\vDashp\mathrm{ for each }p\in\operatorname{Obs}(X)
```

4.2.7 Consider a distribution $\omega \in \mathcal{D}(X)$, a channel $c: X \leadsto Y$ and an observable $q: Y \rightarrow \mathbb{R}$. Show that:

$$
c \gg \omega \vDash q=\omega \vDash(c(-) \vDash q) .
$$

This is known as the 'law of total expectation' or also the 'law of iterated expectation'. An equivalent form using predicate transformation appears in Proposition 4.3.3.
4.2.8 Let $\omega \in \mathcal{D}(X)$ be an arbitrary distribution with two observables $p, q \in$ $\operatorname{Obs}(X)$.
1 Show that there is an inequality:

$$
(\omega \vDash p \& q)^{2} \leq\left(\omega \vDash p^{2}\right) \cdot\left(\omega \vDash q^{2}\right)
$$

Hint: Recall $p^{2}=p \& p$. Handle the case $\left(\omega \vDash p^{2}\right)=0$ separately. Then define $r \in \operatorname{Obs}(X)$ as $r:=q-\frac{\omega \vDash p \& q}{\omega \vDash p^{2}} \cdot p$ and exploit that $\left(\omega \vDash r^{2}\right) \geq 0$ since $r^{2} \geq 0$.
2 Deduce the inequality:

$$
(\omega \vDash p)^{2} \leq \omega \vDash p^{2} .
$$

4.2.9 Let $p$ and $q$ be two arbitrary predicates. Prove that the following use of the partial sum operation $\otimes$ for predicates is justified, that is, yields a well-defined new predicate.

$$
(p \otimes q) \otimes\left(p^{\perp} \otimes q^{\perp}\right)
$$

4.2.10 Let $p_{1}, p_{2}, p_{3}$ be predicates on the same set.

1 Show that:

$$
\left(p_{1}^{\perp} \otimes p_{2}^{\perp}\right)^{\perp}=\left(p_{1} \otimes p_{2}\right) \otimes\left(p_{1} \otimes p_{2}^{\perp}\right) \otimes\left(p_{1}^{\perp} \otimes p_{2}\right)
$$

2 Show also that:

$$
\begin{gathered}
\left(p_{1}^{\perp} \otimes p_{2}^{\perp} \otimes p_{3}^{\perp}\right)^{\perp}=\left(p_{1} \otimes p_{2} \otimes p_{3}\right) \otimes\left(p_{1} \otimes p_{2} \otimes p_{3}^{\perp}\right) \\
\\
\otimes\left(p_{1} \otimes p_{2}^{\perp} \otimes p_{3}\right) \otimes\left(p_{1} \otimes p_{2}^{\perp} \otimes p_{3}^{\perp}\right) \\
\otimes\left(p_{1}^{\perp} \otimes p_{2} \otimes p_{3}\right) \otimes\left(p_{1}^{\perp} \otimes p_{2} \otimes p_{3}^{\perp}\right) \\
\\
\otimes\left(p_{1}^{\perp} \otimes p_{2}^{\perp} \otimes p_{3}\right)
\end{gathered}
$$

3 Generalise to $n$.
4.2.11 For predicates $p, q$ on the same set, define Reichenbach implication $\supset$ as:

$$
p \supset q:=p^{\perp} \otimes(p \& q)
$$

1 Check that:

$$
p \supset q=\left(p \& q^{\perp}\right)^{\perp}
$$

from which it easily follows that:

$$
p \supset \mathbf{0}=p^{\perp} \quad \mathbf{1} \supset q=q \quad p \supset \mathbf{1}=\mathbf{1} \quad \mathbf{0} \supset q=\mathbf{1} .
$$

2 Check also that:

$$
p^{\perp} \leq p \supset q \quad q \leq p \supset q
$$

3 Show that:

$$
p_{1} \supset\left(p_{2} \supset q\right)=\left(p_{1} \& p_{2}\right) \supset q .
$$

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).
4.2.12 Let $p$ be a predicate on a set $X$. Prove that the following statements are equivalent.
$1 p$ is sharp;
$2 p \& p=p$;
$3 p \& p^{\perp}=\mathbf{0}$;
$4 \quad q \leq p$ and $q \leq p^{\perp}$ implies $q=\mathbf{0}$, for each $q \in \operatorname{Pred}(X)$.
4.2.13 Show that an $n$-test $p_{0}, \ldots, p_{n-1}$ on a set $X$, see Definition 4.2.2, can be identified with a channel $c: X \leadsto \boldsymbol{n}$, with $p_{i}=c=\ll \mathbf{1}_{i}$, for $i \in \boldsymbol{n}$.
4.2.14 For a random variable ( $\omega, p$ ), show that the validity of the observable $p-(\omega \vDash p) \cdot \mathbf{1}$ is zero, i.e.,

$$
\omega \vDash(p-(\omega \models p) \cdot \mathbf{1})=0 .
$$

Observables of this form are used later on to define (co)variance in Section 5.1
4.2.15 For a multiset $\varphi \in \mathcal{M}(X)$ on a set $X$ and a factor $p \in \operatorname{Fact}(X)$ 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:

$$
\mathcal{M}(X) \times \operatorname{Fact}(X) \xrightarrow{\bullet} \mathcal{M}(X)
$$

with respect to the multiplicative monoid structure $(\mathbf{1}, \&)$ on factors.
4.2.16 Let $E$ be an arbitrary effect algebra. Prove, from the axioms of an effect algebra, that for elements $e, e^{\prime}, d, d^{\prime}, f \in E$ the following properties hold (see [73] for details).
1 Orthosupplement is an involution: $e^{\perp \perp}=e$;
2 Cancellation holds of the form: $e \oslash d=e \boxtimes d^{\prime}$ implies $d=d^{\prime}$;
3 Zero-sum freeness holds: $e \boxtimes d=0$ implies $e=d=0$;
4 Define an order $\leq$ on $E$ via: $e \leq d$ iff $e \oslash f=d$ for some $f \in E$.
This is a partial order with 1 as top and 0 as bottom element;
$e \leq d$ implies $d^{\perp} \leq e^{\perp}$;
$6 e \otimes d$ is defined iff $e \perp d$ iff $e \leq d^{\perp}$ iff $d \leq e^{\perp}$;
$7 \quad e \leq d$ and $d \perp f$ implies $e \perp f$ and $e \oslash f \leq d \otimes f$;
8 if $e \leq e^{\prime}$ and $d \leq d^{\prime}$ and $e^{\prime} \boxtimes d^{\prime}$ is defined, then also $e \boxtimes d$ is defined.
4.2.17 In an effect algebra $E$, as above, define for elements $e, d \in E$,

$$
\begin{array}{ll}
e \otimes d:=\left(e^{\perp} \boxtimes d^{\perp}\right)^{\perp} & \text { if } e^{\perp} \perp d^{\perp} \\
e \ominus d:=\left(e^{\perp} \boxtimes d\right)^{\perp}=e \otimes d^{\perp} & \text { if } e \geq d .
\end{array}
$$

Show that:

```
\((E, \otimes, 1)\) is a partial commutative monoid;
\(e \leq d\) iff \(e=d \otimes f\) for some \(f\);
\(e \ominus 0=e\) and \(1 \ominus e=e^{\perp}\) and \(e \ominus e=0 ;\)
\(e \boxtimes d=f\) iff \(e=f \ominus d\); in particular, \((f \ominus d) \boxtimes d=f ;\)
\(e \oslash d \leq f\) iff \(e \leq f \ominus d\);
\(e>d\) iff \(e \ominus d>0\);
\(f \ominus e=f \ominus d\) implies \(e=d ;\)
\(e \leq d\) implies \(d \ominus e \leq d\) and \(d \ominus(d \ominus e)=e ;\)
```

9 the function $e \oslash(-)$ 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 $e \otimes\left(\bigvee_{i} d_{i}\right)=\bigvee_{i}\left(e \otimes d_{i}\right)$;
10 Similarly, $e \oplus(-)$ preserves meets.
11 A homomorphism $f: E \rightarrow D$ of effect algebras $E, D$ preserves orthosupplement: $f\left(e^{\perp}\right)=f(e)^{\perp}$, and thus also $f(0)=0, f(e \otimes d)=$ $f(e) \otimes f(d)$ and $f(e \ominus d)=f(e) \ominus f(d)$.
4.2.18 Let $E$ be an effect module. The aim of this exercise is to show that subconvex sums exist in $E$, see [155] Lemma 2.1]. This means that for arbitrary elements $e_{1}, \ldots, e_{n} \in E$ and scalars $r_{1}, \ldots, r_{n} \in[0,1]$ with $\sum_{i} r_{i} \leq 1$ the sum $r_{1} \cdot e_{1} \boxtimes \cdots \otimes r_{n} \cdot e_{n}$ exists in $E$
1 Use induction on $n \geq 1$, and check the base step.
2 For the induction step let $e_{1}, \ldots, e_{n+1}$ and $r_{1}, \ldots, r_{n+1} \in[0,1]$ with $\sum_{i \leq n+1} r_{i} \leq 1$ be given. By induction hypothesis the sum $\mathbb{Q}_{i \leq n} r_{i} \cdot e_{i}$ exists. Use Exercise 4.2 .16 to check that the following chain of inequalities holds and suffices to prove that the sum $\bigotimes_{i \leq n+1} r_{i} \cdot e_{i}$ exists too.

$$
\begin{aligned}
r_{n+1} \cdot e_{n+1} \leq r_{n+1} \cdot 1 & \leq\left(\left(\sum_{i \leq n} r_{i}\right) \cdot 1\right)^{\perp} \\
& =\left(\bigotimes_{i \leq n} r_{i} \cdot 1\right)^{\perp} \leq\left(\bigotimes_{i \leq n} r_{i} \cdot e_{i}\right)^{\perp} .
\end{aligned}
$$

4.2.19 Consider validity as a function $V: \operatorname{Pred}(X) \rightarrow \operatorname{Pred}(\mathcal{D}(X))$, given by $V(p)(\omega):=\omega \vDash p$. Show that $V$ is a map of effect modules.

### 4.3 Transformation of observables

One of the basic operations that we have seen so far is state transformation $\gg$. It is used to transform a state / distribution $\omega$ on the domain $X$ of a channel $c: X \leadsto Y$ into a state $c »=\omega$ on the codomain $Y$ of the channel. This section introduces transformation of observables $=\ll$. It works in the opposite direction of the channel: an observable $q$ on the codomain $Y$ is transformed into an observable $c=\ll q$ on the domain $X$. Thus, state transformation works forwardly, in the direction of the channel, whereas observable transformation $=\lll$ works backwardly, against the direction of the channel. This operation $=<$ 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 $=\ll$ and lists its key mathematical properties. In the next chapter it will be used for probabilistic reasoning, especially in combination with updating.

Definition 4.3.1. Let $c: X \rightsquigarrow Y$ be a channel. An observable $q \in \operatorname{Obs}(Y)$ is transformed into $c=\ll q \in \operatorname{Obs}(X)$ via the definition:

$$
\begin{equation*}
(c=\ll q)(x):=c(x) \vDash q=\sum_{y \in Y} c(x)(y) \cdot q(y) . \tag{4.10}
\end{equation*}
$$

When $q$ is a point predicate $\mathbf{1}_{y}$, for an element $y \in Y$, we get:

$$
\left(c=\ll \mathbf{1}_{y}\right)(x)=c(x)(y) \quad \text { so that } \quad c=\ll \mathbf{1}_{y}=c(-)(y): X \rightarrow[0,1] .
$$

The resulting function $Y \rightarrow \operatorname{Pred}(X)$, given by $y \mapsto c=\ll \mathbf{1}_{y}$, is sometimes called the likelihood function.

There is a whole series of basic facts about $=\ll$.

## Lemma 4.3.2.

1 The operation $c=\ll(-): \operatorname{Obs}(Y) \rightarrow \operatorname{Obs}(X)$ of transforming observables along a channel $c: X \leadsto Y$ restricts, first to factors $c=\ll(-): \operatorname{Fact}(Y) \rightarrow$ $\operatorname{Fact}(X)$, and then to fuzzy predicates $c=\ll(-): \operatorname{Pred}(Y) \rightarrow \operatorname{Pred}(X)$, but not to sharp predicates.
2 Observable transformation $c=\ll(-)$ 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 $\left(c=\ll q_{1}\right) \leq\left(c=\ll q_{2}\right)$.
5 Transformation along the unit channel is the identity: unit $=\ll q=q$.
6 Transformation along a composite channel is successive transformation: $(d \odot c)=\ll q=c=\ll(d=\ll q)$.
7 Transformation along a tuple transforms parallel conjunction $\otimes$ into sequential conjunction $\&:\langle c, d\rangle=\ll(p \otimes q)=(c=\ll p) \&(d=\ll q)$.
8 Transformation along parallel channels preserves parallel conjunctions: $(e \otimes f)=\ll(p \otimes q)=(e=\ll p) \otimes(f=\ll q)$.
9 Transformation along a trivial, deterministic channel $f$ is pre-composition: $f=\ll q=q \circ f$.

Proof. 1 If $q \in \operatorname{Fact}(Y)$ then $q(y) \geq 0$ for all $y$. But then also $(c=\ll q)(x)=$ $\sum_{y} c(x)(y) \cdot q(y) \geq 0$, so that $c=\ll q \in \operatorname{Fact}(X)$. If in addition $q \in \operatorname{Pred}(Y)$, so that $q(y) \leq 1$ for all $y$, then also $(c=\ll q)(x)=\sum_{y} c(x)(y) \cdot q(y) \leq \sum_{y} c(x)(y)=$ 1 , since $c(x) \in \mathcal{D}(Y)$, so that $c=<q \in \operatorname{Pred}(X)$.

The fact that a transformation $c=\lll$ of a sharp predicate $p$ need not be sharp is demonstrated in Exercise 4.3.2.
2 Easy.
$3(c=\mathbb{1})(x)=\sum_{y} c(x)(y) \cdot \mathbf{1}(y)=\sum_{y} c(x)(y) \cdot 1=1$. The fact that \& is not preserved follows from Exercise 4.3.1
4 Easy.
5 Recall that unit $(x)=1|x\rangle$, so that $($ unit $=\ll q)(x)=\sum_{y}$ unit $(x)(y) \cdot q(y)=q(x)$.
6 For $c: X \leadsto Y$ and $d: Y \rightsquigarrow Z$ and $q \in O b s(Z)$ we have:

$$
\begin{aligned}
((d \odot c)=\ll q)(x) & =\sum_{z \in Z}(d \odot c)(x)(z) \cdot q(z) \\
& =\sum_{z \in Z}\left(\sum_{y \in Y} c(x)(y) \cdot d(y)(z)\right) \cdot q(z) \\
& =\sum_{y \in Y} c(x)(y) \cdot\left(\sum_{z \in Z} d(y)(z) \cdot q(z)\right) \\
& =\sum_{y \in Y} c(x)(y) \cdot(d=\ll q)(y) \\
& =(c=\ll(d=\ll q))(x) .
\end{aligned}
$$

7 Let $c: X \leadsto Y, d: X \leadsto Z$, with $p \in \operatorname{Pred}(Y)$ and $q \in \operatorname{Pred}(Z)$. Then for $x \in X$,

$$
\begin{aligned}
(\langle c, d\rangle=\ll(p \otimes q))(x) & =\sum_{y \in Y, z \in Z}\langle c, d\rangle(x)(y, z) \cdot(p \otimes q)(y, z) \\
& =\sum_{y \in Y, z \in Z} c(x)(y) \cdot d(x)(z) \cdot p(y) \cdot q(z) \\
& =\left(\sum_{y \in Y} c(x)(y) \cdot p(y)\right) \cdot\left(\sum_{z \in Z} d(x)(z) \cdot q(z)\right) \\
& =(c \lll p)(x) \cdot(d=\ll q)(x) \\
& =((c=\ll p) \&(d=\ll q))(x) .
\end{aligned}
$$

8 Similarly,

$$
\begin{aligned}
((e \otimes f)=\ll(p \otimes q))(x, y) & =\sum_{u, v}(e \otimes f)(x, y)(u, v) \cdot(p \otimes q)(u, v) \\
& =\sum_{u, v} e(x)(u) \cdot f(y)(v) \cdot p(u) \cdot q(v) \\
& =\left(\sum_{u} e(x)(u) \cdot p(u)\right) \cdot\left(\sum_{v} f(y)(v) \cdot q(v)\right) \\
& =(e=\ll p)(x) \cdot(f=\ll q)(y) \\
& =((e \lll p)(x) \otimes(f=\ll q))(x, y) .
\end{aligned}
$$

9 For a function $f: X \rightarrow Y$, recall that we write $\langle f\rangle:=$ unit $\circ f: X \rightarrow Y$
if we wish to make explicit that we use the function $f$ as a 'deterministic' channel. Then:

$$
(\langle f\rangle=\ll q)(x)=\sum_{y \in Y} \operatorname{unit}(f(x))(y) \cdot q(y)=q(f(x))=(q \circ f)(x) .
$$

There is the following fundamental relationship between transformations $»=$, $=\ll$ and validity $\mid=$.

Proposition 4.3.3. Let $c: X \leadsto Y$ be a channel with a state $\omega \in \mathcal{D}(X)$ on its domain and an observable $q \in \operatorname{Obs}(Y)$ on its codomain. Then:

$$
\begin{equation*}
c \gg \omega \models q=\omega \vDash c=\ll q . \tag{4.11}
\end{equation*}
$$

This equation is essentially the law of total expectation, see Exercise 4.2.7
Proof. The result follows simply by unpacking the relevant definitions:

$$
\begin{aligned}
c \gg=\omega \vDash q=\sum_{y \in Y}(c \gg=\omega)(y) \cdot q(y) & =\sum_{y \in Y}\left(\sum_{x \in X} c(x)(y) \cdot \omega(x)\right) \cdot q(y) \\
& =\sum_{x \in X} \omega(x) \cdot\left(\sum_{y \in Y} c(x)(y) \cdot q(y)\right) \\
& =\sum_{x \in X} \omega(x) \cdot(c=\ll q)(x) \\
& =\omega \vDash c=\ll q .
\end{aligned}
$$

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 (4.7). We can now see this as an instance of 4.11, using a projection $\pi_{i}: X_{1} \times \cdots \times X_{n} \rightarrow X_{i}$ as (trivial) channel, in:

$$
\begin{equation*}
\pi_{i} \gg \omega \vDash p=\omega \vDash \pi_{i}=\ll p \tag{4.12}
\end{equation*}
$$

The left-hand side of 4.12) uses the $i$-th marginal of state $\omega \in \mathcal{D}\left(X_{1} \times \cdots \times\right.$ $X_{n}$ ), namelY:

$$
\pi_{i} \gg=\omega=\omega[0, \ldots, 0,1,0, \ldots, 0] \in \mathcal{D}\left(X_{i}\right)
$$

On the right-hand side of (4.12) the observable $p \in \operatorname{Obs}\left(X_{i}\right)$ is weakenend to the following predicate.

$$
\pi_{i}=\ll p=\mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes p \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1} \in \operatorname{Pred}\left(X_{1} \times \cdots \times X_{n}\right) .
$$

- The first equation in Exercise 4.1.5 is also an instance of 4.11, namely for a trivial channel $f: X \mapsto Y$ given by a function $f: X \rightarrow Y$, as in:

$$
\begin{aligned}
f \gg=\omega \vDash p \stackrel{[4.11]}{=} \omega & =f=\ll p \\
& =\omega \vDash q \circ f \quad \text { by Lemma 4.3.2 (9). }
\end{aligned}
$$

Remark 4.3.4. In a programming context, where a channel $c: X \mapsto Y$ is seen as a program taking inputs from $X$ to outputs in $Y$, one may call $c=\ll q$ the weakest precondition of $q$, commonly written as $w p(c, q)$, see e.g. [47, 116 131]. We briefly explain this view.

A precondition of $q$, w.r.t. channel $c: X \leadsto Y$, may be defined as an observable $p$ on the channel's domain $X$ for which:

$$
\omega \vDash p \leq c \gg=\omega \vDash q, \quad \text { for all states } \omega \text {. }
$$

Proposition 4.3.3 tells that $c=\ll q$ 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:

$$
\begin{aligned}
p(x) & =\operatorname{unit}(x) \vDash p \\
& \leq c »=\operatorname{unit}(x) \vDash q=c(x) \vDash q=(c \approx \ll q)(x) .
\end{aligned}
$$

As a result, $p \leq c=\ll q$.
Lemma 4.3.2 (6) expresses a familiar compositionality property in the theory of weakest preconditions:

$$
w p(d \odot c, q)=(d \odot c)=\ll q=c=\ll(d=\ll q)=w p(c, w p(d, q)) .
$$

We close this section with two topics that dig deeper into the nature of transformations. First, we relate transformation of states and observables in terms of matrix operations. Then we look closer at the categorical aspects of transformation of observables.

Remark 4.3.5. 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 $\boldsymbol{n} \leadsto \boldsymbol{m}$. 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:

$$
M_{c}=\left(\begin{array}{ccc}
c(1)(1) & \cdots & c(n-1)(1) \\
\vdots & & \vdots \\
c(1)(m-1) & \cdots & c(n-1)(m-1)
\end{array}\right) .
$$

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 $M_{\omega}$ of length $n$, as
on the left below. It is then easy to see that the matrix $M_{c \gg=\omega}$ of the transformed state, is obtained by matrix-column multiplication, as on the right:

$$
M_{\omega}=\left(\begin{array}{c}
\omega(0) \\
\vdots \\
\omega(n-1)
\end{array}\right) \quad \text { so that } \quad M_{c »>\omega}=M_{c} \cdot M_{\omega} .
$$

Indeed,

$$
(c \gg=\omega)(j)=\sum_{i} c(i)(j) \cdot \omega(i)=\sum_{i}\left(M_{c}\right)_{j i} \cdot\left(M_{\omega}\right)_{i}=\left(M_{c} \cdot M_{\omega}\right)_{j} .
$$

An observable $q: \boldsymbol{m} \rightarrow \mathbb{R}$ on $\boldsymbol{m}$ can be identified with a row vector $M_{q}=$ $(q(0) \cdots q(m-1))$. Transformation $c=\varangle q$ then corresponds to row-matrix multiplication:

$$
M_{c \approx K q}=M_{q} \cdot M_{c} .
$$

Again, this is checked easily:

$$
(c=\ll q)(i)=\sum_{j} q(j) \cdot c(i)(j)=\sum_{j}\left(M_{q}\right)_{j} \cdot\left(M_{c}\right)_{j, i}=\left(M_{q} \cdot M_{c}\right)_{i} .
$$

We close this section by making the functoriality of observable transformation $=\ll$ explicit, in the style of Proposition 4.2.13. The latter deals with functions, but we now consider functoriality with respect to channels, using the category Chan $=\operatorname{Chan}(\mathcal{D})$ of probabilistic channels. Notice that the case of sharp predicates is omitted from Proposition 4.2.13, simply because sharp predicates are not closed under predicate transformation (see Exercise 4.3.2). Also, conjunction \& is not preserved under transformation, see Exercise 4.3.1 below.

Proposition 4.3.6. Taking particular observables on a set is functorial, namely, via functors:

1 Obs: Chan $\rightarrow$ Vect $^{\text {op }}$;
2 Fact: Chan $\rightarrow$ Cone $^{\text {op }}$;
3 Pred: Chan $\rightarrow$ EMod $^{\text {op }}$;
On a channel $c: X \mapsto Y$, all these functors are given by transformation $c=\ll$ $(-)$, acting in the opposite direction.

Proof. Essentially, all necessary ingredients are already in Lemma 4.3.2 transformation restricts appropriately (item (1)), transformation preserves identies (item (5)) and composition (item (6)), and the relevant structure (items (2) and (3).

## Exercises

4.3.1 Consider the channel $f:\{a, b, c\} \nrightarrow\{u, v\}$ from Example 1.10.2, given by:

$$
f(a)=\frac{1}{2}|u\rangle+\frac{1}{2}|v\rangle \quad f(b)=1|u\rangle \quad f(c)=\frac{3}{4}|u\rangle+\frac{1}{4}|v\rangle .
$$

Take as predicates $p, q:\{u, v\} \rightarrow[0,1]$,

$$
p(u)=\frac{1}{2} \quad p(v)=\frac{2}{3} \quad q(u)=\frac{1}{4} \quad q(v)=\frac{1}{6} .
$$

Alternatively, in the normal-form notation of Lemma 4.2.3(2),

$$
p=\frac{1}{2} \cdot \mathbf{1}_{u}+\frac{2}{3} \cdot \mathbf{1}_{v} \quad q=\frac{1}{4} \cdot \mathbf{1}_{u}+\frac{1}{6} \cdot \mathbf{1}_{v}
$$

Compute:

- $f=\ll p$
- $f=\mathbb{R} q$
- $f=\ll(p \otimes q)$
- $(f=\ll p) \otimes(f=\ll q)$
- $f=\ll(p \& q)$
- $(f=\ll p) \&(f=\ll q)$

This will show that predicate transformation $=\ll$ does not preserve conjunction \&
4.3.2 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 $f=\ll \mathbf{1}_{u}$ on $\{a, b, c\}$ is not sharp. This proves that sharp predicates are not closed under predicate transformation.
2 Let $h: X \rightarrow Y$ be a function, considered as deterministic channel, and let $V \subseteq Y$ be an arbitrary subset (event). Check that:

$$
h=\ll \mathbf{1}_{V}=\mathbf{1}_{h^{-1}(V)}
$$

Conclude that predicate transformation along deterministic channels does preserve sharpness.
4.3.3 In the setting of Exercise 2.1.11 let $X$ be a finite set with $N$ elements. We have a function size $:=\|-\|: \mathcal{N}(X) \rightarrow \mathbb{N}$, and in the other direction a channel $\operatorname{size}^{\dagger}: \mathbb{N} \rightarrow \mathcal{D}(\mathcal{N}(X))$ given by:

$$
\operatorname{size}^{\dagger}(k):=\sum_{\varphi \in \mathcal{N}[k](X)} \frac{(\varphi)}{N^{k}}|\varphi\rangle \in \mathcal{D}(\mathcal{N}(X)) \hookrightarrow \mathcal{D}_{\infty}(\mathcal{N}(X))
$$

This is well-defined by Exercise 1.7.7.
1 Show that size ${ }^{\dagger} \gg=\operatorname{pois}[\lambda]=$ mpois $[\lambda]$, see Examples 2.1.7 (1) and (2).

2 Show that for $p \in \operatorname{Pred}(\mathbb{N})$ and $q \in \operatorname{Pred}(\mathcal{N}(X))$, there is an adjointness equation:

$$
\operatorname{pois}[\lambda] \vDash p \&\left(\operatorname{size}^{\dagger}=\ll q\right)=\operatorname{mpois}[\lambda] \vDash(\text { size }=\ll p) \& q .
$$

3 Draw the corresponding equation between string diagrams.
4.3.4 Let $h: X \rightarrow Y$ be an ordinary function. Recall from Lemma 4.3.2 (9) that $h=\varangle q=q \circ h$, when $h$ is considered as a deterministic channel. Show that transformation along such deterministic channels does preserve conjunctions:

$$
h=\ll\left(q_{1} \& q_{2}\right)=\left(h=\ll q_{1}\right) \&\left(h=\ll q_{2}\right),
$$

in contrast to the findings in Exercise 4.3 .1 for arbitrary channels.
Conclude that weakening preserves conjunction: $\mathbf{1} \otimes(p \& q)=$ $(\mathbf{1} \otimes p) \&(\mathbf{1} \otimes q)$.
4.3.5 Let predicate $q_{i} \in \operatorname{Pred}(Y)$ form a test, see Definition 4.2.2, and let $c: X \leadsto Y$ be channel. Check that the transformed predicates $c=\ll q_{i}$ form a test on $X$.
4.3.6 1 Check that a predicate $p: X \rightarrow[0,1]$ can be identified with a channel $\widehat{p}:=$ flip $\circ p: X \rightarrow \mathbf{2}$, see also Exercise 4.2.13. Describe this $\widehat{p}$ explicitly in terms of $p$.
2 Define a channel orth: $\mathbf{2} \leftrightarrows \mathbf{2}$ such that orth $\odot \widehat{p}=\widehat{p^{\perp}}$.
3 Define also a channel conj: $\mathbf{2} \times \mathbf{2} \leadsto \mathbf{2}$ such that conj $\odot\langle\widehat{p}, \widehat{q}\rangle=$ $\widehat{p \& q}$.
4 Finally, define also a channel scal $(r): \mathbf{2} \rightsquigarrow 2$, for $r \in[0,1]$, so that $\operatorname{scal}(r) \odot \widehat{p}=\widehat{r \cdot p}$.
4.3.7 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 $p: X \rightarrow[0,1]$ can be identified with a channel $X \multimap \mathbf{2}$, see Exercise 4.3.6. Check that under these identifications validity $\omega \vDash p$ can be identified with:

- state transformation $p \gg \omega$;
- predicate transformation $\omega=\ll p$;
- channel composition $p \odot \omega$.
4.3.8 This exercises shows how parallel conjunction $\otimes$ and sequential conjunction \& are inter-definable via predicate transformation $=\ll$, using projection channels $\pi_{i}$ and copy channels $\Delta$, 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 $X_{1} \times X_{2}$,

$$
p_{1} \otimes p_{2}=\left(\pi_{1}=\ll p_{1}\right) \&\left(\pi_{2}=\ll p_{2}\right)=\left(p_{1} \otimes \mathbf{1}\right) \&\left(\mathbf{1} \otimes p_{2}\right)
$$

The last equation occurred already in Lemma 4.2.10 (2).
2 Let $q_{1}, q_{2}$ be observables on the same set $Y$. Prove that on $Y$,

$$
q_{1} \& q_{2}=\Delta=\ll\left(q_{1} \otimes q_{2}\right) .
$$

4.3.9 Let $c: X \leadsto Y$ be a channel, with observables $p$ on $Z$ and $q$ on $Y \times Z$. Check that:

$$
(c \otimes i d)=\ll((\mathbf{1} \otimes p) \& q)=(\mathbf{1} \otimes p) \&((c \otimes i d)=\ll q)
$$

(Recall that \& is not preserved by $=\ll$, see Exercise 4.3.1)
4.3.10 We take a closer look at the functor

$$
\text { Chan } \xrightarrow{\text { Pred }} \text { EMod }^{\text {op }}
$$

from Proposition 4.2.13 It sends a map $c: X \leadsto Y$ in Chan to the function $\operatorname{Pred}(c): \operatorname{Pred}(Y) \rightarrow \operatorname{Pred}(X)$ via:

$$
\operatorname{Pred}(c)(q):=c=\ll q .
$$

1 Check in detail that $\operatorname{Pred}(c)$ is a morphism of effect modules, see Definition 4.2.1 (3).
2 Show that the functor Pred is faithful, in the sense that $\operatorname{Pred}(c)=$ $\operatorname{Pred}\left(c^{\prime}\right)$ implies $c=c^{\prime}$, for channels $c, c^{\prime}: X \mapsto Y$.
3 Let $Y$ be a finite set. Show that for each map $h: \operatorname{Pred}(Y) \rightarrow \operatorname{Pred}(X)$ in the category EMod there is a unique channel $c: X \leadsto Y$ with $\operatorname{Pred}(c)=h$.
Hint: Write a predicate $p$ as finite sum $\bigotimes_{y} p(y) \cdot \mathbf{1}_{y}$, i.e. as the normal form of Lemma 4.2.3 (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 [46, [47, 131], or [69] for an abstract account.

### 4.4 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.

In Example 4.1.4 (4) we have seen the mean of a binomial distribution. A multinomial $m n[K](\omega)$ 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 4.1.3 does not apply: the space is not a subset of the reals. Still, we have described in Proposition 3.3.6 as a generalised mean for multinomials. 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.

Proposition 4.4.1. We fix a set $X$ and consider means in $\mathcal{M}(X)$.
1 For a distribution $\omega \in \mathcal{D}(X)$,

$$
\operatorname{mean}(m n[K](\omega)):=\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi=K \cdot \omega \in \mathcal{M}(X) .
$$

2 For a non-empty urn $v \in \mathcal{N}[L](X)$ of size $L \geq K$,

$$
\operatorname{mean}(h g[K](v)):=\sum_{\varphi \in \leq K v} h g[K](v)(\varphi) \cdot \varphi=K \cdot \operatorname{Flrn}(v) \in \mathcal{M}(X) .
$$

3 For a non-empty urn $v \in \mathcal{N}(X)$,

$$
\operatorname{mean}(p l[K](v)):=\sum_{\varphi \in \mathcal{N}[K](\operatorname{supp}(v))} p l[K](v)(\varphi) \cdot \varphi=K \cdot \operatorname{Flrn}(v) \in \mathcal{M}(X) .
$$

4 For a distribution $\omega \in \mathcal{D}(X)$ and a rate $\lambda \in \mathbb{R}_{>0}$,

$$
\operatorname{mean}(\operatorname{Pmn}[\lambda](\omega)):=\sum_{\varphi \in \mathcal{N}(X)} \operatorname{Pmn}[\lambda](\omega)(\varphi) \cdot \varphi=\lambda \cdot \omega \in \mathcal{M}(X)
$$

The Poisson-iid distribution $\operatorname{Piid}[\lambda](\omega) \in \mathcal{D}_{\infty}(\mathcal{L}(X))$ 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 3.3.6 together with Lemmas 3.4.5 (2) and 3.5.1 (2). Hence we concentrate on the last item,
involving the Poission multinomial.

$$
\begin{aligned}
\operatorname{mean}(\operatorname{Pmn}[\lambda](\omega))(x) & =\sum_{\varphi \in \mathcal{N}(X)} \operatorname{Pmn}[\lambda](\omega)(\varphi) \cdot \varphi(x) \\
& \stackrel{\sqrt{3.52}}{=} \sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{pois}[\lambda](K) \cdot \operatorname{mn}[K](\omega)(\varphi) \cdot \varphi(x) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \varphi(x) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot \operatorname{mean}(\operatorname{mn}[K](\omega))(x) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot K \cdot \omega(x) \\
& =\operatorname{mean}(\operatorname{pois}[\lambda]) \cdot \omega(x) \\
& =\lambda \cdot \omega(x), \quad \text { by Exercise } 4.1 .8 .
\end{aligned}
$$

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:

$$
\begin{equation*}
\mathcal{M}(X) \xrightarrow{e v_{x}} \mathbb{R} \quad \text { by } \quad \quad \mathrm{ev}_{x}(\varphi):=\varphi(x) \tag{4.13}
\end{equation*}
$$

Then we can compute the validity of this observable as:

$$
\begin{aligned}
\operatorname{mn}[K](\omega) \vDash \mathrm{ev}_{x} & \stackrel{\boxed{4.2]}}{=} \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \mathrm{ev}_{x}(\varphi) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \varphi(x) \\
& =K \cdot \omega(x), \quad \text { by Lemma 3.3.2 }
\end{aligned}
$$

Hence we can alternatively write:

$$
\operatorname{mean}(\operatorname{mn}[K](\omega))=\sum_{x \in X}\left(\operatorname{mn}[K](\omega) \models \mathrm{e} v_{x}\right)|x\rangle=K \cdot \omega .
$$

The same can be done for the other draw distributions in Proposition 4.4.1.
Example 4.4.2. 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 1 six: $\mathcal{N}(p i p s) \rightarrow\{0,1\}$ for the sharp predicate that sends a multiset $\varphi$, over the dice space 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 $\varphi$. We can thus model the first option (DM1) as validity:

$$
\begin{equation*}
\operatorname{mn}[4](\text { dice }) \vDash 1 \text { six } . \tag{DM1}
\end{equation*}
$$

The second option (DM2) then becomes:

$$
\begin{equation*}
m n[24](\text { dice } \otimes \text { dice }) \vDash 2 \text { six }, \tag{DM2}
\end{equation*}
$$

where 2 six: $\mathcal{N}($ dice $\times$ dice $) \rightarrow\{0,1\}$ is the obvious predicate with $2 \operatorname{six}(\psi)=1$ 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:

$$
\begin{aligned}
m n[4](\text { dice }) \vDash 1 \text { six } & =\text { acc } »=\text { iid }[4](\text { dice }) \vDash 1 \text { six } \quad \text { by Theorem } 2.6 .7 \\
& =\text { iid }[4](\text { dice }) \vDash \text { acc }=\ll 1 \text { six } \\
& =\text { iid }[4](\text { dice }) \vDash\left(\text { acc } \lll 1 s_{i x}^{\perp}\right)^{\perp} \\
& =\text { dice } \otimes \text { dice } \otimes \text { dice } \otimes \text { dice } \models\left(\mathbf{1}_{6}^{\perp} \otimes \mathbf{1}_{6}^{\perp} \otimes \mathbf{1}_{6}^{\perp} \otimes \mathbf{1}_{6}^{\perp}\right)^{\perp} \\
& =1-\left(\text { dice } \models \mathbf{1}_{6}^{\perp}\right)^{4} \quad \text { by Lemmas } 4.2 .6 \text { and } 4.2 .9 \\
& =1-\left(\frac{5}{6}\right)^{4} \\
& =\frac{671}{1296} \\
& \approx 0.518 .
\end{aligned}
$$

Similarly one can compute (DM2) as:

$$
\begin{aligned}
m n[24](\text { dice } \otimes \text { dice }) \vDash 2 \text { six } & =(\text { dice } \otimes \text { dice })^{24} \vDash\left(\left(\left(\mathbf{1}_{6} \otimes \mathbf{1}_{6}\right)^{\perp}\right)^{24}\right)^{\perp} \\
& =1-\left(\text { dice } \otimes \text { dice } \vDash\left(\mathbf{1}_{6} \otimes \mathbf{1}_{6}\right)^{\perp}\right)^{24} \\
& =1-\left(\frac{35}{36}\right)^{24} \\
& \approx 0.491 .
\end{aligned}
$$

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 $X \rightarrow \mathbb{R}$ 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 4.4.3. Let $p: X \rightarrow \mathbb{R}$ be an observable on a set $X$.
1 The additive extension $\bar{p}^{+}: \mathcal{N}(X) \rightarrow \mathbb{R}$ of $p$ on multisets over $X$ is defined as:

$$
\bar{p}^{+}(\varphi)=\sum_{x \in \operatorname{supp}(\varphi)} \varphi(x) \cdot p(x)=\sum_{x \in X} \varphi(x) \cdot p(x)
$$

2 The multiplicative extension $\bar{p}^{\bullet}: \mathcal{N}(X) \rightarrow \mathbb{R}$ of $p$ is:

$$
\bar{p}^{\bullet}(\varphi)=\prod_{x \in \operatorname{supp}(\varphi)} p(x)^{\varphi(x)}=\prod_{x \in X} p(x)^{\varphi(x)} .
$$

By construction, these extensions are homomorphisms of monoids, so that:

$$
\left\{\begin{array} { r l } 
{ \overline { p } ^ { + } ( \mathbf { 0 } ) } & { = 0 }  \tag{4.14}\\
{ \overline { p } ^ { \bullet } ( \mathbf { 0 } ) } & { = 1 . }
\end{array} \quad \text { and } \quad \left\{\begin{array}{rl}
\bar{p}^{+}(\varphi+\psi) & =\bar{p}^{+}(\varphi)+\bar{p}^{+}(\psi) \\
\bar{p}^{\bullet}(\varphi+\psi) & =\bar{p}^{\bullet}(\varphi) \cdot \bar{p}^{\bullet}(\psi)
\end{array}\right.\right.
$$

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 4.4.4. Consider a random variable given by an observable $p: X \rightarrow$ $\mathbb{R}$ and a state $\omega \in \mathcal{D}(X)$.

1 For $K \in \mathbb{N}$, the additive extension $\bar{p}^{+}$of p forms new random variable, with the multinomial distribution $m n[K](\omega)$ on $\mathcal{N}[K](X)$ as state. The associated validity is:

$$
m n[K](\omega) \vDash \bar{p}^{+}=K \cdot(\omega \vDash p) .
$$

2 The multiplicative extension gives:

$$
m n[K](\omega) \vDash \bar{p}^{\bullet}=(\omega \vDash p)^{K} .
$$

3 For $\lambda \in \mathbb{R}_{>0}$, the validity of the additive extension in a Poisson multinomial is:

$$
\operatorname{Pmn}[\lambda](\omega) \vDash \bar{p}^{+}=\lambda \cdot(\omega \vDash p) .
$$

4 The validity of the multiplicative extension can be expressed as:

$$
\operatorname{Pmn}[\lambda](\omega) \vDash \bar{p}^{\bullet}=e^{-\lambda \cdot\left(\omega \vDash p^{\perp}\right)} .
$$

Proof. 1 We use Lemma 3.3.2 in:

$$
\begin{aligned}
\operatorname{mn}[K](\omega) \vDash \bar{p}^{+} & =\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \bar{p}^{+}(\varphi) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot\left(\sum_{x \in X} \varphi(x) \cdot p(x)\right) \\
& =\sum_{x \in X} p(x) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi(x) \\
& =\sum_{x \in X} p(x) \cdot K \cdot \omega(x) \\
& =K \cdot(\omega \vDash p) .
\end{aligned}
$$

2 By unfolding the multinomial distribution, see 2.40 :

$$
\begin{aligned}
\operatorname{mn}[K](\omega) \vDash \bar{p}^{\bullet} & =\sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \bar{p}^{\bullet}(\varphi) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)}(\varphi) \cdot\left(\prod_{x} \omega(x)^{\varphi(x)}\right) \cdot\left(\prod_{x} p(x)^{\varphi(x)}\right) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)}(\varphi) \cdot \prod_{x}(\omega(x) \cdot p(x))^{\varphi(x)} \\
& \stackrel{\boxed{11.39}}{=}\left(\sum_{x} \omega(x) \cdot p(x)\right)^{K} \\
& =(\omega \vDash p)^{K} .
\end{aligned}
$$

3 The validity of the additive extension $\bar{p}^{+}$in a Poisson point process state is computed as follows, using the first item for multinomials.

$$
\begin{aligned}
\operatorname{Pmn}[\lambda](\omega) \vDash \bar{p}^{+} & =\sum_{K \in \mathbb{N}} \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{pois}[\lambda](K) \cdot \operatorname{mn}[K](\omega)(\varphi) \cdot \bar{p}^{+}(\varphi) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot\left(m n[K](\omega) \vDash \bar{p}^{+}\right) \\
& =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot K \cdot(\omega \vDash p) \\
& =\lambda \cdot(\omega \vDash p) \quad \text { by Exercise } 4.1 .8
\end{aligned}
$$

4 In the multiplicative case we get:

$$
\begin{aligned}
\operatorname{Pmn}[\lambda](\omega) \vDash \bar{p}^{\bullet} & =\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot\left(\operatorname{mn}[K](\omega) \vDash \bar{p}^{\bullet}\right) \\
& \stackrel{\text { 22 }}{=} \sum_{K \in \mathbb{N}} e^{-\lambda} \cdot \frac{\lambda^{K}}{K!} \cdot(\omega \vDash p)^{K} \\
& =\frac{e^{-\lambda}}{e^{-\lambda \cdot(\omega \vDash p)}} \cdot \sum_{K \in \mathbb{N}} e^{-\lambda \cdot(\omega \vDash p)} \cdot \frac{(\lambda \cdot(\omega \vDash p))^{K}}{K!} \\
& =e^{-\lambda \cdot(1-\omega \vDash p)} \\
& =e^{-\lambda \cdot\left(\omega \vDash p^{\perp}\right)} .
\end{aligned}
$$

Remark 4.4.5. A validity $\omega \vDash p$ can be approximated via sampling. This may be useful in situations where the distribution $\omega$ has very large support, so that computing the sum $\sum_{x \in \operatorname{supp}(\omega)} \omega(x) \cdot p(x)=\omega \vDash p$ 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.

$$
\begin{align*}
& \mathrm{v}:=0 \\
& \text { repeat } K \text { times } \\
& \mathrm{x} \leftarrow \omega  \tag{4.15}\\
& \mathrm{v}:=\mathrm{v}+p(\mathrm{x}) \\
& \text { return } v / K
\end{align*}
$$

The justification for this approach is given by Proposition 4.4.4 (1). It takes the probabilities of draws of multisets $\varphi$ into accounts and computes $\bar{p}^{+}(\varphi)=$ $\sum_{x} \varphi(x) \cdot p(x)$ as in the above repeat loop. If we do this for all draws $\varphi$, divided by $K$, with their multinomial probabilities, we get:

$$
\begin{aligned}
\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot\left(\frac{\sum_{x} \varphi(x) \cdot p(x)}{K}\right) & =\frac{1}{K} \cdot\left(m n[K](\omega) \vDash \bar{p}^{+}(\varphi)\right) \\
& =\omega \vDash p .
\end{aligned}
$$

For the hypergeometric and Pólya distributions we have results for the additive extension. They resemble the formulation in Proposition 4.4.4 (1) for multinomials.

Proposition 4.4.6. Consider an observable $p$ and an urn $v$.

$$
h g[K](v) \vDash \bar{p}^{+}=K \cdot(\operatorname{Flrn}(v) \vDash p)=p l[K](v) \vDash \bar{p}^{+} .
$$

Proof. Both equations are obtained as for Proposition 4.4.4 11, this time using Lemmas 3.4.5 (1) and 3.5.1 (1).

For the parallel multinomial law $\mathrm{pml}: \mathcal{M}[K](\mathcal{D}(X)) \rightarrow \mathcal{D}(\mathcal{M}[K](X))$ from Section 3.6 there is a similar result, in the multiplicative case.

Proposition 4.4.7. For an observable $p: X \rightarrow \mathbb{R}$ and for a multiset of distributions $\sum_{i} n_{i}\left|\omega_{i}\right\rangle \in \mathcal{M}[K](\mathcal{D}(X))$,

$$
\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right) \vDash \bar{p}^{\bullet}=\prod_{i}\left(\omega_{i} \vDash p\right)^{n_{i}} .
$$

Proof. We use the second formulation (3.33) of pml in:

$$
\begin{aligned}
\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right) \vDash \bar{p}^{\bullet} & =\sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)}\left(\prod_{i} m n\left[n_{i}\right]\left(\omega_{i}\right)\left(\varphi_{i}\right)\right) \cdot \bar{p}^{\bullet}\left(\sum_{i} \varphi_{i}\right) \\
& =\sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)}\left(\prod_{i} \operatorname{mn}\left[n_{i}\right]\left(\omega_{i}\right)\left(\varphi_{i}\right)\right) \cdot\left(\prod_{i} \bar{p}^{\bullet}\left(\varphi_{i}\right)\right) \\
& =\sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right] \backslash(X)} \prod_{i} m n\left[n_{i}\right]\left(\omega_{i}\right)\left(\varphi_{i}\right) \cdot \bar{p}^{\bullet}\left(\varphi_{i}\right) \\
& =\sum_{i} \operatorname{mn}\left[n_{i}\right]\left(\omega_{i}\right)\left(\varphi_{i}\right) \cdot \bar{p}^{\bullet}\left(\varphi_{i}\right) \\
& =\prod_{i} m n\left[n_{i}\right]\left(\omega_{i}\right) \vDash \bar{p}^{\bullet} \\
& =\prod_{i}\left(\omega_{i} \vDash p\right)^{n_{i}}, \quad \text { by Proposition4.4.4 (2). }
\end{aligned}
$$

For the following result we use the extension $\mathcal{N}[K]:$ Chan $\rightarrow$ Chan of the (natural) multiset functor to the category of probabilistic channels, see Corollary 3.7.8.

Corollary 4.4.8. Let $c: X \leadsto Y$ be a channel, with a factor $q: Y \rightarrow \mathbb{R}_{\geq 0}$ on its codomain. Then we have an equality of predicates on $\mathcal{N}[K](X)$ of the form:

$$
\mathcal{N}[K](c)=\ll \bar{q}^{\bullet}=\overline{c=\lll} \cdot
$$

Proof. For $\varphi \in \mathcal{N}[K](X)$,

$$
\begin{aligned}
\left(\mathcal{N}[K](c)=\ll \bar{q}^{\bullet}\right)(\varphi) & \stackrel{\frac{4.10}{=}}{ } \mathcal{N}[K](c)(\varphi) \vDash \bar{q}^{\bullet} \\
\frac{3.41]}{=} & p m l(\mathcal{N}(c)(\varphi)) \vDash \bar{q}^{\bullet} \\
& =p m l\left(\sum_{x \in X} \varphi(x)|c(x)\rangle\right) \vDash \bar{q}^{\bullet} \\
& =\sum_{x \in X}(c(x) \vDash q)^{\varphi(x)} \quad \text { by Proposition 4.4.7 } \\
& \stackrel{4.10}{=} \sum_{x \in X}(c=\ll q)(x)^{\varphi(x)} \\
& =\overline{c=\ll q} \cdot(\varphi) .
\end{aligned}
$$

## Exercises

4.4.1 Let $p: X \rightarrow \mathbb{R}$ be an observable. Show that:

$$
\begin{aligned}
\bar{p}^{+}\left(\operatorname{acc}\left(x_{1}, \ldots, x_{n}\right)\right) & =p\left(x_{1}\right)+\ldots+p\left(x_{n}\right) \\
\bar{p}^{\bullet}\left(\operatorname{acc}\left(x_{1}, \ldots, x_{n}\right)\right) & =p\left(x_{1}\right) \cdot \ldots \cdot p\left(x_{n}\right) .
\end{aligned}
$$

4.4.2 Show that the additive/multiplicative extension preserves the additive/multiplicative structure of observables:

$$
\overline{p+q}^{+}=\bar{p}^{+}+\bar{q}^{+} \quad \overline{\mathbf{0}}^{+}=\mathbf{0}
$$

and:

$$
\overline{p \& q}=\bar{p}^{\bullet} \& \bar{q}^{\bullet} \quad \overline{\mathbf{1}}^{\bullet}=\mathbf{1}
$$

4.4.3 Let $E \subseteq X$ be a subset (event), corresponding to the sharp indicator predicate $\mathbf{1}_{E}: X \rightarrow[0,1]$. Check that for an arbitrary multiset $\varphi \in$ $\mathcal{N}(X)$,

$$
\overline{\mathbf{1}_{E}^{\bullet}}(\varphi)=1 \Longleftrightarrow \operatorname{supp}(\varphi) \subseteq E
$$

Note that this also holds for $\overline{{\overline{\boldsymbol{1}_{E}}}^{\boldsymbol{\bullet}}} \mathbf{( \mathbf { 0 } )}=1$, since a product over an empty set equals one. Conclude that $\overline{\mathbf{1}_{E}}: \mathcal{N}(X) \rightarrow[0,1]$ is also a sharp predicate.
4.4.4 For a random variable $p: X \rightarrow \mathbb{R}$ with state $\omega \in \mathcal{D}(X)$ consider the validity $m n[-](\omega) \vDash \bar{p}^{+}$as an observable $\mathbb{N} \rightarrow \mathbb{R}$. Show that for a distribution $\sigma \in \mathcal{D}(\mathbb{N})$ one has:

$$
\sigma \vDash(\operatorname{mn}[-](\omega) \models \bar{p})=\operatorname{mean}(\sigma) \cdot(\omega \models p) .
$$

4.4.5 For a random variable $(\omega, p)$ write $\sum(\omega, p): \mathbb{N} \rightarrow \mathbb{R}$ for the summation observable defined by:

$$
\sum(\omega, p)(n):=\omega \vDash p+\cdots+p \quad(n \text { times }) .
$$

Show that for a distribution $\sigma \in \mathcal{D}(\mathbb{N})$ one has:

$$
\sigma \vDash \sum(\omega, p)=\operatorname{mean}(\sigma) \cdot(\omega \vDash p) .
$$

4.4.6 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 4.4.5 to show that the expected number if $\frac{49}{4}$.
2 Obtain this same outcome via Proposition 4.4.4 (1).
4.4.7 Let $p: X \rightarrow \mathbb{R}$ be an observable. In Definition 4.4.3 we have introduced $\bar{p}^{+}$and $\bar{p}^{\bullet}$ 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 $\bar{p}^{+}: \mathcal{L}(X) \rightarrow \mathbb{R}$ and $\bar{p}^{\bullet}: \mathcal{L}(X) \rightarrow \mathbb{R}$ for these extensions.

1 Describe $\bar{p}^{+}\left(\left[x_{1}, \ldots, x_{K}\right]\right)$ and $\bar{p}^{\bullet}\left(\left[x_{1}, \ldots, x_{K}\right]\right)$ concretely.
2 Prove that:

$$
\begin{aligned}
& \operatorname{iid}[K](\omega) \vDash \bar{p}^{+}=K \cdot(\omega \vDash p) \\
& \operatorname{iid}[K](\omega) \vDash \bar{p}^{\bullet}=(\omega \vDash p)^{K} .
\end{aligned}
$$

3 Show next that:

$$
\begin{aligned}
& \operatorname{Piid}[\lambda](\omega) \vDash \bar{p}^{+} \\
&=\lambda \cdot(\omega \vDash p) \\
& \operatorname{Piid}[\lambda](\omega) \vDash \bar{p}^{\bullet}=e^{-\lambda \cdot\left(\omega \vDash p^{\perp}\right)} .
\end{aligned}
$$

### 4.5 Validity-based distances

This section describes two standard distances, between states, and between predicates. It shows that these distances can both be formulated in terms of validity $\vDash$, in a dual form. Basic properties of these distances are included. Earlier, in Section 2.8, we have have seen Kullback-Leibler divergence as a measure of difference between states. But divergence does not form a metric since it is not symmetric, see Remark 4.5 .6 below for an illustration of the difference.

The distances that we focus on in this section are defined as follows. The distance $d\left(\omega_{1}, \omega_{2}\right)$ between two states $\omega_{1}, \omega_{2} \in \mathcal{D}(X)$, on the same set $X$, can be defined as the join of the distances in $[0,1]$, between validities:

$$
\begin{equation*}
d\left(\omega_{1}, \omega_{2}\right):=\bigvee_{p \in \operatorname{Pred}(X)}\left|\omega_{1} \vDash p-\omega_{2} \vDash p\right| \tag{4.16}
\end{equation*}
$$

Similarly, the distance $d\left(p_{1}, p_{2}\right)$ between two predicates $p_{1}, p_{2} \in \operatorname{Pred}(X)$ on the same set, is defined as:

$$
\begin{equation*}
d\left(p_{1}, p_{2}\right):=\bigvee_{\omega \in \mathcal{D}(X)}\left|\omega \vDash p_{1}-\omega \vDash p_{2}\right| . \tag{4.17}
\end{equation*}
$$

Note that the above formulations involve predicates only, not observables in general.


Figure 4.2 Distance graph between distributions, see Example 4.5.2

### 4.5.1 Distance between states

The distance defined in 4.16 is commonly called the total variation distance, which is a special case of the Kantorovich distance, see e.g. [61, 17, 132, 128]. Its two alternative characterisations below are standard. We refer to [95] for more information about the validity-based approach.

Proposition 4.5.1. Let $X$ be an arbitrary set, with states $\omega_{1}, \omega_{2} \in \mathcal{D}(X)$. Then:

$$
d\left(\omega_{1}, \omega_{2}\right)=\max _{U \subseteq X} \omega_{1} \vDash \mathbf{1}_{U}-\omega_{2} \models \mathbf{1}_{U}=\frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right| .
$$

We write maximum 'max' instead of join $V$ to express that the supremum is actually reached by a subset (sharp predicate).

Proof. Let $\omega_{1}, \omega_{2} \in \mathcal{D}(X)$ be two discrete probability distributions on the same set $X$. We will prove the two inequalities labeled $(a)$ and $(b)$ in:

$$
\begin{aligned}
\frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right| & \stackrel{(a)}{\leq} \max _{U \leq X} \omega_{1} \vDash \mathbf{1}_{U}-\omega_{2} \vDash \mathbf{1}_{U} \\
& \leq \bigvee_{p \in \operatorname{Pred}(X)}\left|\omega_{1} \vDash p-\omega_{2} \vDash p\right| \\
& \stackrel{(b)}{\leq} \frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right| .
\end{aligned}
$$

This proves Proposition 4.5 .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 $\omega_{i}(U)=\sum_{x \in U} \omega_{i}(x)=\left(\omega \vDash \mathbf{1}_{U}\right)$. We partition $U$ in three disjoint parts, and take the relevant sums:

$$
\left\{\begin{array} { l } 
{ U _ { > } = \{ x \in U | \omega _ { 1 } ( x ) > \omega _ { 2 } ( x ) \} } \\
{ U _ { > } = \{ x \in U | \omega _ { 1 } ( x ) = \omega _ { 2 } ( x ) \} } \\
{ U _ { < } = \{ x \in U | \omega _ { 1 } ( x ) < \omega _ { 2 } ( x ) \} }
\end{array} \quad \left\{\begin{array}{l}
U \uparrow=\omega_{1}\left(U_{>}\right)-\omega_{2}\left(U_{>}\right) \geq 0 \\
U \downarrow=\omega_{2}\left(U_{<}\right)-\omega_{1}\left(U_{<}\right) \geq 0
\end{array}\right.\right.
$$

We use this notation in particular for $U=X$. In that case we can use:

$$
\begin{aligned}
& 1=\omega_{1}(X)=\omega_{1}\left(X_{>}\right)+\omega_{1}\left(X_{-}\right)+\omega_{1}\left(X_{<}\right) \\
& 1=\omega_{2}(X)=\omega_{2}\left(X_{>}\right)+\omega_{2}\left(X_{=}\right)+\omega_{2}\left(X_{<}\right)
\end{aligned}
$$

Hence by subtraction we obtain, since $\omega_{1}\left(X_{=}\right)=\omega_{2}\left(X_{=}\right)$,

$$
0=\left(\omega_{1}\left(X_{>}\right)-\omega_{2}\left(X_{>}\right)\right)+\left(\omega_{1}\left(X_{<}\right)-\omega_{2}\left(X_{<}\right)\right)
$$

That is,

$$
X \uparrow=\omega_{1}\left(X_{>}\right)-\omega_{2}\left(X_{>}\right)=\omega_{2}\left(X_{<}\right)-\omega_{1}\left(X_{<}\right)=X \downarrow
$$

As a result:

$$
\begin{aligned}
& \frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right| \\
& =\frac{1}{2}\left(\sum_{x \in X_{>}}\left(\omega_{1}(x)-\omega_{2}(x)\right)+\sum_{x \in X_{<}}\left(\omega_{2}(x)-\omega_{1}(x)\right)\right) \\
& =\frac{1}{2}\left(\left(\omega_{1}\left(X_{>}\right)-\omega_{2}\left(X_{>}\right)\right)+\left(\omega_{2}\left(X_{<}\right)-\omega_{1}\left(X_{<}\right)\right)\right) \\
& =\frac{1}{2}(X \uparrow+X \downarrow) \\
& =X \uparrow
\end{aligned}
$$

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:

$$
\begin{aligned}
\frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right| \stackrel{4.18}{=} X \uparrow & =\omega_{1}\left(X_{>}\right)-\omega_{2}\left(X_{>}\right) \\
& =\omega_{1} \vDash \mathbf{1}_{X_{>}}-\omega_{2} \vDash \mathbf{1}_{X_{>}} \\
& \leq \max _{U \subseteq X} \omega_{1} \vDash \mathbf{1}_{U}-\omega_{2} \vDash \mathbf{1}_{U} .
\end{aligned}
$$

(b) Let $p \in \operatorname{Pred}(X)$ be an arbitrary predicate. We have: $\left(\mathbf{1}_{U} \& p\right)(x)=\mathbf{1}_{U}(x)$.

$$
\begin{aligned}
& p(x) \text {, which is } p(x) \text { if } x \in U \text { and } 0 \text { otherwise. Then: } \\
& \left|\omega_{1} \vDash p-\omega_{2} \vDash p\right| \\
& =\mid\left(\omega_{1} \vDash \mathbf{1}_{X_{>}} \& p+\omega_{1} \vDash \mathbf{1}_{X_{\overline{=}}} \& p+\omega_{1} \vDash \mathbf{1}_{X_{<}} \& p\right) \\
& -\left(\omega_{2} \vDash \mathbf{1}_{X_{>}} \& p+\omega_{2} \vDash \mathbf{1}_{X_{-}} \& p+\omega_{2} \vDash \mathbf{1}_{X_{<}} \& p\right) \\
& =\left|\left(\omega_{1} \vDash \mathbf{1}_{X_{>}} \& p-\omega_{2} \vDash \mathbf{1}_{X_{>}} \& p\right)-\left(\omega_{2} \vDash \mathbf{1}_{X_{<}} \& p-\omega_{1} \vDash \mathbf{1}_{X_{<}} \& p\right)\right| \\
& =\left\{\begin{array}{l}
\left(\omega_{1} \vDash \mathbf{1}_{X_{>}} \& p-\omega_{2} \vDash \mathbf{1}_{X_{>}} \& p\right)-\left(\omega_{2} \vDash \mathbf{1}_{X_{<}} \& p-\omega_{1} \vDash \mathbf{1}_{X_{<}} \& p\right) \\
\text { if } \omega_{1} \vDash \mathbf{1}_{X_{>}} \& p-\omega_{2} \vDash \mathbf{1}_{X_{>}} \& p \stackrel{(*)}{\geq} \omega_{2} \vDash \mathbf{1}_{X_{<}} \& p-\omega_{1} \vDash \mathbf{1}_{X_{<}} \& p \\
\left(\omega_{2} \vDash \mathbf{1}_{X_{<}} \& p-\omega_{1} \vDash \mathbf{1}_{X_{<}} \& p\right)-\left(\omega_{1} \vDash \mathbf{1}_{X_{>}} \& p-\omega_{2} \vDash \mathbf{1}_{X_{>}} \& p\right) \\
\text { otherwise }
\end{array}\right. \\
& \leq \begin{cases}\omega_{1} \vDash \mathbf{1}_{X_{>}} \& p-\omega_{2} \vDash \mathbf{1}_{X_{>}} \& p & \text { if }(*) \\
\omega_{2} \vDash \mathbf{1}_{X_{<}} \& p-\omega_{1} \vDash \mathbf{1}_{X_{<}} \& p & \text { otherwise }\end{cases} \\
& = \begin{cases}\sum_{x \in X_{>}}\left(\omega_{1}(x)-\omega_{2}(x)\right) \cdot p(x) & \text { if }(*) \\
\sum_{x \in X_{<}}\left(\omega_{2}(x)-\omega_{1}(x)\right) \cdot p(x) & \text { otherwise }\end{cases} \\
& \leq \begin{cases}\sum_{x \in X_{>}} \omega_{1}(x)-\omega_{2}(x) & \text { if }(*) \\
\sum_{x \in X_{<}} \omega_{2}(x)-\omega_{1}(x) & \text { otherwise }\end{cases} \\
& = \begin{cases}X \uparrow & \text { if }(*) \\
X \downarrow=X \uparrow & \text { otherwise }\end{cases} \\
& =X \uparrow \\
& \text { (4.18) } \frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right| \text {. }
\end{aligned}
$$

This completes the proof.
Example 4.5.2. Consider the set of 'fractional' distributions:

$$
\{\operatorname{Flrn}(\varphi) \mid \varphi \in \mathcal{N}[4](\{a, b, c\})\} .
$$

Its $\left.\binom{3}{4}\right)=15$ elements form a triangle, as described in Figure 4.2, with each edge describing a (total variation) distance of $\frac{1}{4}$.

The sum-formulation in Proposition 4.5.1 is useful in many situations, for instance in order to prove that the above distance function $d$ between states is a metric.

Lemma 4.5.3. The distance $d\left(\omega_{1}, \omega_{2}\right)$ between states $\omega_{1}, \omega_{2} \in \mathcal{D}(X)$ in 4.16 turns the set of distributions $\mathcal{D}(X)$ into a metric space, with $[0,1]$-valued metric.

Proof. If $d\left(\omega_{1}, \omega_{2}\right)=\frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right|=0$, then $\left|\omega_{1}(x)-\omega_{2}(x)\right|=0$ for each $x \in X$, so that $\omega_{1}(x)=\omega_{2}(x)$, and thus $\omega_{1}=\omega_{2}$. Obviously, $d\left(\omega_{1}, \omega_{2}\right)=$
$d\left(\omega_{2}, \omega_{1}\right)$. The triangle inequality holds for $d$ since it holds for the standard distance on $[0,1]$.

$$
\begin{aligned}
d\left(\omega_{1}, \omega_{3}\right) & =\frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{3}(x)\right| \\
& \leq \frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right|+\left|\omega_{2}(x)-\omega_{3}(x)\right| \\
& =\frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right|+\frac{1}{2} \sum_{x \in X}\left|\omega_{2}(x)-\omega_{3}(x)\right| \\
& =d\left(\omega_{1}, \omega_{2}\right)+d\left(\omega_{2}, \omega_{3}\right) .
\end{aligned}
$$

We use this same sum-formulation for the following result. It uses the notion of non-expansive function, which is frequently used as choice of mapping between metric spaces. Explicitly $f:(X, d) \rightarrow(Y, d)$ is non-expansive if $d\left(f(x), f\left(x^{\prime}\right)\right) \leq d\left(x, x^{\prime}\right)$, for all $x, x^{\prime} \in X$. Such non-expansive functions are automatically continuous, since if $d\left(x_{n}, x\right) \rightarrow 0$ as $n$ goes to infinity, then also $d\left(f\left(x_{n}\right), f(x)\right) \rightarrow 0$.

Lemma 4.5.4. State transformation is non-expansive: for a channel $c: X \leadsto Y$ one has:

$$
d\left(c \gg=\omega_{1}, c \gg=\omega_{2}\right) \leq d\left(\omega_{1}, \omega_{2}\right)
$$

Proof. Since:

$$
\begin{aligned}
d\left(c \gg=\omega_{1}, c \gg=\omega_{2}\right) & \left.=\frac{1}{2} \sum_{y \in Y} \right\rvert\,\left(c \gg=\omega_{1}\right)(y)-\left(c \gg=\omega_{2}(y) \mid\right. \\
& =\frac{1}{2} \sum_{y \in Y}\left|\sum_{x \in X} \omega_{1}(x) \cdot c(x)(y)-\sum_{x \in X} \omega_{2}(x) \cdot c(x)(y)\right| \\
& =\frac{1}{2} \sum_{y \in Y}\left|\sum_{x \in X} c(x)(y) \cdot\left(\omega_{1}(x)-\omega_{2}(x)\right)\right| \\
& \leq \frac{1}{2} \sum_{y \in Y} \sum_{x \in X} c(x)(y) \cdot\left|\omega_{1}(x)-\omega_{2}(x)\right| \\
& =\frac{1}{2} \sum_{x \in X}\left(\sum_{y \in Y} c(x)(y)\right) \cdot\left|\omega_{1}(x)-\omega_{2}(x)\right| \\
& =d\left(\omega_{1}, \omega_{2}\right) .
\end{aligned}
$$

We recall from Definition 3.1.5 that a coupling of two states $\omega_{1}, \omega_{2} \in \mathcal{D}(X)$ is a joint state $\sigma \in \mathcal{D}(X \times X)$ that marginalises to $\omega_{1}$ and $\omega_{2}$, i.e. that satisfies $\sigma[1,0]=\omega_{1}$ and $\sigma[0,1]=\omega_{2}$. 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 [180] 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.

Proposition 4.5.5. For states $\omega_{1}, \omega_{2} \in \mathcal{D}(X)$,

$$
\begin{equation*}
d\left(\omega_{1}, \omega_{2}\right)=\bigwedge\left\{\sigma \vDash E q^{\perp} \mid \sigma \text { is a coupling between } \omega_{1}, \omega_{2}\right\} \tag{4.19}
\end{equation*}
$$

where $E q: X \times X \rightarrow[0,1]$ is the equality predicate from Definition 4.1.1 (4).
The binary predicate $E q^{\perp}$ is the discrete distance on the set $X$, since for $x, x^{\prime} \in X$ one has:

$$
\begin{aligned}
E q^{\perp}\left(x, x^{\prime}\right)=1-E q\left(x, x^{\prime}\right) & = \begin{cases}1-1 & \text { if } x=x^{\prime} \\
1-0 & \text { if } x \neq x^{\prime}\end{cases} \\
& = \begin{cases}0 & \text { if } x=x^{\prime} \\
1 & \text { if } x \neq x^{\prime} .\end{cases}
\end{aligned}
$$

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 $E q^{\perp}$.

As we have seen in Exercise 2.3.5 there can be infinitely many couplings of two given distributions. This makes computing the infimum in 4.19) 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 4.5.1 We first prove the inequality $(\leq)$. Write $X_{>}=\left\{x \in X \mid \omega_{1}(x)>\omega_{2}(x)\right\}$ and let $\sigma$ be a coupling of $\omega_{1}, \omega_{2}$. Then:

$$
\begin{aligned}
\omega_{1}(x)=\sum_{y \in X} \sigma(x, y) & =\sigma(x, x)+\sum_{y \neq x} \sigma(x, y) \\
& \leq \omega_{2}(x)+\left(\sigma \vDash\left(\mathbf{1}_{x} \otimes \mathbf{1}\right) \& E q^{\perp}\right) .
\end{aligned}
$$

This means that $\omega_{1}(x)-\omega_{2}(x) \leq \sigma \vDash\left(\mathbf{1}_{x} \otimes \mathbf{1}\right) \& E q^{\perp}$ for $x \in X_{>}$. We similarly have:

$$
\begin{aligned}
\omega_{2}(x)=\sum_{x \in X} \sigma(y, x) & =\sigma(x, x)+\sum_{y \neq x} \sigma(y, x) \\
& \leq \omega_{1}(x)+\left(\sigma \vDash\left(\mathbf{1} \otimes \mathbf{1}_{x}\right) \& E q^{\perp}\right) .
\end{aligned}
$$

[^6]Hence $\omega_{2}(x)-\omega_{1}(x)$ for $x \notin X_{>}$. Putting this together gives:

$$
\begin{aligned}
d\left(\omega_{1}, \omega_{2}\right) & =\frac{1}{2} \sum_{x \in X}\left|\omega_{1}(x)-\omega_{2}(x)\right| \\
& =\frac{1}{2} \sum_{x \in X_{>}} \omega_{1}(x)-\omega_{2}(x)+\frac{1}{2} \sum_{x \in \neg X_{>}} \omega_{2}(x)-\omega_{1}(x) \\
& \leq \frac{1}{2} \sum_{x \in X_{>}} \sigma \vDash\left(\mathbf{1}_{x} \otimes \mathbf{1}\right) \& E q^{\perp}+\frac{1}{2} \sum_{x \in \neg X_{>}} \sigma \vDash\left(\mathbf{1} \otimes \mathbf{1}_{x}\right) \& E q^{\perp} \\
& =\frac{1}{2} \sigma \vDash\left(\mathbf{1}_{X_{>}} \otimes \mathbf{1}\right) \& E q^{\perp}+\frac{1}{2} \sigma \models\left(\mathbf{1} \otimes \mathbf{1}_{\neg X_{>}}\right) \& E q^{\perp} \\
& \leq \frac{1}{2} \sigma \vDash E q^{\perp}+\frac{1}{2} \sigma \vDash E q^{\perp} \\
& =\sigma \vDash E q^{\perp} .
\end{aligned}
$$

For the inequality $(\geq)$ one uses what is called an optimal coupling $\rho \in \mathcal{D}(X \times$ $X)$ of $\omega_{1}, \omega_{2}$. It can be defined as:

$$
\rho(x, y):= \begin{cases}\min \left(\omega_{1}(x), \omega_{2}(x)\right) & \text { if } x=y  \tag{4.20}\\ \frac{\max \left(\omega_{1}(x)-\omega_{2}(x), 0\right) \cdot \max \left(\omega_{2}(y)-\omega_{1}(y), 0\right)}{d\left(\omega_{1}, \omega_{2}\right)} & \text { otherwise }\end{cases}
$$

We first check that this $\rho$ is a coupling. Let $x \in X_{>}$so that $\omega_{1}(x)>\omega_{2}(x)$; then:

$$
\begin{aligned}
& \sum_{y \in X} \rho(x, y) \\
& =\omega_{2}(x)+\left(\omega_{1}(x)-\omega_{2}(x)\right) \cdot \sum_{y \neq x} \frac{\max \left(\omega_{2}(y)-\omega_{1}(y), 0\right)}{d\left(\omega_{1}, \omega_{2}\right)} \\
& =\omega_{2}(x)+\left(\omega_{1}(x)-\omega_{2}(x)\right) \cdot \frac{\sum_{y \in X_{<}} \omega_{2}(y)-\omega_{1}(y)}{d\left(\omega_{1}, \omega_{2}\right)} \\
& =\omega_{2}(x)+\left(\omega_{1}(x)-\omega_{2}(x)\right) \cdot \frac{X \downarrow}{d\left(\omega_{1}, \omega_{2}\right)} \\
& =\omega_{2}(x)+\left(\omega_{1}(x)-\omega_{2}(x)\right) \cdot 1 \quad \text { see the proof of Proposition } 4.5 .1 \\
& =\omega_{1}(x) .
\end{aligned}
$$

If $x \notin X_{>}$, so that $\omega_{1}(x) \leq \omega_{2}(x)$, then it is obvious that $\sum_{y} \rho(x, y)=\omega_{1}(x)+0=$ $\omega_{1}(x)$. This shows $\sigma[1,0]=\omega_{1}$. In a similar way one obtains $\sigma[0,1]=\omega_{2}$. Finally,

$$
\begin{aligned}
\rho \vDash E q=\sum_{x \in X} \rho(x, x) & =\sum_{x \in X} \min \left(\omega_{1}(x), \omega_{2}(x)\right) \\
& =\sum_{x \in X_{>}} \omega_{2}(x)+\sum_{x \notin X_{>}} \omega_{1}(x) \\
& =\omega_{2}\left(X_{>}\right)+1-\omega_{1}\left(X_{>}\right) \\
& =1-\left(\omega_{1}\left(X_{>}\right)-\omega_{2}\left(X_{>}\right)\right)=1-d\left(\omega_{1}, \omega_{2}\right) .
\end{aligned}
$$



Figure 4.3 Visual comparance of distance and divergence between flip states, see Remark 4.5.6 for details.

Hence $d\left(\omega_{1}, \omega_{2}\right)=1-(\rho \vDash E q)=\rho \vDash E q^{\perp}$.
Now that we have a good understanding of the total variation distance on distributions, there are a couple of comparisons to make.

Remark 4.5.6. In Definition 2.8.1 we have seen Kullback-Leibler divergence $D_{K L}$, as a measure of difference between states. However, this $D_{K L}$ is not a proper metric, since it is not symmetric, see Exercise 2.8.1 Nevertheless, it is often used as distance between states, especially in minimisation problems (see e.g. Exercise ??).

Figure 4.3 compares the total variation distance and the Kullback-Leibler divergence between two flip states:

$$
d(\operatorname{flip}(r), \operatorname{flip}(s)) \quad D_{K L}(\operatorname{flip}(r), \text { flip }(s)) .
$$

for $r, s \in[0,1]$, where, recall, 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.

Remark 4.5.7. 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:

$$
\|x\|_{t v}:=\frac{1}{2} \cdot \sum_{i} x_{i}, \quad \text { for } x \in[0,1]^{N} .
$$

However, on $n$-tuples of real numbers the Euclidean norm is common, defined as:

$$
\|x\|_{e u}:=\sqrt{\sum_{i} x_{i}^{2}}, \quad \text { here for } x \in[0,1]^{N} .
$$

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. [30] III, Prop. 1.5]. For this it suffices to show that the norms are related via constants. In our case we have:

$$
\frac{2}{N} \cdot\|x\|_{t v} \leq\|x\|_{e u} \leq 2 \sqrt{N} \cdot\|x\|_{t v}
$$

For the first inequality we use that $x_{i}=\sqrt{x_{i}^{2}} \leq\|x\|_{e u}$, so $\|x\|_{t v}=\frac{1}{2} \cdot \sum_{i} x_{i} \leq$ $\frac{N}{2} \cdot\|x\|_{\text {eu }}$. For the second inequality we abbreviate $s=\|x\|_{t v}$ and observe that $x_{i} \leq 2 s$. Hence $x_{i}^{2} \leq 4 s^{2}$ and thus $\sum_{i} x_{i}^{2} \leq 4 N \cdot s^{2}$. But then: $\|x\|_{\text {eu }}=\sqrt{\sum_{i} x_{i}^{2}} \leq$ $2 \sqrt{N} \cdot s$.

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 numbers in Section 5.5, see esp. Theorem 5.5.4. Here we take a an algorithmic perspective.

Remark 4.5.8. 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 $\operatorname{Flrn}\left(\varphi_{K}\right)$ getting closer and closer to $\omega$, 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 $\omega$ has support $\left\{x_{1}, \ldots, x_{N}\right\}$, of course with $N>0$. Below we use $k \cdot \omega \in \mathcal{M}\left(\left\{x_{1}, \ldots, x_{n}\right\}\right)$, for $k \geq 0$.

- Pick $\varphi_{1}=1\left|x_{j}\right\rangle$ where $\omega$ takes a maximum at $x_{j}$, i.e., $\omega\left(x_{j}\right) \geq \omega\left(x_{i}\right)$ for all $i$. Set variable $c p:=j$, for 'current position'.
- Look for the first position $i$ after $c p$ where $\varphi_{K}\left(x_{i}\right)<(n+1) \cdot \omega\left(x_{i}\right)$. Then you set $\varphi_{K+1}:=\varphi_{K}+1\left|x_{i}\right\rangle$, and $c p:=i$. This search wraps around, if needed. When no $i$ is found we are done and have $\operatorname{Flrn}\left(\varphi_{K}\right)=\omega$.

Concretely, if $\omega=\frac{1}{6}\left|x_{1}\right\rangle+\frac{1}{2}\left|x_{2}\right\rangle+\frac{1}{3}\left|x_{3}\right\rangle$. Then, consecutively,

- $0 \cdot \omega=\mathbf{0}$ and $\varphi_{1}=1\left|x_{2}\right\rangle$
- $1 \cdot \omega=\frac{1}{6}\left|x_{1}\right\rangle+\frac{1}{2}\left|x_{2}\right\rangle+\frac{1}{3}\left|x_{3}\right\rangle$ and $\varphi_{2}=1\left|x_{2}\right\rangle+1\left|x_{3}\right\rangle$
- $2 \cdot \omega=\frac{1}{3}\left|x_{1}\right\rangle+1\left|x_{2}\right\rangle+\frac{2}{3}\left|x_{3}\right\rangle$ and $\varphi_{3}=1\left|x_{1}\right\rangle+1\left|x_{2}\right\rangle+1\left|x_{3}\right\rangle$
- $3 \cdot \omega=\frac{1}{2}\left|x_{1}\right\rangle+\frac{3}{2}\left|x_{2}\right\rangle+1\left|x_{3}\right\rangle$ and $\varphi_{4}=1\left|x_{1}\right\rangle+2\left|x_{2}\right\rangle+1\left|x_{3}\right\rangle$
- $4 \cdot \omega=\frac{2}{3}\left|x_{1}\right\rangle+2\left|x_{2}\right\rangle+\frac{4}{3}\left|x_{3}\right\rangle$ and $\varphi_{5}=1\left|x_{1}\right\rangle+2\left|x_{2}\right\rangle+2\left|x_{3}\right\rangle$
- $5 \cdot \omega=\frac{5}{6}\left|x_{1}\right\rangle+\frac{5}{2}\left|x_{2}\right\rangle+\frac{5}{3}\left|x_{3}\right\rangle$ and $\varphi_{6}=1\left|x_{1}\right\rangle+3\left|x_{2}\right\rangle+2\left|x_{3}\right\rangle$
- $6 \cdot \omega=1\left|x_{1}\right\rangle+3\left|x_{2}\right\rangle+2\left|x_{3}\right\rangle$, giving rise to a halt, since $\varphi_{6}=6 \cdot \omega$ and thus $\operatorname{Flrn}\left(\varphi_{6}\right)=\omega$.

In this case we get a finite sequence of multisets approaching $\omega$. The sequence is infinite for, e.g.,

$$
\omega=\frac{1}{7}\left|x_{1}\right\rangle+\frac{4}{7}\left|x_{2}\right\rangle+\frac{2}{7}\left|x_{3}\right\rangle \approx 0.1429\left|x_{1}\right\rangle+0.5714\left|x_{2}\right\rangle+0.2857\left|x_{3}\right\rangle
$$

Running the above algorithm gives, for instance:

- $\varphi_{10}=2\left|x_{1}\right\rangle+5\left|x_{2}\right\rangle+3\left|x_{3}\right\rangle$
- $\varphi_{100}=15\left|x_{1}\right\rangle+56\left|x_{2}\right\rangle+29\left|x_{3}\right\rangle$
- $\varphi_{1000}=143\left|x_{1}\right\rangle+571\left|x_{2}\right\rangle+286\left|x_{3}\right\rangle$
- $\varphi_{10000}=1429\left|x_{1}\right\rangle+5713\left|x_{2}\right\rangle+2858\left|x_{3}\right\rangle$.

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. 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 [89].

## Theorem 4.5.9.

1 For an arbitrary set $X$, the set $\mathcal{D}(X)$ has a dense subset:

$$
\bigcup_{K \in \mathbb{N}} \mathcal{D}[K](X) \subseteq \mathcal{D}(X) \quad \text { where } \quad \mathcal{D}[K](X):=\{F \operatorname{lrn}(\varphi) \mid \varphi \in \mathcal{N}[K](X)\} .
$$

We often refer to the elements of $\mathcal{D}[K](X)$ as fractional distributions.
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.
4 When $X$ is finite, the space $\mathcal{D}(X)$ is totally bounded, and thus compact.

Proof. 1 Let $\omega \in \mathcal{D}(X)$ and $\varepsilon>0$. We need to find a multiset $\varphi \in \mathcal{N}(X)$ with $d(\omega, \operatorname{Flrn}(\varphi))<\varepsilon$. There is a systematic way to find such multisets via the decimal representation of the probabilities in $\omega$. 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:

$$
\begin{aligned}
\varphi_{1}:= & 3|a\rangle+4|b\rangle+2|c\rangle & \text { with } & d\left(\omega, \operatorname{Flrn}\left(\varphi_{1}\right)\right) \leq \frac{1}{2} \cdot 3 \cdot 10^{-1} \\
\varphi_{2}:= & 38|a\rangle+40|b\rangle+20|c\rangle & \text { with } & d\left(\omega, \operatorname{Flrn}\left(\varphi_{2}\right)\right) \leq \frac{1}{2} \cdot 3 \cdot 10^{-2} \\
\varphi_{3}:= & 383|a\rangle+406|b\rangle+209|c\rangle & \text { with } & d\left(\omega, \operatorname{Flrn}\left(\varphi_{3}\right)\right) \leq \frac{1}{2} \cdot 3 \cdot 10^{-3} \\
& \text { etc. } & &
\end{aligned}
$$

In general, for a distribution $\omega$ with $\operatorname{supp}(\omega)=\left\{x_{1}, \ldots, x_{M}\right\}$ we can thus construct a sequence of multisets $\varphi_{n} \in \mathcal{N}\left(\left\{x_{1}, \ldots, x_{M}\right\}\right)$ with $d\left(\omega, \operatorname{Flrn}\left(\varphi_{n}\right)\right) \leq$ $\frac{1}{2} \cdot M \cdot 10^{-n}$. 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 $X=\left\{x_{1}, \ldots, x_{M}\right\}$ and $\omega_{i} \in \mathcal{D}(X)$ be a Cauchy sequence. Fix $n$. Then for all $i, j$,

$$
\left|\omega_{i}\left(x_{n}\right)-\omega_{j}\left(x_{n}\right)\right| \leq 2 \cdot d\left(\omega_{i}, \omega_{j}\right)
$$

Hence, the sequence $\omega_{i}\left(x_{n}\right) \in[0,1]$ is a Cauchy sequence, say with limit $r_{n} \in[0,1]$. Take $\omega=\sum_{n} r_{n}\left|x_{n}\right\rangle \in \mathcal{D}(X)$. This is the limit of the distributions $\omega_{i}$.
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.
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 $\varepsilon$-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 $d(\omega, \operatorname{Flrn}(\varphi))<\varepsilon$. This shows:

$$
\begin{aligned}
\mathcal{D}(X) & \subseteq \bigcup_{\varphi \in \mathcal{N}[K](X)}\{\omega \in \mathcal{D}(X) \mid d(\omega, \operatorname{Flrn}(\varphi))<\varepsilon\} \\
& =\bigcup_{\varphi \in \mathcal{N}[K](X)} B_{\varepsilon}(\operatorname{Flrn}(\varphi)) .
\end{aligned}
$$



Figure 4.4 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

Notice that the sequence of multisets $\varphi_{K}$ approaching $\omega$ described in Remark 4.5.8 has the special property that $\left\|\varphi_{K}\right\|=K$. This does not hold for the sequence $\varphi_{n}$ in the above proof. Such a size property is not needed there.

The denseness of the fractional distributions is illustrated in Figure 4.4
A different way to approximate a distribution $\omega$ is described in Section 5.5 , via what is called the law of large numbers.

### 4.5.2 Distance between predicates

What we have to say about the validity-based distance (4.17) between predicates is rather brief. First, there is also a pointwise formulation.

Lemma 4.5.10. For two predicates $p_{1}, p_{2} \in \operatorname{Pred}(X)$,

$$
d\left(p_{1}, p_{2}\right)=\bigvee_{x \in X}\left|p_{1}(x)-p_{2}(x)\right| .
$$

This distance function d makes the set $\operatorname{Pred}(X)$ into a metric space.

Proof. First, we have for each $x \in X$,

$$
\begin{aligned}
d\left(p_{1}, p_{2}\right) & \stackrel{\boxed{4.17}}{=} \bigvee_{\omega \in \mathcal{D}(X)}\left|\omega \vDash p_{1}-\omega \vDash p_{2}\right| \\
& \geq\left|\operatorname{unit}(x) \vDash p_{1}-\operatorname{unit}(x) \vDash p_{2}\right| \\
& =\left|p_{1}(x)-p_{2}(x)\right| .
\end{aligned}
$$

Hence $d\left(p_{1}, p_{2}\right) \geq \bigvee_{x}\left|p_{1}(x)-p_{2}(x)\right|$.
The other direction follows from:

$$
\begin{aligned}
\left|\omega \vDash p_{1}-\omega \vDash p_{2}\right| & =\left|\sum_{z \in X} \omega(z) \cdot p_{1}(z)-\sum_{z \in X} \omega(z) \cdot p_{2}(z)\right| \\
& \leq \sum_{z \in X} \omega(z) \cdot\left|p_{1}(z)-p_{2}(z)\right| \\
& \leq \sum_{z \in X} \omega(z) \cdot \bigvee_{x \in X}\left|p_{1}(x)-p_{2}(x)\right| \\
& =\left(\sum_{z \in X} \omega(z)\right) \cdot \bigvee_{x \in X}\left|p_{1}(x)-p_{2}(x)\right| \\
& =\bigvee_{x \in X}\left|p_{1}(x)-p_{2}(x)\right| .
\end{aligned}
$$

The fact that we get a metric space is now straightforward.
There is an analogue of Lemma 4.5.4.
Lemma 4.5.11. Predicate transformation is also non-expansive: for a channel $c: X \leadsto Y$ one has, for predicates $p_{1}, p_{2} \in \operatorname{Pred}(Y)$,

$$
d\left(c=\ll p_{1}, c=\ll p_{2}\right) \leq d\left(p_{1}, p_{2}\right)
$$

Proof. Via the formulation of Lemma 4.5.10 we get:

$$
\begin{aligned}
d\left(c=\ll p_{1}, c=\ll p_{2}\right) & =\bigvee_{x \in X}\left|\left(c=\ll p_{1}\right)(x)-\left(c=\ll p_{2}\right)(x)\right| \\
& =\bigvee_{x \in X}\left|\sum_{y \in Y} c(x)(y) \cdot p_{1}(y)-\sum_{y \in Y} c(x)(y) \cdot p_{2}(y)\right| \\
& \leq \bigvee_{x \in X} \sum_{y \in Y} c(x)(y) \cdot\left|p_{1}(y)-p_{2}(y)\right| \\
& \leq \bigvee_{x \in X}\left(\sum_{y \in Y} c(x)(y)\right) \cdot d\left(p_{1}, p_{2}\right) \\
& =d\left(p_{1}, p_{2}\right) .
\end{aligned}
$$

## Exercises

4.5.1 Prove, analogously to Excercise 3.3.2, that the total variation distance satisfies, for distributions $\omega_{1}, \omega_{2} \in \mathcal{D}(X)$ and $K \in \mathbb{N}$,

$$
d\left(\omega_{1}, \omega_{2}\right) \leq d\left(m n[K]\left(\omega_{1}\right), m n[K]\left(\omega_{2}\right)\right) .
$$

4.5.2 1 Prove that for states $\omega, \omega^{\prime} \in \mathcal{D}(X)$ and $\rho, \rho^{\prime} \in \mathcal{D}(Y)$ there is an inequality:

$$
d\left(\omega \otimes \rho, \omega^{\prime} \otimes \rho^{\prime}\right) \leq d\left(\omega, \omega^{\prime}\right)+d\left(\rho, \rho^{\prime}\right)
$$

(In this situation there is an actual equality for Kullback-Leibler divergence, see Lemma 2.8.2 (2).)
2 Prove similarly that for predicates $p, p^{\prime} \in \operatorname{Pred}(X)$ and $q, q^{\prime} \in$ $\operatorname{Pred}(Y)$ one gets:

$$
d\left(p \otimes q, p^{\prime} \otimes q^{\prime}\right) \leq d\left(p, p^{\prime}\right)+d\left(q, q^{\prime}\right)
$$

4.5.3 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 $\frac{1}{K} \mathrm{in}$ :

$$
d(\operatorname{Flrn}(\varphi), \operatorname{Flrn}(\psi)) \geq \frac{1}{K}
$$

(See also Figure 4.2)
4.5.4 1 Show that for a state $\omega \in \mathcal{D}(X)$, the "validity in $\omega$ " function (predicate):

$$
\operatorname{Pred}(X) \xrightarrow{\omega \vDash(-)}[0,1]
$$

is non-expansive.
2 Similarly, show that for a predicate $p \in \mathcal{D}(X)$ the "validity of $p$ " function is non-expansive:

$$
\mathcal{D}(X) \xrightarrow{(-) \vDash p}[0,1]
$$

4.5.5 In the context of Remark 4.5.6, check that:

$$
\begin{aligned}
d(f l i p(0), \operatorname{flip}(1)) & =1=d(\operatorname{flip}(1), \operatorname{flip}(0)) \\
D_{K L}(\operatorname{flip}(0), \operatorname{flip}(1)) & =0=D_{K L}(\operatorname{flip}(1), \operatorname{flip}(0)) .
\end{aligned}
$$

(Using that the logarithm of zero is defined to be zero.)
4.5.6 This exercise uses the distance between a joint state and the product of its marginals as measure of entwinedness, like in [77].

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:

$$
d\left(\sigma_{2}, \sigma_{2}[1,0] \otimes \sigma_{2}[0,1]\right)=\frac{1}{2}
$$

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:

$$
d\left(\sigma_{3}, \sigma_{3}[1,0,0] \otimes \sigma_{3}[0,1,0] \otimes \sigma_{3}[0,0,1]\right)=\frac{3}{4} .
$$

3 Now define $\sigma_{n} \in \mathcal{D}\left(\mathbf{2}^{n}\right)$ for $n \geq 2$ as:

$$
\sigma_{n}:=\frac{1}{2}|\underbrace{0 \cdots 0}_{n \text { times }}\rangle+\frac{1}{2}|\underbrace{1 \cdots 1}_{n \text { times }}\rangle .
$$

Show that:

- each marginal $\pi_{i} \gg=\sigma_{n}$ equals $\frac{1}{2}|0\rangle+\frac{1}{2}|1\rangle$;
- the product $\bigotimes_{i}\left(\pi_{i} \gg \sigma_{n}\right)$ of the marginals is the uniform state
- on $\left.2^{n} ;{ }^{\prime}, \sigma_{i}\left(\pi_{i} \gg=\sigma_{n}\right)\right)=\frac{2^{n-1}-1}{2^{n-1}}$.

Note that the latter distance goes to 1 as $n$ goes to infinity.
4.5.7 The next 'splitting lemma' is attributed to Jones [105], see e.g. [107, 128]. For $\omega_{1}, \omega_{2} \in \mathcal{D}(X)$ with distance $d:=d\left(\omega_{1}, \omega_{2}\right)$ one can find distributions $\omega_{1}^{\prime}, \omega_{2}^{\prime}, \sigma \in \mathcal{D}(X)$ so that both $\omega_{1}$ and $\omega_{2}$ can be written as convex sum:

$$
\omega_{i}=d \cdot \omega_{i}^{\prime}+(1-d) \cdot \sigma .
$$

Prove this result.
Hint: Use the optimal coupling $\rho$ from (4.20) to define $\sigma(x)=\frac{\rho(x, x)}{1-d}$.

## 5

## Variance and covariance

The previous chapter introduced validity $\omega \vDash p$, of an observable $p$ in a state / distribution $\omega$. 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 $(\omega, p)$, the variance $\operatorname{Var}(\omega, p)$ describes the extent to which the observable $p$ differs from the expected value $\omega \vDash p$, that is, how much much $p$ varies or is spread out. Together, $\omega \vDash p$ and $\operatorname{Var}(\omega, p)$ 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 ( $\omega, p_{1}$ ) and ( $\omega, p_{2}$ ), with (possibly) different observables $p_{1}, p_{2}: X \rightarrow \mathbb{R}$, but with the same shared state $\omega \in \mathcal{D}(X)$;
2 a joint state $\tau \in \mathcal{D}\left(X_{1} \times X_{2}\right)$ together with two observables $q_{1}: X_{1} \rightarrow \mathbb{R}$ and $q_{2}: X_{2} \rightarrow \mathbb{R}$ on the two components $X_{1}, X_{2}$. Via weakening of the observables we get two random variables:

$$
\left(\tau, q_{1} \otimes \mathbf{1}\right) \quad\left(\tau, \mathbf{1} \otimes q_{2}\right)
$$

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 short chapter first introduces the basic definitions and results for variance, and for covariance in shared-state form. They are applied to draw distributions, from Chapter3, 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 the next chapter. In presence of such correlation, updating in one (product) component has crossover influence in the other component.

At the end of this chapter we use what we have seen about variance, to formulate what is called the weak law of large numbers. It shows that by accumulating repeated draws from a distribution one comes arbitrary close to that distribution. This is an alternative way of expressing the denseness of fractional distributions among all distributions, as formulated in Theorem4.5.9.

### 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 $\vDash$ and the operations on observables from Section 4.2. Recall that we understand a random variable here as a pair ( $\omega, p$ ) consisting of a state $\omega \in \mathcal{D}(X)$ and an observable $p: X \rightarrow \mathbb{R}$. The validity $\omega \vDash p$ is a real number, and can thus be used as a scalar, in the sense of Section 4.2 The truth predicate forms an observable $\mathbf{1} \in \operatorname{Obs}(X) ;$ scalar multiplication yields a new observable $(\omega \vDash p) \cdot \mathbf{1} \in \operatorname{Obs}(X)$.

It can be subtracted ${ }^{1}$ from $p$, and then squared, giving an observable:

$$
(p-(\omega \vDash p) \cdot \mathbf{1})^{2}=(p-(\omega \vDash p) \cdot \mathbf{1}) \&(p-(\omega \vDash p) \cdot \mathbf{1}) \in \operatorname{Obs}(X) .
$$

This observable denotes the function that sends $x \in X$ to $(p(x)-(\omega \vDash p))^{2} \in$ $\mathbb{R}_{\geq 0}$. It is thus a factor. Its validity in the original state $\omega$ 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 $(\omega, p)$, the variance $\operatorname{Var}(\omega, p)$ is the non-negative number defined by:

$$
\operatorname{Var}(\omega, p):=\omega \vDash(p-(\omega \vDash p) \cdot \mathbf{1})^{2}
$$

When the underlying sample space $X$ is a subset of $\mathbb{R}$, say via an obvious inclusion function incl: $X \hookrightarrow \mathbb{R}$, we simply write $\operatorname{Var}(\omega)$ for $\operatorname{Var}(\omega$, incl).

The name standard deviation is used for the square root of the variance; thus:

$$
\operatorname{StDev}(\omega, p):=\sqrt{\operatorname{Var}(\omega, p)}
$$

Example 5.1.2. 1 We recall Example 4.1.4 (1), with distribution $\operatorname{flip}\left(\frac{3}{10}\right)=$ $\frac{3}{10}|1\rangle+\frac{7}{10}|0\rangle$ and observable $v(0)=-50$ and $v(1)=100$. We had $\omega \vDash v=$ -5 , and so we get:

$$
\begin{aligned}
\operatorname{Var}\left(f l i p\left(\frac{3}{10}\right), v\right) & =\sum_{x \in\{0,1\}} \operatorname{flip}\left(\frac{3}{10}\right)(x) \cdot(v(x)+5)^{2} \\
& =\frac{3}{10} \cdot(100+5)^{2}+\frac{7}{10} \cdot(-50+5)^{2}=4725
\end{aligned}
$$

The standard deviation is around 68.7.
2 For a (fair) dice we have pips $=\{1,2,3,4,5,6\} \hookrightarrow \mathbb{R}$ and mean $($ dice $)=\frac{7}{2}$ so that:

$$
\begin{aligned}
\operatorname{Var}(\text { dice }) & =\sum_{x \in \operatorname{pips}} \operatorname{dice}(x) \cdot\left(x-\frac{7}{2}\right)^{2} \\
& =\frac{1}{6} \cdot\left(\left(\frac{5}{2}\right)^{2}+\left(\frac{3}{2}\right)^{2}+\left(\frac{1}{2}\right)^{2}+\left(\frac{1}{2}\right)^{2}+\left(\frac{3}{2}\right)^{2}+\left(\frac{5}{2}\right)^{2}\right)=\frac{35}{12}
\end{aligned}
$$

Via a suitable shift-and-rescale one can standardise an observable so that its validity becomes 0 and its variance becomes 1 , see Exercise 5.1.11.

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.
${ }^{1}$ Subtraction expressions like these occur more frequently in mathematics. For instance, an eigenvalue $\lambda$ 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 4.2.14

Lemma 5.1.3. Variance satisfies:

$$
\operatorname{Var}(\omega, p)=\left(\omega \vDash p^{2}\right)-(\omega \vDash p)^{2}
$$

Proof. We have:
$\operatorname{Var}(\omega, p)$
$=\omega \vDash(p-(\omega \vDash p) \cdot \mathbf{1})^{2}$
$=\sum_{x \in X} \omega(x) \cdot(p(x)-(\omega \models p))^{2}$
$=\sum_{x \in X} \omega(x) \cdot\left(p(x)^{2}-2(\omega \vDash p) \cdot p(x)+(\omega \vDash p)^{2}\right)$
$=\left(\sum_{x \in X} \omega(x) p^{2}(x)\right)-2(\omega \vDash p) \cdot\left(\sum_{x \in X} \omega(x) \cdot p(x)\right)+\left(\sum_{x \in X} \omega(x) \cdot(\omega \vDash p)^{2}\right)$
$=\left(\omega \vDash p^{2}\right)-2(\omega \vDash p) \cdot(\omega \vDash p)+(\omega \vDash p)^{2}$
$=\left(\omega \vDash p^{2}\right)-(\omega \vDash p)^{2}$.
Corollary 5.1.4. 1 For a distribution $\omega \in \mathcal{D}(X)$ and an observable $p$ on $X$, there is an inequality:

$$
\omega \vDash p^{2} \geq(\omega \vDash p)^{2} .
$$

2 For a channel $c: X \mapsto Y$ and an observable $q$ on $Y$ there is a pointwise inequality (see Subsection 4.2.1):

$$
c=\ll q^{2} \geq(c=\ll q)^{2} \quad \text { i.e. } \quad c=\ll(q \& q) \geq(c=\ll q) \&(c=\ll q) .
$$

Proof. The first 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 4.2.8(2).

For the second item, let $x \in X$ be arbitrary. Then:

$$
\begin{array}{rlrl}
(c \approx<(q \& q))(x) & =c(x) \vDash q^{2} & & \text { see Definition4.3.1 } \\
& \geq(c(x) \vDash q)^{2} & & \text { by the previous item } \\
& =((c=\ll q)(x)) \cdot((c=\ll q)(x)) & & \\
& =((c=\ll q) \&(c=<q))(x) . &
\end{array}
$$

This result is used to obtain the variances of a draw distributions in a later section. Also, it is used in the following example in which discrete distributions are introduced that approximate continuous 'normal' distributions, see Section??.

Example 5.1.5. We fix a number $N \in \mathbb{N}_{>0}$ and form the set:

$$
S_{N}:=\{0,1, \ldots, 2 N\}=\{0, \ldots, N-1, N, N+1, \ldots, 2 N\} .
$$

It has $2 N+1$ elements, with $N$ sitting in the middle.
For $K \in \mathbb{N}_{>0}$ we consider the average function:

$$
\left(S_{N}\right)^{K} \xrightarrow{\operatorname{avg}[K]} \mathbb{Q} \quad \text { where } \quad \operatorname{avg}[K]\left(i_{1}, \ldots, i_{K}\right):=\frac{i_{1}+\cdots+i_{K}}{K} .
$$

We define a sequence of distributions $\omega_{K} \in \mathcal{D}(\mathbb{Q})$, as:

$$
\begin{align*}
\omega_{K} & :=\mathcal{D}(\operatorname{avg}[K])\left(\text { unif }_{S_{N}} \otimes \cdots \otimes \text { unif }_{S_{N}}\right) \\
& =\frac{\text { unif }_{S_{N}}+\cdots+\text { unif }_{S_{N}}}{K} \quad \text { in the style of Proposition 2.7.2, } \tag{5.1}
\end{align*}
$$

The support of this distribution $\omega_{K}$ is given by the set of numbers from 0 to $2 N$, with steps of $\frac{1}{K}$. Thus:

$$
\left\{0, \frac{1}{K}, \ldots, N-\frac{1}{K}, N, N+\frac{1}{K}, \ldots, 2 N-\frac{1}{K}, 2 N\right\} .
$$

It has $2 K \cdot N+1$ 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 ${ }^{2}$ 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 can now derive the mean and variance of the distributions $\omega_{K} \in \mathcal{D}\left(S_{N}\right)$, namely:

$$
\begin{equation*}
\operatorname{mean}\left(\omega_{K}\right)=N \quad \text { and } \quad \operatorname{Var}\left(\omega_{K}\right)=\frac{N \cdot(N+1)}{3 K} \tag{5.2}
\end{equation*}
$$

The first equation is as expected. The proof involves the inclusion function incl : $S_{N} \rightarrow \mathbb{R}$ and uses Proposition 1.2.6 (3):

$$
\begin{aligned}
\operatorname{mean}\left(\omega_{K}\right)=\omega_{K} \vDash \text { incl } & =\sum_{0 \leq i_{1}, \ldots, i_{K} \leq 2 N}\left(\text { unif }_{S_{N}}\right)^{K}\left(i_{1}, \ldots, i_{K}\right) \cdot \operatorname{avg}\left(i_{1}, \ldots, i_{K}\right) \\
& =\sum_{0 \leq i_{1}, \ldots, i_{K} \leq 2 N} \frac{1}{(2 N+1)^{K}} \cdot \frac{i_{1}+\cdots+i_{K}}{K} \\
& =\frac{1}{(2 N+1)^{K} \cdot K} \cdot \frac{K \cdot 2 N \cdot(2 N+1)^{K}}{2}=N .
\end{aligned}
$$

To prove the variance equation in 5.2 we use Lemma 5.1 .3 and Proposi-

[^7]

Figure 5.1 Averages $\omega_{K}$ 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 $\frac{1}{K} \cdot \operatorname{Norm}(N, \sqrt{N \cdot(N+1) / 3 K})$, see Section ??.
tion 1.2.6(4):

$$
\begin{aligned}
\operatorname{Var}\left(\omega_{K}\right) & =\left(\omega_{K} \vDash \text { incl }^{2}\right)-\left(\omega_{K} \vDash \text { incl }\right)^{2} \\
& =\sum_{0 \leq i_{1}, \ldots, i_{K} \leq 2 N} \frac{1}{(2 N+1)^{K}} \cdot\left(\frac{i_{1}+\cdots+i_{K}}{K}\right)^{2}-N^{2} \\
& =\frac{1}{(2 N+1)^{K} \cdot K^{2}} \cdot \frac{K \cdot 2 N \cdot(2 N+1)^{K} \cdot((3 K+1) \cdot 2 N+2)}{12}-N^{2} \\
& =\frac{N \cdot((3 K+1) \cdot 2 N+2)-6 K \cdot N^{2}}{6 K} \\
& =\frac{2 N^{2}+2 N}{6 K}=\frac{N \cdot(N+1)}{3 K} .
\end{aligned}
$$

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 shifted to an arbitrary position via a translation function.

We continue with covariance and correlation, which involve two random variables, instead one, as for variance. We can distinguish situations, namely:

- The two random variables are of the form $\left(\omega, p_{1}\right)$ and $\left(\omega, p_{2}\right)$, where they share their state $\omega$.
- There is a joint state $\tau \in \mathcal{D}\left(X_{1} \times X_{2}\right)$ together with two observables $q_{1}: X_{1} \rightarrow$ $\mathbb{R}$ and $q_{2}: X_{2} \rightarrow \mathbb{R}$ 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: $\pi_{1}=\ll q_{1}=q_{1} \otimes \mathbf{1}$ and $\pi_{2}=\ll q_{2}=\mathbf{1} \otimes q_{2}$. In this way we obtain two random variable with a shared state:

$$
\left(\tau, \pi_{1}=\ll q_{1}\right) \quad \text { and } \quad\left(\tau, \pi_{2}=\ll q_{2}\right)
$$

These observable transformations $\pi_{i}=\ll q_{i}$ along a deterministic channel $\pi_{i}: X_{1} \times X_{2} \rightarrow X_{i}$ can also be described simply as function composition $q_{i} \circ \pi_{i}: X_{1} \times X_{2} \rightarrow \mathbb{R}$, see Lemma 4.3.2 (9).

We start with the situation in the first bullet above, and deal with the second bullet in Definition 5.3.1 in Section5.3

Definition 5.1.6. Let $\left(\omega, p_{1}\right)$ and $\left(\omega, p_{2}\right)$ 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:

$$
\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right):=\omega \vDash\left(p_{1}-\left(\omega \vDash p_{1}\right) \cdot \mathbf{1}\right) \&\left(p_{2}-\left(\omega \vDash p_{2}\right) \cdot \mathbf{1}\right) .
$$

2 The correlation between $\left(\omega, p_{1}\right)$ and $\left(\omega, p_{2}\right)$ is the covariance divided by their standard deviations:

$$
\operatorname{Cor}\left(\omega, p_{1}, p_{2}\right):=\frac{\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)}{\operatorname{StDev}\left(\omega, p_{1}\right) \cdot \operatorname{StDev}\left(\omega, p_{2}\right)}
$$

Notice that variance $\operatorname{Var}(\omega, p)$ is a special case of covariance $\operatorname{Cov}(\omega, p, p)$, namely with equal observables. Hence if there is an inclusion incl: $X \hookrightarrow \mathbb{R}$ and we would use this inclusion twice to compute covariance, we are in fact computing variance.

Correlation is normalised covariance, so that the outcome is in the interval $[-1,1]$, see Exercise 5.1 .12 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. 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 5.1.7. Covariance can be reformulated as:

$$
\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)=\left(\omega \vDash p_{1} \& p_{2}\right)-\left(\omega \vDash p_{1}\right) \cdot\left(\omega \vDash p_{2}\right)
$$

Example 5.1.8. 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.

$$
a=[5,10,15,20,25] \quad b=[10,8,10,15,12]
$$

We will identify $a$ and $b$ with random variables, namely with $a, b: \mathbf{5} \rightarrow$ $\mathbb{R}$, 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 $f_{5}$ on the set 5 :

$$
\operatorname{avg}(a)=\operatorname{unif}_{5} \vDash a=15 \quad \operatorname{avg}(a)=\text { unif }_{5} \vDash a=11
$$

We will calculate the covariance between $a$ and $b$ wrt. the uniform state $u_{n i f_{5}}$, as:

$$
\begin{aligned}
\operatorname{Cov}\left(\text { unif }_{\mathbf{5}}, a, b\right) & =\text { unif }_{5} \vDash\left(a-\left(\text { unif }_{5} \models a\right) \cdot \mathbf{1}\right) \&\left(b-\left(\text { unif }_{\mathbf{5}} \vDash b\right) \cdot \mathbf{1}\right) \\
& =\sum_{i} \frac{1}{5} \cdot(a(i)-15) \cdot(b(i)-11) \\
& =11 .
\end{aligned}
$$

2 In order to obtain the correlation between $a, b$, we first need to compute their variances:

$$
\begin{aligned}
& \operatorname{Var}\left(\text { unif }_{5}, a\right)=\sum_{i} \frac{1}{5}(a(i)-15)^{2}=50 \\
& \operatorname{Var}\left(\text { unif }_{5}, b\right)=\sum_{i} \frac{1}{5}(b(i)-11)^{2}=5.6
\end{aligned}
$$

Then:

$$
\operatorname{Cor}\left(\text { unif }_{5}, a, b\right)=\frac{\operatorname{Cov}\left(\text { unif }_{5}, a, b\right)}{\sqrt{\operatorname{Var}\left(\text { unif }_{5}, a\right)} \cdot \sqrt{\operatorname{Var}\left(\text { unif }_{5}, b\right)}}=\frac{11}{\sqrt{50} \cdot \sqrt{5.6}} \approx 0.66
$$

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 5.1.9. Consider a state $\omega \in \mathcal{D}(X)$, with observables $p, p_{1}, p_{2} \in$ $\operatorname{Obs}(X)$ and numbers $r, s \in \mathbb{R}$.

1 Covariance satisfies:

$$
\begin{aligned}
\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right) & =\operatorname{Cov}\left(\omega, p_{2}, p_{1}\right) \\
\operatorname{Cov}\left(\omega, p_{1}, \mathbf{1}\right) & =0 \\
\operatorname{Cov}\left(\omega, r \cdot p_{1}, p_{2}\right) & =r \cdot \operatorname{Cov}\left(\omega, p_{1}, p_{2}\right) \\
\operatorname{Cov}\left(\omega, p, p_{1}+p_{2}\right) & =\operatorname{Cov}\left(\omega, p, p_{1}\right)+\operatorname{Cov}\left(\omega, p, p_{2}\right) \\
\operatorname{Cov}\left(\omega, p_{1}+r \cdot \mathbf{1}, p_{2}+s \cdot \mathbf{1}\right) & =\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right) .
\end{aligned}
$$

2 Variance satisfies:

$$
\begin{aligned}
\operatorname{Var}(\omega, r \cdot p) & =r^{2} \cdot \operatorname{Var}(\omega, p) \\
\operatorname{Var}(\omega, p+r \cdot \mathbf{1}) & =\operatorname{Var}(\omega, p) \\
\operatorname{Var}\left(\omega, p_{1}+p_{2}\right) & =\operatorname{Var}\left(\omega, p_{1}\right)+2 \cdot \operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)+\operatorname{Var}\left(\omega, p_{2}\right)
\end{aligned}
$$

3 For correlation we have:

$$
\operatorname{Cor}\left(\omega, p_{1}, p_{2}\right)=\operatorname{Cor}\left(\omega, p_{2}, p_{1}\right)
$$

$\operatorname{Cor}\left(\omega, r \cdot p_{1}, s \cdot p_{2}\right)= \begin{cases}\operatorname{Cor}\left(\omega, p_{1}, p_{2}\right) & \text { if } r, \text { s have the same sign } \\ -\operatorname{Cor}\left(\omega, p_{1}, p_{2}\right) & \text { otherwise. }\end{cases}$
$\operatorname{Cor}\left(\omega, p_{1}+r \cdot \mathbf{1}, p_{2}+s \cdot \mathbf{1}\right)=\operatorname{Cor}\left(\omega, p_{1}, p_{2}\right)$.
Proof. 1 Obviously, covariance is symmetric and covariance with truth is 0 . Covariance preserves scalar multiplication in each (observable) argument since by Lemma 5.1.7.

$$
\begin{aligned}
\operatorname{Cov}\left(\omega, r \cdot p_{1}, p_{2}\right) & =\left(\omega \vDash\left(r \cdot p_{1}\right) \& p_{2}\right)-\left(\omega \vDash r \cdot p_{1}\right) \cdot\left(\omega \vDash p_{2}\right) \\
& =r \cdot\left(\omega \vDash p_{1} \& p_{2}\right)-r \cdot\left(\omega \vDash p_{1}\right) \cdot\left(\omega \vDash p_{2}\right) \\
& =r \cdot \operatorname{Cov}\left(\omega, p_{1}, p_{2}\right) .
\end{aligned}
$$

For preservation of sums we reason from the definition:

$$
\begin{aligned}
& \operatorname{Cov}\left(\omega, p, p_{1}+p_{2}\right) \\
& =\left(\omega \vDash p \&\left(p_{1}+p_{2}\right)\right)-(\omega \vDash p) \cdot\left(\omega \vDash p_{1}+p_{2}\right) \\
& =\left(\omega \vDash\left(p \& p_{1}\right)+\left(p \& p_{2}\right)\right)-(\omega \vDash p) \cdot\left(\left(\omega \vDash p_{1}\right)+\left(\omega \vDash p_{2}\right)\right) \\
& =\left(\omega \vDash p \& p_{1}\right)+\left(\omega \vDash p \& p_{2}\right) \\
& \quad-(\omega \vDash p) \cdot\left(\omega \vDash p_{1}\right)-(\omega \vDash p) \cdot\left(\omega \vDash p_{2}\right) \\
& =\operatorname{Cov}\left(\omega, p, p_{1}\right)+\operatorname{Cov}\left(\omega, p, p_{2}\right)
\end{aligned}
$$

The equation $\operatorname{Cov}\left(\omega, p_{1}+r \cdot \mathbf{1}, p_{2}+s \cdot \mathbf{1}\right)=\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)$ follows from the previous equations.
2 The first property holds by what we have just seen:

$$
\operatorname{Var}(\omega, r \cdot p)=\operatorname{Cov}(\omega, r \cdot p, r \cdot p)=r^{2} \cdot \operatorname{Cov}(\omega, p, p)=r^{2} \cdot \operatorname{Var}(\omega, p)
$$

Similarly, $\operatorname{Var}(\omega, p+r \cdot \mathbf{1})=\operatorname{Var}(\omega, p)$. Next:

$$
\begin{aligned}
& \operatorname{Var}\left(\omega, p_{1}+p_{2}\right) \\
& =\operatorname{Cov}\left(\omega, p_{1}+p_{2}, p_{1}+p_{2}\right) \\
& =\operatorname{Cov}\left(\omega, p_{1}+p_{2}, p_{1}\right)+\operatorname{Cov}\left(\omega, p_{1}+p_{2}, p_{2}\right) \\
& =\operatorname{Cov}\left(\omega, p_{1}, p_{1}\right)+\operatorname{Cov}\left(\omega, p_{2}, p_{1}\right)+\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)+\operatorname{Cov}\left(\omega, p_{2}, p_{2}\right) \\
& =\operatorname{Var}\left(\omega, p_{1}\right)+2 \cdot \operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)+\operatorname{Var}\left(\omega, p_{2}\right) .
\end{aligned}
$$

3 Symmetry of correlation is obvious. By unpacking the definition of correlation and using the previous two items we get:

$$
\begin{aligned}
\operatorname{Cor}\left(\omega, r \cdot p_{1}, s \cdot p_{2}\right) & =\frac{\operatorname{Cov}\left(\omega, r \cdot p_{1}, s \cdot p_{2}\right)}{\sqrt{\operatorname{Var}\left(\omega, r \cdot p_{1}\right)} \cdot \sqrt{\operatorname{Var}\left(\omega, s \cdot p_{2}\right)}} \\
& =\frac{r \cdot s \cdot \operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)}{\sqrt{r^{2} \cdot \operatorname{Var}\left(\omega, p_{1}\right)} \cdot \sqrt{s^{2} \cdot \operatorname{Var}\left(\omega, p_{2}\right)}} \\
& =\frac{r \cdot s \cdot \operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)}{|r| \cdot \sqrt{\operatorname{Var}\left(\omega, p_{1}\right)} \cdot|s| \cdot \sqrt{\operatorname{Var}\left(\omega, p_{2}\right)}} \\
& = \begin{cases}\operatorname{Cor}\left(\omega, p_{1}, p_{2}\right) & \text { if } r, s \text { have the same sign } \\
-\operatorname{Cor}\left(\omega, p_{1}, p_{2}\right) & \text { otherwise. }\end{cases}
\end{aligned}
$$

(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 $\operatorname{Cor}\left(\omega, p_{1}+r \cdot \mathbf{1}, p_{2}+s \cdot \mathbf{1}\right)=\operatorname{Cor}\left(\omega, p_{1}, p_{2}\right)$ holds since both variance and covariance are closed under addition of constants.

## Exercises

5.1.1 Let $\omega$ be a state and $p$ be a factor on the same set. Define for $v \in \mathbb{R}_{\geq 0}$,

$$
f(v):=\omega \vDash(p-v \cdot \mathbf{1})^{2} .
$$

Show that the function $f: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ takes its minimum value at $\omega \vDash p$.
5.1.2 Let $\omega \in \mathcal{D}(X)$ and $p \in \operatorname{Obs}(X)$. Prove that:

$$
\operatorname{Var}(\omega, p)=0 \Longleftrightarrow p \text { is constant on the subset } \operatorname{supp}(\omega) \subseteq X .
$$

5.1.3 Let $\tau \in \mathcal{D}(X \times Y)$ be a joint state with an observable $p: X \rightarrow \mathbb{R}$. We can turn them into a random variable in two ways, by marginalisation and weakening:

$$
(\tau[1,0], p) \quad \text { and } \quad\left(\tau, \pi_{1}=\ll p\right),
$$

where $\pi_{1}=\ll p=p \otimes \mathbf{1}$ is an observable on $X \times Y$.

These two random variables have the same expected value, by (4.7). Show that they have the same variance too:

$$
\operatorname{Var}(\tau[1,0], p)=\operatorname{Var}\left(\tau, \pi_{1}=\ll p\right) .
$$

As a result, the standard deviations are also the same.
5.1.4 Define $g(r):=\operatorname{flip}(r) \otimes$ flip $(1-r) \in \mathcal{D}(\mathbf{2} \times \mathbf{2})$ and show that for each $r \in[0,1]$,

$$
\operatorname{Cov}(g(r))=\operatorname{Cov}\left(g(r), \pi_{1}, \pi_{2}\right)=0 .
$$

5.1.5 Use Jensen's inequality, see Lemma 2.8.3, to prove the inequality $\omega \models$ $p^{2} \geq(\omega \vDash p)^{2}$ in Corollary 5.1.4 (1).
5.1.6 Let $\lambda \in \mathbb{R}_{>0}$.

1 Prove that $\sum_{k \geq 1} \operatorname{pois}[\lambda](k) \cdot k \cdot(k-1)=\lambda^{2}$;
2 Use this equation to show that $\sum_{k \geq 1} \operatorname{pois}[\lambda](k) \cdot k^{2}=\lambda^{2}+\lambda$;
3 Deduce that $\operatorname{Var}(\operatorname{pois}[\lambda])=\lambda$.
5.1.7 Prove Lemma 5.1.7 along the lines of the proof of Lemma 5.1.3
5.1.8 Show for a predicate $p$,
$1 \quad \operatorname{Cov}\left(\omega, p, p^{\perp}\right)=-\operatorname{Var}(\omega, p) \leq 0 ;$
$2 \operatorname{Var}\left(\omega, p^{\perp}\right)=\operatorname{Var}(\omega, p)$.
5.1.9 Let $h: X \rightarrow Y$ be a function, with a state $\omega \in \mathcal{D}(X)$ on its domain an an observable $q: Y \rightarrow \mathbb{R}$ on its codomain. Show that:

$$
\operatorname{Var}(\omega, q \circ h)=\operatorname{Var}(\mathcal{D}(h)(\omega), q) .
$$

Hint: Recall Exercise 4.3 .4
5.1.10 Consider a distribution $\omega \in \mathcal{D}(X)$, a channel $c: X \leadsto Y$ and an observable $q: Y \rightarrow \mathbb{R}$. Prove the 'law of total variance', in analogy with the law of total expectation in Exercise 4.2.7.

$$
\operatorname{Var}(c \gg=\omega, q)=(\omega \vDash \operatorname{Var}(c(-), q))+\operatorname{Var}(\omega, c=\Omega) .
$$

Conclude that $\operatorname{Var}(c \gg=\omega, q) \geq \operatorname{Var}(\omega, c=\ll q)$.
5.1.11 Let $(\omega, p)$ be a random variable on a space $X$. Define a new standard score observable $\operatorname{StSc}(\omega, p): X \rightarrow \mathbb{R}$ by:

$$
\operatorname{StSc}(\omega, p)(x):=\frac{p(x)-(\omega \vDash p)}{\operatorname{StDev}(\omega, p)} .
$$

This pair of $\omega$ with $\operatorname{StSc}(\omega, p)(x)$ is also called the Z-random variable. It is normalised in the sense that:
$1 \omega \vDash \operatorname{StSc}(\omega, p)=0$;
$2 \operatorname{Var}(\omega, \operatorname{StSc}(\omega, p))=\operatorname{StDev}(\omega, \operatorname{StSc}(\omega, p))=1$.
Prove these two items.
5.1.12 Recall the Cauchy-Schwarz inequality, for real numbers $a_{i}, b_{i} \in \mathbb{R}$,

$$
\left(\sum_{i} a_{i} b_{i}\right)^{2} \leq\left(\sum_{i} a_{i}^{2}\right) \cdot\left(\sum_{i} b_{i}^{2}\right)
$$

Use this inequality to prove that correlation is in the interval $[-1,1]$.
5.1.13 Let $\omega \in \mathcal{D}(X)$ be a state with an $n$-test $\vec{p}=p_{1}, \ldots, p_{n}$, see Definition 4.2.2

1 Define the (symmetric) covariance matrix as:

$$
\operatorname{CovMat}(\omega, \vec{p}):=\left(\begin{array}{ccc}
\operatorname{Cov}\left(\omega, p_{1}, p_{1}\right) & \cdots & \operatorname{Cov}\left(\omega, p_{1}, p_{n}\right) \\
\vdots & \vdots \\
\operatorname{Cov}\left(\omega, p_{n}, p_{1}\right) & \cdots & \operatorname{Cov}\left(\omega, p_{n}, p_{n}\right)
\end{array}\right)
$$

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:

$$
v:=\left(\begin{array}{c}
\omega \vDash p_{1} \\
\vdots \\
\omega \vDash p_{n}
\end{array}\right) \quad A:=\left(\begin{array}{c}
\omega \vDash p_{1} \& p_{1} \cdots \omega \\
\vdots \\
\omega \vDash p_{n} \& p_{1} \cdots \\
\cdots \vDash p_{1} \& \\
\vdots
\end{array}\right)
$$

Check that $\operatorname{CovMat}(\omega, \vec{p})=A-v \cdot v^{T}$, where $(-)^{T}$ is transpose.
5.1.14 In linear regression a finite collection $\left(a_{i}, b_{i}\right)_{1 \leq i \leq n}$ 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 $y=v x+w$ that best approximates these points. The error that is minimised is the 'sum of squared residuals' given as:

$$
f(v, w):=\sum_{i}\left(b_{i}-\left(v a_{i}+w\right)\right)^{2}
$$

We redescribe the $a_{i}, b_{i}$ as observables $a, b:\{1,2, \ldots, n\} \rightarrow \mathbb{R}$, 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:

$$
\bar{a}:=\frac{1}{n} \sum_{i} a_{i} \quad \bar{b}:=\frac{1}{n} \sum_{i} b_{i}
$$

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 $\left(a_{i}, b_{i}\right)$ via coefficients given by the familiar formulas:

$$
\hat{v}=\frac{\sum_{i} a_{i}\left(b_{i}-\bar{b}\right)}{\sum_{i} a_{i}\left(a_{i}-\bar{a}\right)} \quad \hat{w}=\bar{b}-\hat{v} \bar{a} .
$$

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:

$$
\hat{v}=\frac{\sum_{i}\left(a_{i}-\bar{a}\right)\left(b_{i}-\bar{b}\right)}{\sum_{i}\left(a_{i}-\bar{a}\right)^{2}} .
$$

3 Check that this yields:

$$
\hat{v}=\frac{\operatorname{Cov}(\text { unif, } a, b)}{\operatorname{Var}(\text { unif }, a)} .
$$

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 5.1.11, namely as:

$$
b-(\hat{v} \cdot a+\hat{w} \cdot \mathbf{1})=\operatorname{StDev}(b) \cdot(\operatorname{StSc}(b)-\operatorname{Cor}(a, b) \cdot \operatorname{StSc}(a)),
$$

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 $\left(a_{i}, b_{i}\right)$. 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 ?? 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 for draw distributions multinomial, hypergeometric and Pólya, but also of Poisson multinomials. We continue the approach of Section 4.4 and use inclusion functions incl : $\mathcal{N}[K](X) \hookrightarrow$ $\mathcal{M}(X)$, 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). In the literature one sometimes finds descriptions of such variances as vectors, but they presuppose an ordering on the points of the underlying space. The multiset description given below does not require such an ordering. It forms an analogue of Proposition 4.4.1.

Proposition 5.2.1. 1 For a distribution $\omega \in \mathcal{D}(X)$,

$$
\operatorname{Var}(m n[K](\omega))=K \cdot \omega \cdot(\mathbf{1}-\omega) \in \mathcal{M}(X)
$$

2 For a non-empty urn $v \in \mathcal{N}[L](X)$ of size $L \geq K$,

$$
\operatorname{Var}(h g[K](v))=K \cdot \frac{L-K}{L-1} \cdot \operatorname{Flrn}(v) \cdot(\mathbf{1}-\operatorname{Flrn}(v)) \in \mathcal{M}(X)
$$

3 For a non-empty urn $v \in \mathcal{N}(X)$,

$$
\operatorname{Var}(p l[K](v))=K \cdot \frac{L+K}{L+1} \cdot \operatorname{Flrn}(v) \cdot(\mathbf{1}-\operatorname{Flrn}(v))
$$

4 For a distribution $\omega \in \mathcal{D}(X)$ and a rate $\lambda \in \mathbb{R}_{>0}$,

$$
\operatorname{Var}(\operatorname{Pmn}[\lambda](\omega))=\lambda \cdot \omega \in \mathcal{M}(X)
$$

Proof. 1 We use the formulation of Lemma 5.1.3 and compute in $\mathcal{M}(X)$,

$$
\begin{aligned}
\operatorname{Var}(m n[K](\omega)) & =\left(m n[K](\omega) \vDash \text { incl }^{2}\right)-(m n[K](\omega) \vDash \text { incl })^{2} \\
& =\left(\sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \varphi^{2}\right)-\operatorname{mean}(\operatorname{mn}[K](\omega))^{2} \\
& =K \cdot(K-1) \cdot \omega^{2}+K \cdot \omega-(K \cdot \omega)^{2} \\
& \quad \text { by Exercise 3.3.8 and Proposition 4.4.1)(1) } \\
& =K \cdot \omega \cdot(\mathbf{1}-\omega) .
\end{aligned}
$$

2 For a non-empty urn $v \in \mathcal{N}[L](X)$ of size $L \geq K$, we use Exercise 3.4.4 and Proposition 4.4.1 (2).

$$
\begin{aligned}
\operatorname{Var}(h g[K](v)) & =\left(\sum_{\varphi \leq K^{v}} h g[K](v)(\varphi) \cdot \varphi^{2}\right)-\operatorname{mean}(h g[K](v))^{2} \\
& =K \cdot \operatorname{Flrn}(v) \cdot \frac{(K-1) \cdot v+(L-K)}{L-1}-(K \cdot \operatorname{Flrn}(v))^{2} \\
& =K \cdot \operatorname{Flrn}(v) \cdot\left(\frac{L-K}{L-1}+\frac{L \cdot(K-1) \cdot v-(L-1) \cdot K \cdot v}{L \cdot(L-1)}\right) \\
& =K \cdot \operatorname{Flrn}(v) \cdot\left(\frac{L-K}{L-1}-\frac{(L-K) \cdot v}{L \cdot(L-1)}\right) \\
& =K \cdot \frac{L-K}{L-1} \cdot \operatorname{Flrn}(v) \cdot(\mathbf{1}-\operatorname{Flrn}(v)) .
\end{aligned}
$$

3 Similarly, for a non-empty urn $v \in \mathcal{N}[L](X)$, we now use Exercise 3.5.1 and

Proposition 4.4.1 (3).

$$
\begin{aligned}
\operatorname{Var}(p l[K](v)) & =\left(\sum_{\varphi \in \mathcal{N}[K] \operatorname{supp}(v))} p l[K](v)(\varphi) \cdot \varphi^{2}\right)-\operatorname{mean}(p l[K](v))^{2} \\
& =K \cdot \operatorname{Flrn}(v) \cdot \frac{(K-1) \cdot v+(L+K)}{L+1}-(K \cdot F \operatorname{lrn}(v))^{2} \\
& =K \cdot \operatorname{Flrn}(v) \cdot\left(\frac{L+K}{L+1}+\frac{L \cdot(K-1) \cdot v-(L+1) \cdot K \cdot v}{L \cdot(L+1)}\right) \\
& =K \cdot F \operatorname{Fln}(v) \cdot\left(\frac{L+K}{L+1}-\frac{(L+K) \cdot v}{L \cdot(L+1)}\right) \\
& =K \cdot \frac{L+K}{L+1} \cdot \operatorname{Flrn}(v) \cdot(\mathbf{1}-\operatorname{Flrn}(v)) .
\end{aligned}
$$

4 Finally, using Proposition 4.4.1 (4) together with Exercises 3.3.8 and 5.1.6

$$
\begin{aligned}
& \operatorname{Var}(\operatorname{Pmn}[\lambda](\omega)) \\
& =\left(\sum_{\varphi \in \mathcal{N}(X)} \operatorname{Pmn}[\lambda](\omega)(\varphi) \cdot \varphi^{2}\right)-\operatorname{mean}(\operatorname{Pmn}[\lambda](\omega))^{2} \\
& =\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \varphi^{2}\right)-(\lambda \cdot \omega)^{2} \\
& =\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot\left(K \cdot(K-1) \cdot \omega^{2}+K \cdot \omega\right)\right)-\lambda^{2} \cdot \omega^{2} \\
& =\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot K \cdot(K-1)\right) \cdot \omega^{2}+\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot K\right) \cdot \omega-\lambda^{2} \cdot \omega^{2} \\
& =\lambda^{2} \cdot \omega^{2}+\lambda \cdot \omega-\lambda^{2} \cdot \omega^{2} \\
& =\lambda \cdot \omega .
\end{aligned}
$$

Recall from (4.13) the multiset-evaluation observable $\mathrm{e} v_{x}: \mathcal{M}(X) \rightarrow \mathbb{R}_{\geq 0}$ given by $\mathrm{ev}_{x}(\varphi)=\varphi(x)$. The above formulations of variance as multisets can also be written as:

$$
\begin{equation*}
\operatorname{Var}(m n[K](\omega))=\sum_{x \in X} \operatorname{Var}\left(m n[K](\omega), \mathrm{ev}_{x}\right)|x\rangle \in \mathcal{M}(X) . \tag{5.3}
\end{equation*}
$$

(And similarly for $h g[K](v), p l[K](v)$ and $\operatorname{Pmn}[\lambda](\omega)$.)
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 $\omega \in \mathcal{D}(X)$,

$$
\begin{aligned}
\operatorname{Cov}(m n[K](\omega)) & :=\sum_{y, z \in X} \operatorname{Cov}\left(m n[K]\left(\omega, \mathrm{ev}_{y}, \mathrm{ev}_{z}\right)|y, z\rangle\right. \\
& =-K \cdot(\omega \otimes \omega) \in \mathcal{M}(X \times X)
\end{aligned}
$$

2 For an urn $v \in \mathcal{N}[L](X)$ of size $L \geq K$,

$$
\begin{aligned}
\operatorname{Cov}(h g[K](v)) & :=\sum_{y, z \in X} \operatorname{Cov}\left(h g[K](v), \mathrm{ev}_{y}, \mathrm{ev}_{z}\right)|y, z\rangle \\
& =-K \cdot \frac{L-K}{L-1} \cdot(\operatorname{Flrn}(v) \otimes \operatorname{Flrn}(v)) \in \mathcal{M}(X \times X)
\end{aligned}
$$

3 In the Pólya case:

$$
\begin{aligned}
\operatorname{Cov}(p l[K](v)) & :=\sum_{y, z \in X} \operatorname{Cov}\left(p l[K](v), \mathrm{ev}_{y}, \mathrm{ev}_{z}\right)|y, z\rangle \\
& =-K \cdot \frac{L+K}{L+1} \cdot(\operatorname{Flrn}(v) \otimes \operatorname{Flrn}(v)) \in \mathcal{M}(X \times X)
\end{aligned}
$$

4 The Poisson point processes always has zero covariance:

$$
\begin{aligned}
\operatorname{Cov}(\operatorname{Pmn}[\lambda](\omega)): & =\sum_{y, z \in X} \operatorname{Cov}\left(\operatorname{Pmn}[\lambda](\omega), \mathrm{ev}_{y}, \mathrm{ev}_{z}\right)|y, z\rangle \\
& =\mathbf{0} \in \mathcal{M}(X \times X)
\end{aligned}
$$

Proof. 1 Using the covariance formula of Lemma 5.1.7 and Exercise 3.3.8 we compute:

$$
\begin{aligned}
& \operatorname{Cov}\left(m n[K](\omega), \mathrm{ev}_{y}, \mathrm{ev}_{z}\right) \\
& =\left(m n[K](\omega) \vDash \mathrm{ev}_{y} \& \mathrm{ev}_{z}\right)-\left(m n[K](\omega) \vDash \mathrm{ev}_{y}\right) \cdot\left(m n[K](\omega) \vDash \mathrm{ev}_{z}\right) \\
& =K \cdot(K-1) \cdot \omega(y) \cdot \omega(z)-K \cdot \omega(y) \cdot K \cdot \omega(z) \\
& =-K \cdot \omega(y) \cdot \omega(z) . \\
& =-K \cdot(\omega \otimes \omega)(y, z) .
\end{aligned}
$$

2 In the same way, using Exercise 3.4.4 and Proposition 4.4.1 (2):

$$
\begin{aligned}
& \operatorname{Cov}\left(h g[K](v), \mathrm{ev}_{y}, \mathrm{ev}_{z}\right) \\
& =\left(h g[K](v) \vDash \mathrm{ev}_{y} \& \mathrm{ev}_{z}\right)-\left(h g[K](v) \vDash \mathrm{ev}_{y}\right) \cdot\left(h g[K](v) \vDash \mathrm{ev}_{z}\right) \\
& =K \cdot(K-1) \cdot \operatorname{Flrn}(v)(y) \cdot \frac{v(z)}{L-1}-K \cdot \operatorname{Flrn}(v)(y) \cdot K \cdot \operatorname{Flrn}(v)(z) \\
& =K \cdot \operatorname{Flrn}(v)(y) \cdot \frac{L \cdot(K-1) \cdot v(z)-K \cdot(L-1) \cdot v(z)}{L \cdot(L-1)} \\
& =-K \cdot \frac{L-K}{L-1} \cdot \operatorname{Flrn}(v)(y) \cdot \operatorname{Flrn}(v)(z) .
\end{aligned}
$$

3 Similarly, via Exercise 3.5.1 and Proposition 4.4.1 (3):

$$
\begin{aligned}
& \operatorname{Cov}\left(p l[K](v), \mathrm{ev}_{y}, \mathrm{ev}_{z}\right) \\
& =\left(p l[K](v) \vDash \mathrm{ev}_{y} \& \mathrm{ev}_{z}\right)-\left(p l[K](v) \vDash \mathrm{ev}_{y}\right) \cdot\left(p l[K](v) \vDash \mathrm{ev}_{z}\right) \\
& =K \cdot(K-1) \cdot \operatorname{Flrn}(v)(y) \cdot \frac{v(z)}{L+1}-K \cdot \operatorname{Flrn}(v)(y) \cdot K \cdot \operatorname{Flrn}(v)(z) \\
& =K \cdot \operatorname{Flrn}(v)(y) \cdot \frac{L \cdot(K-1) \cdot v(z)-K \cdot(L+1) \cdot v(z)}{L \cdot(L+1)} \\
& =-K \cdot \frac{L+K}{L+1} \cdot \operatorname{Flrn}(v)(y) \cdot \operatorname{Flrn}(v)(z) .
\end{aligned}
$$

4 Via Proposition 4.4.1 (4) together with Exercises 3.3.8 and 5.1.6 we get:

$$
\begin{aligned}
& \operatorname{Cov}\left(\operatorname{Pmn}[\lambda](\omega), \mathrm{ev}_{y}, \mathrm{ev}_{z}\right) \\
&=\left(\operatorname{Pmn}[\lambda](\omega) \vDash \mathrm{ev}_{y} \& \mathrm{ev}_{z}\right)-\left(\operatorname{Pmn}[\lambda](\omega) \vDash \mathrm{ev}_{y}\right) \cdot\left(\operatorname{Pmn}[\lambda](\omega) \vDash \mathrm{ev}_{z}\right) \\
&=\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot\left(\operatorname{mn}[K](\omega) \vDash \mathrm{ev}_{y} \& \mathrm{ev}_{z}\right)\right) \\
& \quad-\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot\left(\operatorname{mn}[K](\omega) \vDash \mathrm{ev}_{y}\right)\right) \\
& \cdot\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot\left(\operatorname{mn}[K](\omega) \models \mathrm{ev} v_{z}\right)\right) \\
&=\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot K \cdot(K-1) \cdot \omega(y) \cdot \omega(z)\right) \\
& \quad-\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot K \cdot \omega(y)\right) \cdot\left(\sum_{K \in \mathbb{N}} \operatorname{pois}[\lambda](K) \cdot K \cdot \omega(z)\right) \\
&= \lambda^{2} \cdot \omega(y) \cdot \omega(z)-\lambda \cdot \omega(y) \cdot \lambda \cdot \omega(z) \\
&=0 .
\end{aligned}
$$

On the diagonal in these joint multisets, where $y=z$, one recovers the variances. When the elements of the space $X$ are ordered, 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 christiandemocractic party $(C)$. We thus have a set of vote options $V=\{G, L, C\}$ with a (natural) voter multiset $v=25|G\rangle+10|L\rangle+15|C\rangle$.

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 4.4.1)(2). It gives the multiset:

$$
\operatorname{mean}(h g[5](v))=5 \cdot \operatorname{Flrn}(v)=\frac{5}{2}|G\rangle+\frac{3}{2}|L\rangle+1|C\rangle .
$$

The deviations of the mean are given by the variance, see Proposition 5.2.1 (2), as a multiset on $V$.

$$
\begin{aligned}
\operatorname{Var}(h g[5](v)) & =5 \cdot \frac{50-5}{50-1} \cdot \sum_{x \in V} \operatorname{Flrn}(v)(x) \cdot(1-\operatorname{Flrn}(v)(x))|x\rangle \\
& =\frac{225}{196}|G\rangle+\frac{27}{28}|L\rangle+\frac{36}{49}|C\rangle \\
& \approx 1.148|G\rangle+0.9643|L\rangle+0.7347|C\rangle .
\end{aligned}
$$

The covariances give a 2-dimensional multiset on the product $V \times V$, see Proposition 5.2.2 (2).

$$
\begin{aligned}
\operatorname{Cov}(\operatorname{hg}[5](v))= & \frac{225}{196}|G, G\rangle-\frac{135}{196}|G, L\rangle-\frac{45}{98}|G, C\rangle \\
& -\frac{135}{196}|L, G\rangle+\frac{27}{28}|L, L\rangle-\frac{27}{98}|L, C\rangle \\
& -\frac{45}{98}|C, G\rangle-\frac{27}{98}|C, L\rangle+\frac{36}{49}|C, C\rangle \\
\approx & 1.148|G, G\rangle-0.6888|G, L\rangle-0.4592|G, C\rangle \\
& -0.6888|L, G\rangle+0.9643|L, L\rangle-0.2755|L, C\rangle \\
& -0.4592|C, G\rangle-0.2755|C, L\rangle+0.7347|C, C\rangle .
\end{aligned}
$$

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 4.4.3 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 $\bar{p}^{+}$interacts well with (co)variance in the multinomial case, like with validity in Proposition 4.4.4 (1).

Proposition 5.2.4. Let $p, q: X \rightarrow \mathbb{R}$ be observables on a set $X$, with their addtive extensions $\bar{p}^{+}, \bar{q}^{+}: \mathcal{N}(X) \rightarrow \mathbb{R}$. For a distribution $\omega \in \mathcal{D}(X)$ and a number $K \in \mathbb{N}$,

1 Variance of $\bar{p}^{+}$over multinomial draws is related to variance of $p$ over $X$, via:

$$
\operatorname{Var}\left(m n[K](\omega), \bar{p}^{+}\right)=K \cdot \operatorname{Var}(\omega, p) .
$$

2 Similarly for covariance:

$$
\operatorname{Cov}\left(m n[K](\omega), \bar{p}^{+}, \bar{q}^{+}\right)=K \cdot \operatorname{Cov}(\omega, p, q)
$$

Proof. 1 By Proposition 4.4.4 (1) and Exercise 3.3.8.

$$
\begin{aligned}
& \operatorname{Var}\left(m n[K](\omega), \bar{p}^{+}\right) \\
& =m n[K](\omega) \vDash \bar{p}^{+} \& \bar{p}^{+}-\left(m n[K](\omega) \vDash \bar{p}^{+}\right)^{2} \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \bar{p}^{+}(\varphi) \cdot \bar{p}^{+}(\varphi)-K^{2} \cdot(\omega \vDash p)^{2} \\
& =\sum_{x, y \in X} p(x) \cdot p(y) \cdot \sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi(x) \cdot \varphi(y)-K^{2} \cdot(\omega \vDash p)^{2} \\
& =\sum_{x, y \in X} p(x) \cdot p(y) \cdot\left\{\begin{array}{l}
K \cdot(K-1) \cdot \omega(x) \cdot \omega(y) \quad \text { if } x \neq y \\
K \cdot(K-1) \cdot \omega(x)^{2}+K \cdot \omega(x) \text { if } x=y
\end{array}\right\}-K^{2} \cdot(\omega \vDash p)^{2} \\
& =K \cdot(K-1) \cdot(\omega \vDash p)^{2}+K \cdot(\omega \vDash p \& p)-K^{2} \cdot(\omega \vDash p)^{2} \\
& =K \cdot(\omega \vDash p \& p)-K \cdot(\omega \vDash p)^{2} \\
& =K \cdot \operatorname{Var}(\omega, p) .
\end{aligned}
$$

2 Similarly.

### 5.2.1 Distributions of validities and of variances

Fix a random variable $(\omega, p)$ on a set $X$ and a number $K$. We can form the multinomial distribution $m n[K](\omega)$ on the set $\mathcal{N}[K](X)$ of natural multisets $\varphi$ of size $K$, over $X$. Applying frequentist learning to such multisets $\varphi$ gives new distributions Flrn $(\varphi) \in \mathcal{D}(X)$, and thus new random variables $(F l r n(\varphi), p)$. We can look a the validity and variance of the latter. This gives what we call distributions of validity and distributions of variance. They are defined as distributions on $\mathbb{R}$ via:

$$
\begin{align*}
\operatorname{dval}[K](\omega, p) & :=\mathcal{D}(F \operatorname{lrn}(-) \vDash p)(m n[K](\omega)) \\
& =\sum_{\varphi \in \mathcal{M}[K](X)} m n[K](\omega)(\varphi)|F \operatorname{lrn}(\varphi) \vDash p\rangle \\
\operatorname{dvar}[K](\omega, p) & :=\mathcal{D}(\operatorname{Var}(F \operatorname{lrn}(-), p))(\operatorname{mn}[K](\omega))  \tag{5.4}\\
& =\sum_{\varphi \in \mathcal{M}[K](X)} \operatorname{mn}[K](\omega)(\varphi)|\operatorname{Var}(\operatorname{Flrn}(\varphi), p)\rangle .
\end{align*}
$$

Such distributions are useful in hypothesis testing in statistics, for instance when $\omega$ is a huge distribution for which we wish we check the validity of the observable (or predicate) $p$. We can then take small samples $\varphi$ from $\omega$ via the multinomial distribution and check the validity of $p$ in the normalised sample $\operatorname{Flrn}(\varphi)$. Proposition 5.2.5 (1) below says that the mean of all such samples equals the validity $\omega \vDash p$.

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:

$$
\omega \vDash p=7 \quad \text { and } \quad \operatorname{Var}(\omega, p)=13 .
$$

The distributions of validities is:

$$
\begin{aligned}
& \left.\left.d v a l[2](\omega, p)=\frac{1}{36} \right\rvert\, \operatorname{Flrn}(2|a\rangle) \vDash p\right\rangle+\frac{1}{6}|F \operatorname{lrn}(1|a\rangle+1|b\rangle) \vDash p\rangle \\
& \left.\left.+\frac{1}{4} \right\rvert\, \operatorname{Flrn}(2|b\rangle) \vDash p\right\rangle+\frac{1}{9}|\operatorname{Flrn}(1|a\rangle+1|c\rangle) \vDash p\rangle \\
& \left.\left.+\frac{1}{3}|\operatorname{Flrn}(1|b\rangle+1|c\rangle) \vDash p\rangle+\frac{1}{9} \right\rvert\, \operatorname{Flrn}(2|c\rangle) \vDash p\right\rangle \\
& \left.\left.\left.\left.=\frac{1}{36}|1| a\right\rangle \vDash p\right\rangle+\frac{1}{6}\left|\frac{1}{2}\right| a\right\rangle+\frac{1}{2}|b\rangle \vDash p\right\rangle \\
& \left.\left.\left.\left.+\frac{1}{4}|1| b\right\rangle \vDash p\right\rangle+\frac{1}{9}\left|\frac{1}{2}\right| a\right\rangle+\frac{1}{2}|c\rangle \vDash p\right\rangle \\
& \left.\left.\left.\left.+\frac{1}{3}\left|\frac{1}{2}\right| b\right\rangle+\frac{1}{2}|c\rangle \vDash p\right\rangle+\frac{1}{9}|1| c\right\rangle \vDash p\right\rangle \\
& =\frac{1}{36}|6\rangle+\frac{1}{6}|5\rangle+\frac{1}{4}|4\rangle+\frac{1}{9}|9\rangle+\frac{1}{3}|8\rangle+\frac{1}{9}|12\rangle
\end{aligned}
$$

It is not hard to see that its mean satisfies:

$$
\frac{1}{36} \cdot 6+\cdots+\frac{1}{9} \cdot 12=7=\omega \vDash p .
$$

As to the variance:

$$
\left(\frac{1}{36} \cdot 6^{2}+\cdots+\frac{1}{9} \cdot 12^{2}\right)-7^{2}=\frac{13}{2}=\frac{1}{2} \cdot \operatorname{Var}(\omega, p) .
$$

These outcomes are in line with Proposition 5.2.5 (1) (2) below.
Next, the distribution of variances is:

$$
\left.\left.\begin{array}{rl}
d \operatorname{var}[2](\omega, p)= & \frac{1}{36} \\
\mid & \operatorname{Var}(1|a\rangle, p)\rangle+\frac{1}{6}\left|\operatorname{Var}\left(\frac{1}{2}|a\rangle+\frac{1}{2}|b\rangle, p\right)\right\rangle \\
& \left.\left.+\frac{1}{4} \right\rvert\, \operatorname{Var}(1|b\rangle, p)\right\rangle+\frac{1}{9}\left|\operatorname{Var}\left(\frac{1}{2}|a\rangle+\frac{1}{2}|c\rangle, p\right)\right\rangle \\
& \left.\left.+\frac{1}{3}\left|\operatorname{Var}\left(\frac{1}{2}|b\rangle+\frac{1}{2}|c\rangle, p\right)\right\rangle+\frac{1}{9} \right\rvert\, \operatorname{Var}(1|c\rangle, p)\right\rangle \\
= & \left.\frac{1}{36} \right\rvert\,
\end{array} 6^{2}-6^{2}\right\rangle+\frac{1}{6}\left|26-5^{2}\right\rangle+\frac{1}{4}\left|4^{2}-4^{2}\right\rangle\right)
$$

The mean $\frac{7}{18} \cdot 0+\cdots+\frac{1}{3} \cdot 16$ of this distribution equals $\frac{13}{2}=\frac{1}{2} \cdot \operatorname{Var}(\omega, p)$. This is an instance of Proposition 5.2.5 (3).

Diagrammatically the definitions 5.4 involve the composites:


The distribution of validities $d v a l[K]$ generalises the distribution of means that
is often used in hypothesis testing in statistics (see e.g. [172]). This distribution of means uses the validiy $(-) \vDash p$ to compute the mean, by taking the inclusion incl : $X \hookrightarrow \mathbb{R}$ as observable, as in Definition4.1.3. 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 3.3.3 which says that transforming a multinomial distribution along frequentist learning yields the orginal (urn) distribution.

Proposition 5.2.5. Let $(\omega, p)$ 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.

$$
\operatorname{mean}(d v a l[K](\omega, p))=\omega \vDash p .
$$

2 The variance of the distribution of validities satisfies:

$$
\operatorname{Var}(d v a l[K](\omega, p))=\frac{\operatorname{Var}(\omega, p)}{K} .
$$

3 The mean of the distribution of variances is:

$$
\operatorname{mean}(d \operatorname{var}[K](\omega, p))=\frac{K-1}{K} \cdot \operatorname{Var}(\omega, p) .
$$

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 Easy, via the following diagrammatic proof.


The rectangle on the right commutes by Exercise 4.1.9(3), and the triangle on the left by Theorem 3.3.3
2 The second item requires more work. Let's assume $\operatorname{supp}(\omega)=\left\{x_{1}, \ldots, x_{n}\right\}$. We first prove the auxiliary result (*) below, via the Multinomial Theorem (1.40), and via Exercise 3.3.8 Usage of this exercise is denoted below by the marked equation $\stackrel{(E)}{=}$.

$$
\begin{equation*}
\sum_{\varphi \in \mathcal{M}[K](X)} m n[K](\omega)(\varphi)(F \operatorname{lrn}(\varphi) \vDash p)^{2}=\frac{(K-1)(\omega \vDash p)^{2}+\left(\omega \vDash p^{2}\right)}{K} . \tag{*}
\end{equation*}
$$

We reason as follows.

$$
\begin{aligned}
& \sum_{\varphi \in \mathcal{M}[K](X)} m n[K](\omega)(\varphi)(F \operatorname{lrn}(\varphi) \vDash p)^{2} \\
& =\sum_{\varphi \in \mathcal{M}[K](X)} m n[K](\omega)(\varphi)\left(\sum_{i} \frac{\varphi\left(x_{i}\right)}{K} \cdot p\left(x_{i}\right)\right)^{2} \\
& \stackrel{[1.40}{-} \frac{1}{K^{2}} \cdot \sum_{\varphi \in \mathcal{M}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \sum_{\psi \in \mathcal{N}[2](\{11, \ldots, n)}(\psi) \cdot \prod_{i}\left(\varphi\left(x_{i}\right) \cdot p\left(x_{i}\right)\right)^{\psi(i)} \\
& =\frac{1}{K^{2}} \cdot\left(2 \sum_{i \neq j} \sum_{\varphi \in \mathcal{M}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi\left(x_{i}\right) \cdot p\left(x_{i}\right) \cdot \varphi\left(x_{j}\right) \cdot p\left(x_{j}\right)\right. \\
& \left.+\sum_{i} \sum_{\varphi \in \mathcal{M}[K](X)} m n[K](\omega)(\varphi) \cdot \varphi\left(x_{i}\right)^{2} \cdot p\left(x_{i}\right)^{2}\right) \\
& \stackrel{(E)}{=} \frac{1}{K^{2}} \cdot\left(2 \sum_{i \neq j} K \cdot(K-1) \cdot \omega\left(x_{i}\right) \cdot p\left(x_{i}\right) \cdot \omega\left(x_{j}\right) \cdot p\left(x_{j}\right)\right. \\
& \left.+\sum_{i} K \cdot(K-1) \cdot \omega\left(x_{i}\right)^{2} \cdot p\left(x_{i}\right)^{2}+K \cdot \omega\left(x_{i}\right) \cdot p\left(x_{i}\right)^{2}\right) \\
& \stackrel{\text { 1.40] }}{=} \frac{K-1}{K} \cdot\left(\sum_{i} \omega\left(x_{i}\right) \cdot p\left(x_{i}\right)\right)^{2}+\frac{1}{K} \cdot\left(\sum_{i} \omega\left(x_{i}\right) \cdot p\left(x_{i}\right)^{2}\right) \\
& =\frac{(K-1)(\omega \vDash p)^{2}+\left(\omega \vDash p^{2}\right)}{K} \text {. }
\end{aligned}
$$

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 (*).

$$
\begin{aligned}
& \operatorname{Var}(\operatorname{dval}[K](\omega, p)) \\
& =\sum_{\varphi \in \mathcal{M}[K](X)} \operatorname{mn}[K](\omega)(\varphi)(\operatorname{Flrn}(\varphi) \vDash p)^{2}-(\operatorname{mean}(\operatorname{dval}[K](\omega, p)))^{2} \\
& \stackrel{(*)}{=} \frac{(K-1)(\omega \vDash p)^{2}+\left(\omega \vDash p^{2}\right)}{K}-(\omega \vDash p)^{2} \\
& =\frac{-(\omega \vDash p)^{2}+\left(\omega \vDash p^{2}\right)}{K} \\
& =\frac{\operatorname{Var}(\omega, p)}{K} .
\end{aligned}
$$

3 We use item (1) and (*).

$$
\begin{aligned}
& \operatorname{mean}(\operatorname{dvar}[K](\omega, p)) \\
& =\sum_{\varphi \in \mathcal{M}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot \operatorname{Var}(\operatorname{Flrn}(\varphi), p) \\
& \left.=\sum_{\varphi \in \mathcal{M}[K](X)} \operatorname{mn}[K](\omega)(\varphi) \cdot\left(F \operatorname{lrn}(\varphi) \vDash p^{2}\right)-(F \operatorname{lrn}(\varphi) \vDash p)^{2}\right) \\
& \stackrel{(*)}{=}\left(\omega \vDash p^{2}\right)-\frac{(K-1)(\omega \vDash p)^{2}+\left(\omega \vDash p^{2}\right)}{K} \\
& =\frac{(K-1)\left(\omega \vDash p^{2}\right)}{K}-\frac{(K-1)(\omega \vDash p)^{2}}{K} \\
& =\frac{K-1}{K} \cdot \operatorname{Var}(\omega, p) .
\end{aligned}
$$

## Exercises

5.2.1 In the distribution of validities (5.4) we have used multinomial samples. We can also use hypergeometric ones. Show that in that case the analogue of Proposition 5.2.5 (1) still holds: for an urn $v \in \mathcal{N}(X)$, an observable $p: X \rightarrow \mathbb{R}$ and a number $K \leq\|v\|$,

$$
\operatorname{mean}(\mathcal{D}(\operatorname{Flrn}(-) \vDash p)(h g[K](v)))=\operatorname{Flrn}(v) \vDash p .
$$

5.2.2 1 Prove Equation (5.3).

2 Formulate and prove similar equations for the distributions $h g[K](v)$, $p l[K](v)$ and $\operatorname{Pmn}[\lambda](\omega)$.

### 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 $\tau \in \mathcal{D}\left(X_{1} \times X_{2}\right)$ be a joint state on sets $X_{1}, X_{2}$ and let $q_{1} \in$
$\operatorname{Obs}\left(X_{1}\right)$ and $q_{2} \in \operatorname{Obs}\left(X_{2}\right)$ be two observables on these two sets separately. In this situation the joint covariance is defined as:

$$
\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right):=\operatorname{Cov}\left(\tau, \pi_{1}=\ll q_{1}, \pi_{2}=\ll q_{2}\right) .
$$

Thus we use the weakenings $\pi_{1} \lll q_{1}=q_{1} \otimes \mathbf{1}$ and $\pi_{2} \lll q_{2}=\mathbf{1} \otimes q_{2}$ 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 Definition5.1.6 applies.

Similarly, the joint correlation is:

$$
\operatorname{JCor}\left(\tau, q_{1}, q_{2}\right):=\operatorname{Cor}\left(\tau, \pi_{1}=\ll q_{1}, \pi_{2}=\ll q_{2}\right)
$$

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 $\operatorname{JCov}(\tau)$ and $\operatorname{JCor}(\tau)$.

Joint covariance can be reformulated in different ways, including in the style that we have seen before, in Lemma 5.1.3 and 5.1.7.

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 $\tau \in \mathcal{D}\left(X_{1} \times X_{2}\right)$, namely as:

$$
\begin{aligned}
\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right) & =\tau \vDash\left(q_{1}-\left(\tau[1,0] \vDash q_{1}\right) \cdot \mathbf{1}\right) \otimes\left(q_{2}-\left(\tau[0,1] \vDash q_{2}\right) \cdot \mathbf{1}\right) . \\
& =\left(\tau \vDash q_{1} \otimes q_{2}\right)-\left(\tau[1,0] \vDash q_{1}\right) \cdot\left(\tau[0,1] \vDash q_{2}\right) .
\end{aligned}
$$

As a consequence, when $\tau$ is non-entwined, the joint covariance is zero, whatever the obervables $q_{1}, q_{2}$ are:

$$
\operatorname{JCov}\left(\tau_{1} \otimes \tau_{2}, q_{1}, q_{2}\right)=0
$$

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 4.3.8.

$$
\begin{aligned}
& \operatorname{JCov}\left(\tau, q_{1}, q_{2}\right)= \operatorname{Cov}\left(\tau, \pi_{1}=\ll q_{1}, \pi_{2}=\ll q_{2}\right) . \\
&= \tau \vDash\left(\left(\pi_{1}=\ll q_{1}\right)-\left(\tau \vDash \pi_{1}=\ll q_{1}\right) \cdot \mathbf{1}\right) \\
& \&\left(\left(\pi_{2}=\ll q_{2}\right)-\left(\tau \vDash \pi_{2}=\ll q_{2}\right) \cdot \mathbf{1}\right) \\
& \stackrel{4.12]}{=} \tau \vDash\left(\left(\pi_{1}=\ll q_{1}\right)-\left(\tau[1,0] \vDash q_{1}\right) \cdot\left(\pi_{1}=\ll \mathbf{1}\right)\right) \\
& \&\left(\left(\pi_{2}=\ll q_{2}\right)-\left(\tau[0,1] \vDash q_{2}\right) \cdot\left(\pi_{2}=\ll \mathbf{1}\right)\right) \\
&= \tau \vDash\left(\pi_{1}=\ll\left(q_{1}-\left(\tau[1,0] \vDash q_{1}\right) \cdot \mathbf{1}\right)\right) \\
& \&\left(\pi_{2} \lll\left(q_{2}-\left(\tau[0,1] \vDash q_{2}\right) \cdot \mathbf{1}\right)\right) \\
&= \tau \vDash\left(q_{1}-\left(\tau[1,0] \vDash q_{1}\right) \cdot \mathbf{1}\right) \otimes\left(q_{2}-\left(\tau[0,1] \vDash q_{2}\right) \cdot \mathbf{1}\right) .
\end{aligned}
$$

Via this first equation in Lemma 5.3.2 we prove the second one:

$$
\begin{aligned}
& \operatorname{JCov}\left(\tau, q_{1}, q_{2}\right) \\
& =\quad \tau \vDash\left(q_{1}-\left(\tau[1,0] \vDash q_{1}\right) \cdot \mathbf{1}\right) \otimes\left(q_{2}-\left(\tau[0,1] \vDash q_{2}\right) \cdot \mathbf{1}\right) \\
& =\sum_{x_{1} \in X_{1}, x_{2} \in X_{2}} \tau\left(x_{1}, x_{2}\right) \cdot\left(q_{1}-\left(\tau[1,0] \vDash q_{1}\right) \cdot \mathbf{1}\right)\left(x_{1}\right) \\
& =\sum_{x_{1} \in X_{1}, x_{2} \in X_{2}} \tau\left(x_{1}, x_{2}\right) \cdot\left(q_{1}\left(x_{1}\right)-\left(\tau[0,1] \vDash q_{2}\right) \cdot \mathbf{1}\right)\left(x_{2}\right) \\
& =\left(\sum_{x_{1} \in X_{1}, x_{2} \in X_{2}} \tau\left(x_{1}, x_{2}\right) \cdot q_{1}\left(x_{1}\right) \cdot q_{2}\left(x_{2}\right)\right) \\
& \quad-\left(\sum_{x_{1} \in X_{1}, x_{2} \in X_{2}} \tau\left(x_{1}, x_{2}\right) \cdot q_{1}\left(x_{1}\right) \cdot\left(\tau[0,1] \vDash q_{2}\right)\right) \\
& \quad \quad-\left(q_{2}\left(x_{2}\right)-\left(\tau[0,1] \vDash q_{2}\right)\right) \\
& \left.\quad \sum_{x_{1} \in X_{1}, x_{2} \in X_{2}} \tau\left(x_{1}, x_{2}\right) \cdot\left(\tau[1,0] \vDash q_{1}\right) \cdot q_{2}\left(x_{2}\right)\right) \\
& \quad+\left(\tau[1,0] \vDash q_{1}\right) \cdot\left(\tau[0,1] \vDash q_{2}\right) \\
& =\left(\tau \vDash q_{1} \otimes q_{2}\right)-\left(\tau[1,0] \vDash q_{1}\right) \cdot\left(\tau[0,1] \vDash q_{2}\right) .
\end{aligned}
$$

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:

$$
\tau=\frac{1}{4}|1,1\rangle+\frac{1}{4}|1,2\rangle+\frac{1}{4}|2,2\rangle+\frac{1}{4}|2,3\rangle .
$$

Its two marginals are:

$$
\tau[1,0]=\frac{1}{2}|1\rangle+\frac{1}{2}|2\rangle \quad \tau[0,1]=\frac{1}{4}|1\rangle+\frac{1}{2}|2\rangle+\frac{1}{4}|3\rangle
$$

Since both $X \hookrightarrow \mathbb{R}$ and $Y \hookrightarrow \mathbb{R}$ we get means as validities of the inclusions:

$$
\operatorname{mean}(\tau[1,0])=\frac{3}{2} \quad \operatorname{mean}(\tau[1,0])=2
$$

Now we can compute the joint covariance in the joint state $\tau$ as:

$$
\begin{aligned}
\operatorname{JCov}(\tau) & =\tau \vDash(\text { incl }-\operatorname{mean}(\tau[1,0]) \cdot \mathbf{1}) \otimes(\text { incl }-\operatorname{mean}(\tau[1,0]) \cdot \mathbf{1}) \\
& =\sum_{x \in X, y \in Y} \tau(x, y) \cdot\left(x-\frac{3}{2}\right) \cdot(y-2) \\
& =\frac{1}{4}\left(-\frac{1}{2} \cdot-1+\frac{1}{2} \cdot 1\right)=\frac{1}{4} .
\end{aligned}
$$

2 In order to calculate the (joint) correlation of $\tau$ we first need the variances of its marginals:

$$
\begin{aligned}
& \operatorname{Var}(\tau[1,0])=\sum_{x \in X} \tau[1,0](x) \cdot\left(x-\frac{3}{2}\right)^{2}=\frac{1}{4} \\
& \operatorname{Var}(\tau[0,1])=\sum_{y \in Y}, \tau[0,1](y) \cdot(y-2)^{2}=\frac{1}{2}
\end{aligned}
$$

Then:

$$
\operatorname{JCor}(\tau)=\frac{\operatorname{JCov}(\tau)}{\sqrt{\operatorname{Var}(\tau[1,0])} \cdot \sqrt{\operatorname{Var}(\tau[0,1])}}=\frac{1 / 4}{1 / 2 \cdot 1 / \sqrt{2}}=\frac{1}{2} \sqrt{2} .
$$

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 $\Delta \gg \omega \neq \omega \otimes \omega$, in general, see Subsection 2.3.2

Proposition 5.3.4. Shared state covariance can be expressed as joint covariance via copying:

$$
\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)=\operatorname{JCov}\left(\Delta \gg=\omega, p_{1}, p_{2}\right)
$$

More generally, for suitably typed channels $c, d$,

$$
\operatorname{Cov}(\omega, c=\ll p, d=\ll q)=\operatorname{JCov}(\langle c, d\rangle \gg=\omega, p, q) .
$$

The latter equation is reminiscent of Proposition 4.3.3 with validity $\vDash$ 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 Lemma5.3.2 we get:

$$
\begin{aligned}
& \operatorname{JCov}(\langle c, d\rangle \gg=\omega, p, q) \\
& =\langle c, d\rangle \gg=\omega \vDash(p-((\langle c, d\rangle\rangle=\omega)[1,0] \vDash p) \cdot \mathbf{1}) \\
& \otimes(q-((\langle c, d\rangle \gg \omega)[0,1] \vDash q) \cdot \mathbf{1}) \\
& =\omega \vDash\langle c, d\rangle=\ll((p-(c \gg \omega \vDash p) \cdot \mathbf{1}) \otimes(q-(d \gg=\omega \vDash q) \cdot \mathbf{1})) \\
& =\omega \vDash(c=\ll(p-(\omega \models c=\ll p) \cdot \mathbf{1})) \&(d=\ll(q-(\omega \vDash d=\ll q) \cdot \mathbf{1})) \\
& \text { by Lemma 4.3.2 7 } 7 \\
& =\omega \vDash((c=\ll p)-(\omega \vDash c=\ll p) \cdot \mathbf{1}) \&((d=\ll q)-(\omega \vDash d=\ll q) \cdot \mathbf{1}) \\
& \text { since } c \lll(-) \text { is linear and preserves } \mathbf{1} \\
& =\operatorname{Cov}(\omega, c=\ll p, d=\ll q) .
\end{aligned}
$$

We started with 'ordinary' covariance in Definition 5.1 .6 for two random variables with a shared state. It was subsequently used to define the 'joint' version in Definition 5.3.1 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. As we shall see below, there are notable differences between shared-state and joint-state versions, see Proposition 5.4.3 and Theorem5.4.6 in the next section.

But first we formulate a joint-state analogue for the linearity properties of Theorem 5.1.9.

Theorem 5.3.5. Consider a state $\tau \in \mathcal{D}\left(X_{1} \times X_{2}\right)$, with observables $q_{1} \in$ $\operatorname{Obs}\left(X_{1}\right), q_{2}, q_{3} \in \operatorname{Obs}\left(X_{2}\right)$ and numbers $r, s \in \mathbb{R}$.

1 Joint-state covariance satisfies:

$$
\begin{aligned}
\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right) & =\operatorname{JCov}\left(\tau, q_{2}, q_{1}\right) \\
\operatorname{JCov}\left(\tau, q_{1}, \mathbf{1}\right) & =0 \\
\operatorname{JCov}\left(\tau, r \cdot q_{1}, q_{2}\right) & =r \cdot \operatorname{JCov}\left(\tau, q_{1}, q_{2}\right) \\
\operatorname{JCov}\left(\omega, q_{1}, q_{2}+q_{3}\right) & =\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right)+\operatorname{Cov}\left(\tau, q_{1}, q_{3}\right) . \\
\operatorname{JCov}\left(\tau, q_{1}+r \cdot \mathbf{1}, q_{2}+s \cdot \mathbf{1}\right) & =\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right) .
\end{aligned}
$$

2 For joint-state correlation we have:

$$
\begin{aligned}
\operatorname{JCor}\left(\tau, r \cdot q_{1}, s \cdot q_{2}\right) & = \begin{cases}\operatorname{JCor}\left(\tau, q_{1}, q_{2}\right) & \text { if } r, \text { s have the same sign } \\
-\operatorname{JCor}\left(\tau, q_{1}, q_{2}\right) & \text { otherwise. }\end{cases} \\
\operatorname{JCor}\left(\tau, q_{1}+r \cdot \mathbf{1}, q_{2}+s \cdot \mathbf{1}\right) & =\operatorname{JCor}\left(\omega, q_{1}, q_{2}\right) .
\end{aligned}
$$

Proof. This follows directly from Theorem 5.1.9, using that predicate transformation $\pi_{i} \lll(-)$ is linear and thus preserves sums and scalar multiplications (and also truth), see Lemma 4.3.2 (2).

We conclude this section with a medical example about the correlation between disease and test.

Example 5.3.6. We start with a space $D=\left\{d, d^{\perp}\right\}$ 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:

$$
[0,1] \xrightarrow{\text { prev }} D \quad \text { with } \quad \operatorname{prev}(r):=r|d\rangle+(1-r)\left|d^{\perp}\right\rangle
$$

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 : $D \rightsquigarrow T$ with:

$$
\operatorname{test}(d)=\frac{9}{10}|p\rangle+\frac{1}{10}|n\rangle \quad \text { test }\left(d^{\perp}\right)=\frac{1}{20}|p\rangle+\frac{19}{20}|n\rangle .
$$

We can now form the joint 'graph' state:

$$
j \operatorname{joint}(r):=\langle\text { id, test }\rangle \gg=\operatorname{prev}(r) \in \mathcal{D}(D \times T)
$$

Exercise 5.3 .4 below tells that it does not really matter which observables we choose, so we simply take $\mathbf{1}_{d}: D \rightarrow[0,1]$ and $\mathbf{1}_{p}: T \rightarrow[0,1]$, mapping $d$ and $p$ to 1 and $d^{\perp}$ and $n$ to 0 . We are thus interested in (joint-state) correlation function:

$$
[0,1] \ni r \longmapsto \operatorname{JCor}\left(\operatorname{joint}(r), \mathbf{1}_{d}, \mathbf{1}_{p}\right) \in[-1,1]
$$

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 $(s e, s p): D \rightsquigarrow T$ with parameters $s e, s p \in[0,1]$.

$$
\begin{aligned}
\operatorname{test}(s e, s p)(d) & =s e|p\rangle+(1-s e)|n\rangle \\
\operatorname{test}(s e, s p)\left(d^{\perp}\right) & =(1-s p)|p\rangle+s p|n\rangle
\end{aligned}
$$

As before we form a joint state, but now with different parameters.

$$
\operatorname{joint}(s e, s p):=\langle\operatorname{id}, \operatorname{test}(s e, s p)\rangle \gg \operatorname{prev}\left(\frac{1}{5}\right) \in \mathcal{D}(D \times T) .
$$

We are then interested in the function:

$$
[0,1] \times[0,1] \ni(s e, s p) \longmapsto J \operatorname{JCor}\left(\text { joint }(\text { se }, s p), \mathbf{1}_{d}, \mathbf{1}_{p}\right) \in[-1,1] .
$$

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.


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 5.3.6for details.

## 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:

$$
\operatorname{JCor}\left(\tau, q_{1}, q_{2}\right)=\frac{\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right)}{\operatorname{StDev}\left(\tau[1,0], q_{1}\right) \cdot \operatorname{StDev}\left(\tau[0,1], q_{2}\right)}
$$

5.3.3 Consider distributions $\tau_{i} \in \mathcal{D}\left(X_{i}\right)$ with observables $p_{i} \in \operatorname{Obs}\left(X_{i}\right)$ for $i=1,2$. Use Theorem $5.1 .9(2)$ to prove that:

$$
\operatorname{Var}\left(\tau_{1} \otimes \tau_{2},\left(\pi_{1}=\ll p_{1}\right)+\left(\pi_{2}=\ll p_{2}\right)\right)=\operatorname{Var}\left(\tau_{1}, p_{1}\right)+\operatorname{Var}\left(\tau_{2}, p_{2}\right)
$$

where the observable $\left(\pi_{1}=\ll p_{1}\right)+\left(\pi_{2}=\ll p_{2}\right): X_{1} \times X_{2} \rightarrow \mathbb{R}$ sends a pair $\left(x_{1}, x_{2}\right)$ to $p_{1}\left(x_{1}\right)+p\left(x_{2}\right)$.
5.3.4 Consider two two-element sample spaces $X_{1}=\left\{a_{1}, b_{1}\right\}$ and $X_{2}=$ $\left\{a_{2}, b_{2}\right\}$ with a joint state $\tau \in \mathcal{D}\left(X_{1} \times X_{2}\right)$. Prove that in this binary case joint-state covariance and correlation do not depend on the observables, that is:

$$
\operatorname{JCor}\left(\tau, p_{1}, p_{2}\right)= \pm \operatorname{JCor}\left(\tau, q_{1}, q_{2}\right)
$$

for all non-constant observables $p_{1}, q_{1} \in \operatorname{Obs}\left(X_{1}\right), p_{2}, q_{2} \in \operatorname{Obs}\left(X_{2}\right)$. Hint: Use Theorem 5.3.5 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 5.3.6.

1 for each $s \in[0,1]$ one has:

$$
\operatorname{JCor}\left(\operatorname{joint}\left(s, 1-s, \mathbf{1}_{d}, \mathbf{1}_{p}\right)=0\right.
$$

2 for all se, $s p \in[0,1]$,

$$
s e+s p \geq 1 \Longleftrightarrow \operatorname{JCor}\left(\operatorname{joint}(\operatorname{se}, s p), \mathbf{1}_{d}, \mathbf{1}_{p}\right) \geq 0
$$

### 5.4 Independence for random variables

Earlier we have called a joint state / distribution entwined when it cannot be written as 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,

$$
\begin{equation*}
P\left[p_{1}=a, p_{2}=b\right]=P\left[p_{1}=a\right] \cdot P\left[p_{2}=b\right] \tag{5.6}
\end{equation*}
$$

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 Subsection 4.1.1, the expression $P[p=a]$ can be interpreted as transformation along the observable $p: X \rightarrow \mathbb{R}$, considered as deterministic channel:

$$
P[p=a]:=\mathcal{D}(p)(\omega)(a)=p \gg=\omega \vDash \mathbf{1}_{a} \stackrel{\text { 4.11] }}{=} \omega \vDash p=<\mathbf{1}_{a} .
$$

where $\omega \in \mathcal{D}(X)$ is the implicit distribution.
We can then translate the above requirement in Equation (5.6) into the condition:

$$
\begin{equation*}
\left\langle p_{1}, p_{2}\right\rangle \gg=\omega=\left(p_{1} \gg=\omega\right) \otimes\left(p_{2} \gg=\omega\right) . \tag{5.7}
\end{equation*}
$$

But this says that the joint state $\left\langle p_{1}, p_{2}\right\rangle \gg \omega=\omega$ on $\mathbb{R} \times \mathbb{R}$, transformed along
$\left\langle p_{1}, p_{2}\right\rangle: X \rightarrow \mathbb{R} \times \mathbb{R}$, is non-entwined: it is the product of its marginals. Indeed, its (first) marginal is:

$$
\begin{aligned}
\left(\left\langle p_{1}, p_{2}\right\rangle \gg \omega\right)[1,0] & =\pi_{1} \gg=\left(\left\langle p_{1}, p_{2}\right\rangle \gg=\omega\right) \\
& =\left(\pi_{1} \odot\left\langle p_{1}, p_{2}\right\rangle\right) \gg=\omega \\
& =p_{1} \gg \omega .
\end{aligned}
$$

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 $\left(\omega, p_{1}\right)$ and $\left(\omega, p_{2}\right)$ be two random variables with a common, shared state $\omega$. These random variables will be called independent if the transformed state $\left\langle p_{1}, p_{2}\right\rangle \gg \omega$ 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:

$$
\left.\left.\left\langle p_{1}, p_{2}\right\rangle\right\rangle=\omega=\left(p_{1} \gg=\omega\right) \otimes\left(p_{2}\right\rangle=\omega\right) .
$$

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 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 $\omega=f l i p \otimes f l i p$ with underlying set $2 \times 2$, where $2=\{0,1\}$. We will define two observables $p_{1}, p_{2}: 2 \times 2 \rightarrow \mathbb{R}$, to be used as random variables for this same, shared distribution $\omega$.
We first define an auxiliary observable $p: 2 \rightarrow \mathbb{R}$ for the amount of the bet:

$$
p(1)=100 \quad p(0)=50 .
$$

We then define $p_{1}=p \otimes \mathbf{1}=p \circ \pi_{1}: 2 \times 2 \rightarrow \mathbb{R}$, via weakening, as on the left below. The observable $p_{2}$ is defined on the right.

$$
p_{1}(x, y):=p(x)=\left\{\begin{array}{ll}
100 & \text { if } x=1 \\
50 & \text { if } x=0
\end{array} \quad \quad p_{2}(x, y):= \begin{cases}p(x) & \text { if } y=1 \\
-p(x) & \text { if } y=0 .\end{cases}\right.
$$

We claim that $\left(\omega, p_{1}\right)$ and $\left(\omega, p_{2}\right)$ 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:

$$
\begin{aligned}
\left\langle p_{1}, p_{2}\right\rangle \gg \omega= & \mathcal{D}\left(\left\langle p_{1}, p_{2}\right\rangle\right)(\text { flip } \otimes \text { flip }) \\
= & \frac{1}{4}\left|p_{1}(1,1), p_{2}(1,1)\right\rangle+\frac{1}{4}\left|p_{1}(1,0), p_{2}(1,0)\right\rangle \\
& \quad+\frac{1}{4}\left|p_{1}(0,1), p_{2}(0,1)\right\rangle+\frac{1}{4}\left|p_{1}(0,0), p_{2}(0,0)\right\rangle \\
= & \frac{1}{4}|100,100\rangle+\frac{1}{4}|100,-100\rangle+\frac{1}{4}|50,50\rangle+\frac{1}{4}|50,-50\rangle .
\end{aligned}
$$

Its two marginals are, in $\mathcal{D}(\mathbb{R})$,

$$
\begin{aligned}
& p_{1} \gg=\omega=\left(\left\langle p_{1}, p_{2}\right\rangle \gg=\omega\right)[1,0]=\frac{1}{2}|100\rangle+\frac{1}{2}|50\rangle \\
& p_{2} \gg=\omega=\left(\left\langle p_{1}, p_{2}\right\rangle \gg=\omega\right)[0,1]=\frac{1}{4}|100\rangle+\frac{1}{4}|-100\rangle+\frac{1}{4}|50\rangle+\frac{1}{4}|-50\rangle .
\end{aligned}
$$

It is not hard to see that the parallel product $\otimes$ of these two marginal distributions differs from the joint distribution $\left.\left\langle p_{1}, p_{2}\right\rangle\right\rangle=\omega$ on $\mathbb{R} \times \mathbb{R}$.

Proposition 5.4.3. The shared-state covariance of shared-state independent random variables is zero: if random variables $\left(\omega, p_{1}\right)$ and ( $\omega, p_{2}$ ) are independent, then $\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)=0$.

The converse does not hold.
Proof. If $\left(\omega, p_{1}\right)$ and $\left(\omega, p_{2}\right)$ are independent, then, by definition, $\left.\left\langle p_{1}, p_{2}\right\rangle\right\rangle=$ $\omega=\left(p_{1} \gg \omega\right) \otimes\left(p_{2} \gg \omega\right)$. The calculation belows shows that covariance is then zero. It uses multiplication $\&: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ as observable. It can also be described as the parallel product id $\otimes i d$ of the observable id: $\mathbb{R} \rightarrow \mathbb{R}$ with itself.

```
\(\operatorname{Cov}\left(\omega, p_{1}, p_{2}\right)\)
\(=\left(\omega \vDash p_{1} \& p_{2}\right)-\left(\omega \vDash p_{1}\right) \cdot\left(\omega \vDash p_{2}\right) \quad\) by Lemma5.1.7
\(=\left(\omega \vDash \& \circ\left\langle p_{1}, p_{2}\right\rangle\right)\)
    \(-\left(p_{1} \gg \omega=i d\right) \cdot\left(p_{2} \gg \omega \models i d\right) \quad\) by Exercise 4.1.5
\(=\left(\left\langle p_{1}, p_{2}\right\rangle \gg \omega=\omega\right)\)
    \(-\left(\left(p_{1} \gg \omega\right) \otimes\left(p_{2} \gg=\omega\right) \vDash\right.\) id \(\left.\otimes \mathrm{id}\right) \quad\) by Lemma4.2.9,4.1.5
\(=\left(\left\langle p_{1}, p_{2}\right\rangle \gg \omega \models \&\right)-\left(\left\langle p_{1}, p_{2}\right\rangle \gg=\omega \vDash \&\right) \quad\) by assumption
\(=0\).
```

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 $\left(\omega, p_{1}\right)$ and $\left(\omega, p_{2}\right)$, with shared state $\omega=$ flip $\otimes$ flip. We show here that they (nevertheless) have covariance zero. This proves the second claim of Proposition5.4.3, namely that zero-covariance need not imply independence, in the shared-state context.

We first compute the validities (expected values):

$$
\begin{aligned}
\omega \vDash p_{1} & =\frac{1}{4} \cdot p_{1}(1,1)+\frac{1}{4} \cdot p_{1}(1,0)+\frac{1}{4} \cdot p_{1}(0,1)+\frac{1}{4} \cdot p_{1}(0,0) \\
& =\frac{1}{4} \cdot 100+\frac{1}{4} \cdot 100+\frac{1}{4} \cdot 50+\frac{1}{4} \cdot 50=75 \\
\omega \vDash p_{2} & =\frac{1}{4} \cdot 100+\frac{1}{4} \cdot-100+\frac{1}{4} \cdot 50+\frac{1}{4} \cdot-50=0
\end{aligned}
$$

Then:

$$
\begin{aligned}
& \operatorname{Cov}\left(\omega, p_{1}, p_{2}\right) \\
& =\omega \vDash\left(p_{1}-\left(\omega \vDash p_{1}\right) \cdot \mathbf{1}\right) \&\left(p_{2}-\left(\omega \vDash p_{2}\right) \cdot \mathbf{1}\right) \\
& =\frac{1}{4}((100-75) \cdot 100+(100-75) \cdot-100+(50-75) \cdot 50+(50-75) \cdot-50) \\
& =0 .
\end{aligned}
$$

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 $\tau \in \mathcal{D}\left(X_{1} \times X_{2}\right)$ be a joint state and with two observables $q_{1} \in \operatorname{Obs}\left(X_{1}\right)$ and $q_{2} \in \operatorname{Obs}\left(X_{2}\right)$. We say that there is joint-state independence of $q_{1}, q_{2}$ if the two random variables $\left(\tau, \pi_{1}=\ll q_{1}\right)$ and $\left(\tau, \pi_{2}=\ll q_{2}\right)$ are (sharedstate) independent, as described in Definition 5.4.1

Concretely, this means that:

$$
\begin{equation*}
\left(q_{1} \otimes q_{2}\right) \gg \tau=\left(q_{1} \gg=\tau[1,0]\right) \otimes\left(q_{2} \gg=\tau[0,1]\right) . \tag{5.8}
\end{equation*}
$$

Equation (5.8) is an instance of the formulation (5.7) used in Definition 5.4.1 since $\pi_{i}=\ll q_{i}=q_{i} \circ \pi_{i}$ and:

$$
\begin{aligned}
\left\langle q_{1} \circ \pi_{1}, q_{2} \circ \pi_{2}\right\rangle \gg \tau=\tau & \mathcal{D}\left(\left\langle q_{1} \circ \pi_{1}, q_{2} \circ \pi_{2}\right\rangle\right)(\tau) \\
& =\mathcal{D}\left(q_{1} \times q_{2}\right)(\tau) \\
\stackrel{\text { 2.25. }}{=} & \left(q_{1} \otimes q_{2}\right) \gg \tau \\
\left(\left(q_{1} \circ \pi_{1}\right) \gg \tau\right) \otimes\left(\left(q_{2} \circ \pi_{2}\right) \gg \tau\right)= & \left(q_{1} \gg\left(\pi_{1} \gg \tau\right)\right) \otimes\left(q_{2} \gg=\left(\pi_{2} \gg \tau\right)\right) \\
& =\left(q_{1} \gg \tau[1,0]\right) \otimes\left(q_{2} \gg \tau[0,1]\right) .
\end{aligned}
$$

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 $\tau \in \mathcal{D}\left(X_{1} \times X_{2}\right)$ the following three statements are equivalent.
$1 \tau$ is non-entwined, i.e. $\tau$ is the product of its marginals;
2 the two observables $q_{i} \in \operatorname{Obs}\left(X_{i}\right)$ are joint-state independent wrt. $\tau$;
3 the joint-state covariance $\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right)$ is zero, for all observables $q_{i} \in$ $\operatorname{Obs}\left(X_{i}\right)$ - or equivalently, all correlations $\operatorname{JCor}\left(\tau, q_{1}, q_{2}\right)$ are zero.

Proof. Let joint state $\tau \in \mathcal{D}\left(X_{1} \times X_{2}\right)$ be given. We write $\tau_{i}:=\pi_{i} \gg \tau \in \mathcal{D}\left(X_{i}\right)$ for its marginals.
(1) $\Rightarrow$ (2). If $\tau$ is non-entwined, then $\tau=\tau_{1} \otimes \tau_{2}$. Hence for all observables $q_{1} \in \operatorname{Obs}\left(X_{1}\right)$ and $q_{2} \in \operatorname{Obs}\left(X_{2}\right)$ we have that $\sigma:=\left(q_{1} \otimes q_{2}\right) »=\tau$ is nonentwined. To see this, first note that $\pi_{i} \gg=\sigma=q_{i} \gg \tau_{i}$. Then, by Exercise 2.4.8.

$$
\begin{aligned}
\left(\pi_{1} \gg=\sigma\right) \otimes\left(\pi_{2} \gg \sigma\right) & =\left(q_{1} \gg=\tau_{1}\right) \otimes\left(q_{2} \gg=\tau_{2}\right) \\
& =\left(q_{1} \otimes q_{2}\right) \gg=\left(\tau_{1} \otimes \tau_{2}\right)=\left(q_{1} \otimes q_{2}\right) \gg \tau=\sigma .
\end{aligned}
$$

(2) $\Rightarrow$ (3). Let $q_{1} \in \operatorname{Obs}\left(X_{1}\right)$ and $q_{2} \in \operatorname{Obs}\left(X_{2}\right)$ be two observables. We may assume that $q_{1}, q_{2}$ are independent wrt. $\tau$, that is, $\left(q_{1} \otimes q_{2}\right) \gg \tau=\left(q_{1} \gg=\right.$ $\left.\tau_{1}\right) \otimes\left(q_{2} \gg=\tau_{2}\right)$ as in 5.8 . We must prove $\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right)=0$. Consider, like in the proof of Proposition 5.4.3, the multiplication map $\&: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, given by $\&\left(r_{1}, r_{2}\right)=r_{1} \cdot r_{2}$, as an observable on $\mathbb{R} \times \mathbb{R}$. We consider its validity:

$$
\begin{aligned}
\left(q_{1} \otimes q_{2}\right) \gg \tau \vDash \& & =\sum_{r_{1}, r_{2} \in \mathbb{R}}\left(\left(q_{1} \otimes q_{2}\right) \gg=\tau\right)\left(r_{1}, r_{2}\right) \cdot \&\left(r_{1}, r_{2}\right) \\
& =\sum_{r_{1}, r_{2} \in \mathbb{R}} \mathcal{D}\left(q_{1} \times q_{2}\right)(\tau)\left(r_{1}, r_{2}\right) \cdot r_{1} \cdot r_{2} \\
& =\sum_{r_{1}, r_{2} \in \mathbb{R}}\left(\sum_{\left(x_{1}, x_{2}\right) \in\left(q_{1} \times q_{2}\right)^{-1}(r, s)} \tau\left(x_{1}, x_{2}\right)\right) \cdot r \cdot s \\
& =\sum_{r_{1}, r_{2} \in \mathbb{R}} \sum_{x_{1} \in q_{1}^{-1}(r), x_{2} \in q_{2}^{-1}(s)} \tau\left(x_{1}, x_{2}\right) \cdot q_{1}\left(x_{1}\right) \cdot q_{2}\left(x_{2}\right) \\
& =\sum_{x_{1}, x_{2} \in X} \tau\left(x_{1}, x_{2}\right) \cdot\left(q_{1} \otimes q_{2}\right)\left(x_{1}, x_{2}\right) \\
& =\tau \models q_{1} \otimes q_{2}
\end{aligned}
$$

In the same way one proves $\left(q_{1} \gg \tau_{1}\right) \otimes\left(q_{2} \gg \tau_{2}\right) \vDash \&=\left(\tau_{1} \vDash q_{1}\right)$. $\left(\tau_{2} \vDash q_{2}\right)$. But then we are done via the formulation of binary covariance from Lemma 5.3.2.

$$
\begin{aligned}
\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right) & =\left(\tau \vDash q_{1} \otimes q_{2}\right)-\left(\tau_{1} \vDash q_{1}\right) \cdot\left(\tau_{2} \vDash q_{2}\right) \\
& =\left(\left(q_{1} \otimes q_{2}\right) \gg \tau \vDash \&\right)-\left(\left(q_{1} \gg=\tau_{1}\right) \otimes\left(q_{2} \gg=\tau_{2}\right) \vDash \&\right) \\
& =\left(\left(q_{1} \otimes q_{2}\right) \gg \tau \vDash \&\right)-\left(\left(q_{1} \otimes q_{2}\right) \gg \tau \vDash \&\right) \\
& =0 .
\end{aligned}
$$

(3) $\Rightarrow$ (1). Let joint-state covariance $\operatorname{JCov}\left(\tau, q_{1}, q_{2}\right)$ be zero for all observables $q_{1}, q_{2}$. In order to prove that $\tau$ 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 Lemma5.3.2. Then, since, by assumption, the binary covariance is zero, so that:

$$
\tau(x, y)=\tau \vDash \mathbf{1}_{x} \otimes \mathbf{1}_{y}=\left(\tau_{1} \vDash \mathbf{1}_{x}\right) \cdot\left(\tau_{2} \vDash \mathbf{1}_{y}\right)=\tau_{1}(x) \cdot \tau_{2}(y) .
$$

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 the setting of Definition 5.4.5 that the first marginal $\left(\left(q_{1} \otimes\right.\right.$ $\left.\left.q_{2}\right) \gg \tau\right)[1,0]$ of the transformed state along $q_{1} \otimes q_{2}$ is equal to the transformed marginal $q_{1} \gg=\tau[1,0]$.

### 5.5 The law of large numbers, in weak form

In this section we describe what is called the weak law of large numbers. The strong version appears later on, in ??. This weak law captures limit behaviour of probabilistic operations, such as: if we throw a fair dice many, many times, we expect to see each number of pips $\frac{1}{6}$ of the time. It is sometimes also called Bernoulli's Theorem.

We shall describe this law of large numbers in two forms: as binary version, which is most familiar, and as multivariate version. Both versions use results about means and variances that we have seen before.

First we need to introduce two famous inequalities, due to Markov and to Chebyshev. If we have an observable $p: X \rightarrow \mathbb{R}$ and a number $a \in \mathbb{R}$ we introduce a sharp predicate $[p \geq a]$ on $X$, defined in an obvious way as:

$$
[p \geq a](x):= \begin{cases}1 & \text { if } p(x) \geq a  \tag{5.9}\\ 0 & \text { otherwise }\end{cases}
$$

We can similarly write $[p>a],[p \leq a]$ and $[p<a]$.
Lemma 5.5.1. Let $\omega \in \mathcal{D}(X)$ be a state, $p \in \operatorname{Obs}(X)$ an observable, and $a \in \mathbb{R}$ an arbitrary number. Then:

1 Chebyshev's inequality holds: $a \cdot(\omega \vDash[p \geq a]) \leq \omega \vDash p$;
2 Markov's inequality holds: $a^{2} \cdot(\omega \vDash[|p-(\omega \vDash p) \cdot \mathbf{1}| \geq a]) \leq \operatorname{Var}(\omega, p)$.
Proof. 1 Because:

$$
\begin{aligned}
\omega \vDash p \geq \omega \vDash p \&[p \geq a] & \geq \omega \vDash(a \cdot \mathbf{1}) \&[p \geq a] \\
& =\omega \vDash a \cdot(\mathbf{1} \&[p \geq a]) \\
& =\omega \vDash a \cdot[p \geq a] \\
& =a \cdot(\omega \vDash[p \geq a]) .
\end{aligned}
$$

2 Write $q:=p-(\omega \vDash p) \cdot \mathbf{1}: X \rightarrow \mathbb{R}$, so that $q(x)=p(x)-(\omega \vDash p)$. Then:

$$
\begin{aligned}
{[|q| \geq a](x)=1 } & \Longleftrightarrow|q(x)| \geq a \\
& \Longleftrightarrow q(x)^{2} \geq a^{2} \Longleftrightarrow\left[q^{2} \geq a^{2}\right](x)=1 .
\end{aligned}
$$

This gives an inequality of (sharp) predicates: $(|q| \geq a) \leq\left(q^{2} \geq a^{2}\right)$. Hence, by the previous item (Markov's inequality),

$$
\begin{aligned}
a^{2} \cdot(\omega \vDash[|p-(\omega \vDash p) \cdot \mathbf{1}| \geq a]) & =a^{2} \cdot(\omega \vDash[|q| \geq a]) \\
& \leq a^{2} \cdot\left(\omega \vDash\left[q^{2} \geq a^{2}\right]\right) \\
& \leq \omega \vDash q^{2} \quad \text { by item (1) } \\
& =\operatorname{Var}(\omega, p) .
\end{aligned}
$$

We turn to the weak law of large numbers, in binary form. To start, fix the single and parallel coin states:

$$
\sigma:=\operatorname{flip}\left(\frac{1}{2}\right)=\frac{1}{2}|1\rangle+\frac{1}{2}|0\rangle \in \mathcal{D}(2) \quad \text { and } \quad \sigma^{n}:=\sigma \otimes \cdots \otimes \sigma \in \mathcal{D}\left(2^{n}\right)
$$

with an average predicate on $2^{n}$, namely:

$$
\begin{equation*}
2^{n} \xrightarrow{\operatorname{avg}_{n}}[0,1] \quad \text { defined by } \quad \operatorname{avg}_{n}\left(x_{1}, \ldots, x_{n}\right):=\frac{x_{1}+\cdots+x_{n}}{n} \tag{5.10}
\end{equation*}
$$

The predicate $\operatorname{avg}_{n}$ captures the average number of heads (as 1) in $n$ coin flips.
Then:

$$
\begin{aligned}
& \sigma \vDash \operatorname{avg}_{1}=\frac{1}{2} \cdot 1+\frac{1}{2} \cdot 0=\frac{1}{2} \\
& \sigma^{2} \vDash \operatorname{avg}_{2}=\frac{\frac{1}{4} \cdot(1+1)+\frac{1}{4} \cdot(1+0)+\frac{1}{4} \cdot(0+1)+\frac{1}{4} \cdot(0+0)}{2}=\frac{1}{2} \\
& \quad \vdots \\
& \sigma^{n} \vDash \operatorname{avg}_{n}=\frac{\sum_{1 \leq k \leq n}\binom{n}{k} \cdot \frac{k}{2^{n}}}{n}=\frac{\operatorname{mean}\left(b n[n]\left(\frac{1}{2}\right)\right)}{n}=\frac{n \cdot 1 / 2}{n}=\frac{1}{2} .
\end{aligned}
$$

For the last equation, see Example 4.1.4 (4). These equations express that in $n$ coin flips the average number of heads is $\frac{1}{2}$. This makes perfect sense.

The weak law of large numbers involves a more subtle statement, namely that for each $\varepsilon>0$ the validity:

$$
\begin{equation*}
\sigma^{n} \vDash\left[\left|\operatorname{avg}_{n}-\frac{1}{2} \cdot \mathbf{1}\right| \geq \varepsilon\right] . \tag{5.11}
\end{equation*}
$$

goes to zero as $n$ goes to infinity. One may think that this convergence to zero is obvious, but it is not, see Figure 5.3. The above validities 5.11) do go down, but not monotonously.


Figure 5.3 Example validities (5.11) with $\varepsilon=\frac{1}{10}$, from $n=1$ to $n=20$.

Theorem 5.5.2 (Weak law of large numbers). Using the fair $\operatorname{coin} \sigma=\operatorname{flip}\left(\frac{1}{2}\right)=$ $\frac{1}{2}|1\rangle+\frac{1}{2}|0\rangle$ and the average predicate $\operatorname{avg}_{n}: 2^{n} \rightarrow[0,1]$ from (5.10), for each $\varepsilon>0$,

$$
\lim _{n \rightarrow \infty} \sigma^{n} \vDash\left[\left|\operatorname{avg}_{n}-\frac{1}{2} \cdot \mathbf{1}\right| \geq \varepsilon\right]=0
$$

Proof. We use Chebyshev's inequality from Lemma 5.5.1 (2). Since $\sigma^{n} \vDash$ $\operatorname{avg}_{n}=\frac{1}{2}$, as we have seen above, it gives an inequality in:

$$
\sigma^{n} \vDash\left[\left|\operatorname{avg}_{n}-\frac{1}{2} \cdot \mathbf{1}\right| \geq \varepsilon\right] \leq \frac{\operatorname{Var}\left(\sigma^{n}, \operatorname{avg}_{n}\right)}{\varepsilon^{2}} \stackrel{(*)}{=} \frac{1}{4 n \varepsilon^{2}}
$$

It suffices to prove the marked equality $\stackrel{(*)}{=}$, since then it is clear that the limit in the theorem goes to zero, as $n$ goes to infinity.

The product $2^{n}$ comes with $n$ projection functions $\pi_{i}: 2^{n} \rightarrow 2$. Using these we can write the predicate $\operatorname{avg}_{n}: 2^{n} \rightarrow[0,1]$ as $\frac{1}{n} \cdot \sum_{i} \pi_{i}$. Thus:

$$
\begin{aligned}
\sigma^{n} \vDash \operatorname{avg}_{n} \& \operatorname{avg}_{n} & =\frac{1}{n^{2}} \cdot \sum_{i, j} \sigma^{n} \vDash \pi_{i} \& \pi_{j} \\
& =\frac{1}{n^{2}} \cdot\left(\sum_{i} \sigma^{n} \vDash \pi_{i} \& \pi_{i}+\sum_{i, j, i \neq j} \sigma^{n} \vDash \pi_{i} \& \pi_{i}\right) \\
& =\frac{1}{n^{2}} \cdot\left(\frac{n}{2}+\frac{n^{2}-n}{4}\right) \\
& =\frac{n+1}{4 n} .
\end{aligned}
$$

Hence, by Lemma 5.1.3

$$
\operatorname{Var}\left(\sigma^{n}, \operatorname{avg}_{n}\right)=\left(\sigma^{n} \vDash \operatorname{avg}_{n} \& \operatorname{avg}_{n}\right)-\left(\sigma^{n} \vDash \operatorname{avg}_{n}\right)^{2}=\frac{n+1}{4 n}-\frac{1}{4}=\frac{1}{4 n}
$$

This proves the marked equation $\stackrel{(*)}{=}$, and thus the theorem.

An obvious next step is to extend this result to arbitrary distributions. So let us fix a distribution $\omega$. For a number $n \in \mathbb{N}$ and for an element $y \in X$ we use the accumulation predicate $\operatorname{accfrac}(y): X^{n} \rightarrow[0,1]$, given by:

$$
\begin{equation*}
\operatorname{accfrac}(y)\left(x_{1}, \ldots, x_{n}\right):=\frac{\operatorname{acc}\left(x_{1}, \ldots, x_{n}\right)(y)}{n}=\frac{\left|\left\{i \mid x_{i}=y\right\}\right|}{n} \tag{5.12}
\end{equation*}
$$

Thus, the predicate $\operatorname{accfrac}(y)$ outputs the fraction of $y$ 's in an input sequence. For instance, $\operatorname{accfrac}(c)(a, b, c, c, b)=\frac{2}{5}$.

As one may expect, the validity of accfrac $(y)$ converges to $\omega(y)$ in product states $\omega^{n}=\omega \otimes \cdots \otimes \omega$ when $n$ goes to infinity. This is the content of item (3) below, which is the multivariate version of the weak law of large numbers.

Proposition 5.5.3. Consider the above situation, with a distribution $\omega$, an element $y \in X$, and the predicate $\operatorname{accfrac}(y)$ from (5.12).

1 The validity is given by:

$$
\omega^{n} \vDash \operatorname{accfrac}(y)=\omega(y) .
$$

2 The formula for the variance is:

$$
\operatorname{Var}\left(\omega^{n}, \operatorname{accfrac}(y)\right)=\frac{\omega(y) \cdot(1-\omega(y))}{n}
$$

3 For each $\varepsilon>0$,

$$
\lim _{n \rightarrow \infty} \omega^{n} \vDash[|\operatorname{accfrac}(y)-\omega(y) \cdot \mathbf{1}| \geq \varepsilon]=0
$$

Proof. 1 Notice that we can write:

$$
\operatorname{accfrac}(y)=\frac{1}{n} \cdot\left(X^{n} \xrightarrow{\mathrm{acc}} \mathcal{N}[n](X) \xrightarrow{\mathrm{ev}_{y}} \mathbb{R}_{\geq 0}\right),
$$

where $e v_{x_{i}}$ is the point evaluation map from 4.13. Then, using acc as a deterministic channel,

$$
\begin{array}{rlrl}
\omega^{n} \vDash \operatorname{accfrac}(y) & =\frac{1}{n} \cdot\left(\omega^{n} \vDash \operatorname{acc}=\ll \mathrm{ev}_{y}\right) & & \\
& =\frac{1}{n} \cdot\left(\operatorname{acc} \gg=\operatorname{iid}[n](\omega) \vDash \mathrm{ev}_{y}\right) & \\
& =\frac{1}{n} \cdot\left(\operatorname{mn}[n](\omega) \vDash \mathrm{ev}_{y}\right) & & \text { by Theorem 2.6.7 } \\
& =\omega(y) & & \text { by Lemma3.3.2. }
\end{array}
$$

2 Via a combination of several earlier results we get:

$$
\begin{aligned}
\operatorname{Var}\left(\omega^{n}, \operatorname{accfrac}(y)\right) & =\operatorname{Var}\left(\omega^{n}, \frac{1}{n} \cdot\left(\mathrm{ev} v_{y} \circ \operatorname{acc}\right)\right) & & \\
& =\frac{1}{n^{2}} \cdot \operatorname{Var}\left(\omega^{n}, \mathrm{ev}_{y} \circ \mathrm{acc}\right) & & \text { by Theorem 5.1.9] } 2] \\
& =\frac{1}{n^{2}} \cdot \operatorname{Var}\left(\mathcal{D}(\operatorname{acc})\left(\omega^{n}\right), \mathrm{ev}_{y}\right) & & \text { by Exercise5.1.9 } \\
& =\frac{1}{n^{2}} \cdot \operatorname{Var}\left(m n[n](\omega), \mathrm{ev}_{y}\right) & & \text { see Theorem 2.6.7 } \\
& =\frac{\omega(y) \cdot(1-\omega(y))}{n} . & &
\end{aligned}
$$

The latter equation involves a combination of Proposition 5.2.1 (1) and Equation (5.3).
3 Via Chebyshev's inequality from Lemma 5.5.1, 22, in combination with the previous two items, we get:

$$
\begin{aligned}
\lim _{n \rightarrow \infty} \omega^{n} \vDash[|\operatorname{accfrac}(y)-\omega(y) \cdot \mathbf{1}| \geq \varepsilon] & \leq \lim _{n \rightarrow \infty} \frac{\operatorname{Var}\left(\omega^{n}, \operatorname{accfrac}(y)\right)}{\varepsilon^{2}} \\
& =\lim _{n \rightarrow \infty} \frac{\omega(y) \cdot(1-\omega(y))}{n \cdot \varepsilon^{2}} \\
& =0
\end{aligned}
$$

There is an alternative formulation. In the above situation we use the accumulation map acc: $X^{n} \rightarrow \mathcal{N}[n](X)$. We shall use it in combination with frequentist learning, namely as function:

$$
\text { Flrn } \circ \text { acc }: X^{n} \longrightarrow \mathcal{D}(X)
$$

For instance $($ Flrn $\circ \operatorname{acc})(a, c, a, a, c, b)=\operatorname{Flrn}(\operatorname{acc}(a, c, a, a, c, b))=\frac{1}{2}|a\rangle+$ $\frac{1}{6}|b\rangle+\frac{1}{3}|c\rangle$. This composite Flrn $\circ$ acc is related to the predicate accfrac that we use above (5.12), since, for $\vec{x} \in X^{n}$,

$$
(\text { Flrn } \circ \operatorname{acc})(\vec{x})=\sum_{y \in X} \frac{\operatorname{acc}(\vec{x})(y)}{n}|y\rangle=\sum_{y \in X} \operatorname{accfrac}(y)(\vec{x})|y\rangle .
$$

For a given state $\omega \in \mathcal{D}(X)$ we can look at the total variation distance $d$, see Subsection 4.5.1, as a predicate $X^{n} \rightarrow[0,1]$, given by:

$$
\begin{align*}
\vec{x} \longmapsto d(\operatorname{Flrn}(\operatorname{acc}(\vec{x})), \omega) & =\frac{1}{2} \cdot \sum_{y \in X}|\operatorname{Flrn}(\operatorname{acc}(\vec{x}))(y)-\omega(y)| \\
& =\frac{1}{2} \cdot \sum_{y \in X}|\operatorname{accfrac}(y)(\vec{x})-\omega(y)| . \tag{5.13}
\end{align*}
$$

This predicate is used in the following multivariate weak large number theorem. Informally it says: the distribution that is obtained by accumulating $n$ samples from $\omega$ has a total variation distance to $\omega$ that goes to zero as $n$ goes
to infinity. This holds 'in probability', since it is expressed as a predicate that is evaluated in a (product) state $\omega^{n}$.

Theorem 5.5.4. For a state $\omega \in \mathcal{D}(X)$ and a number $\varepsilon>0$ one has:

$$
\lim _{n \rightarrow \infty} \omega^{n} \vDash[d(\operatorname{Flrn}(\operatorname{acc}(-)), \omega) \geq \varepsilon]=0
$$

Alternatively, in terms of multinomial distributions:

$$
\lim _{n \rightarrow \infty} m n[n](\omega) \vDash[d(\operatorname{Flrn}(-), \omega) \geq \varepsilon]=0 .
$$

Proof. Let $\operatorname{supp}(\omega) \subseteq X$ have $\ell \in \mathbb{N}_{>0}$ elements. There is an inequality of sharp predicates on $X^{n}$ of the form:

$$
\begin{equation*}
[d(F \operatorname{lrn}(\operatorname{acc}(-)), \omega) \geq \varepsilon] \leq \sum_{y \in \operatorname{supp}(\omega)}\left[|\operatorname{accfrac}(y)-\omega(y) \cdot \mathbf{1}| \geq \frac{2 \varepsilon}{\ell}\right] . \tag{*}
\end{equation*}
$$

Indeed, if the right-hand-side in $(*)$ is 0 , then for each $\vec{x} \in X^{n}$ and $y \in \operatorname{supp}(\omega)$ we have:

$$
|F \operatorname{lrn}(\operatorname{acc}(\vec{x}))(y), \omega(y)|=|\operatorname{accfrac}(y)(\vec{x}), \omega(y)|<\frac{\varepsilon}{2 \ell} .
$$

The predicate (5.13) now satisfies:

$$
\begin{aligned}
d(\operatorname{Flrn}(\operatorname{acc}(-)), \omega)(\vec{x}) & =\frac{1}{2} \cdot \sum_{y \in \operatorname{supp}(\omega)}|\operatorname{Flrn}(\operatorname{acc}(\vec{x}))(y), \omega(y)| \\
& <\frac{1}{2} \cdot \sum_{y \in \operatorname{supp}(\omega)}^{\frac{2 \varepsilon}{\ell}}=\varepsilon .
\end{aligned}
$$

Thus the left-hand-side in $(*)$ satisfies: $[d(F \operatorname{lrn}(\operatorname{acc}(-)), \omega) \geq \varepsilon](\vec{x})=0$.
Now we can prove the first limit result in the theorem:

$$
\begin{aligned}
& \lim _{n \rightarrow \infty} \omega^{n} \vDash[d(\operatorname{Flrn}(\operatorname{acc}(-)), \omega) \geq \varepsilon] \\
& \leq \lim _{n \rightarrow \infty} \omega^{n} \vDash \sum_{y \in \operatorname{supp}(\omega)}\left[|\operatorname{accfrac}(y)-\omega(y) \cdot \mathbf{1}| \geq \frac{2 \varepsilon}{\ell}\right] \quad \text { by }(*) \\
& =\sum_{y \in \operatorname{supp}(\omega)} \lim _{n \rightarrow \infty} \omega^{n} \vDash\left[|\operatorname{accfrac}(y)-\omega(y) \cdot \mathbf{1}| \geq \frac{2 \varepsilon}{\ell}\right] \\
& =0 \quad \text { by Proposition } 5.5 .3 \text { 33). }
\end{aligned}
$$

The second limit in the theorem is now obtained via Proposition 4.3.3 and Theorem 2.6.7.

Figure 5.4 gives an impression of the terms in a limit like in Theorem 5.5.4, for $\omega=\frac{1}{4}|a\rangle+\frac{5}{12}|b\rangle+\frac{1}{3}|c\rangle$. The plot covers only a small initial fragment because of the exponential growth of the sample space. For instance, the space of the parallel product $\omega^{12}$ has $3^{12}$ elements, which is more than half a million.


Figure 5.4 Initial segment of the limit from Theorem 5.5.4 with $\omega=\frac{1}{4}|a\rangle+$ $\frac{5}{12}|b\rangle+\frac{1}{3}|c\rangle$ and $\varepsilon=\frac{1}{10}$, from $n=1$ to $n=12$.

This initial segment suggests that the validities go down, but the suggestion is weak; Theorem 5.5.4 provides certainty.

In the end we recall from Theorem 3.3.3 the equation:

$$
\omega=F \operatorname{lrn} \gg m n[K](\omega)=\sum_{\varphi \in \mathcal{N}[K](X)} m n[K](\omega)(\varphi) \cdot \operatorname{Flrn}(\varphi) .
$$

This says that an arbitrary distribution $\omega$ is actually equal to a convex sum of normalised multinomial draws from it. This works for any number $K$, and not only in the limit. We can formulate the second equation in the above Theorem 5.5.4 in a somewhat similar way as:

$$
\lim _{n \rightarrow \infty} \sum\{m n[n](\omega)(\varphi) \mid \varphi \in \mathcal{N}[n](X) \text { with } d(\operatorname{Flrn}(\varphi), \omega) \geq \varepsilon\}=0
$$

We can read this as: for each very large multiset $\varphi$, if its normalisation $\operatorname{Flrn}(\varphi)$ differs more than a little bit $(\varepsilon)$ from $\omega$, then it is an unlikely multinomial draw from $\omega$. Or, turned around, the normalisation of a large multiset $\varphi$ drawn from $\omega$, looks very much like $\omega$. This the basis of sampling from a distribution, see Subsection 2.2.1.

## Exercises

5.5.1 For a number $a \in \mathbb{R}$, write $\mathbf{1}_{\geq a}: \mathbb{R} \rightarrow[0,1]$ for the sharp predicate with $\mathbf{1}_{\geq a}(r)=1$ iff $r \geq a$. Consider an observable $p: X \rightarrow \mathbb{R}$ as a deterministic channel. Check that the sharp predicate $[p \geq a]$ from (5.9) can also be described as predicate transformation $p=\ll \mathbf{1}_{\geq a}$.
5.5.2 Prove that (Flrn ○ acc) $>=\omega^{n}=\omega$.
5.5.3 Show how Theorem 5.5.2 is an instance of Theorem 5.5.3 (3).

## 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 $\omega$, both on the same sample space, one can form a new updated (conditioned) distribution, written as $\left.\omega\right|_{p}$. It absorps the evidence $p$ into $\omega$. This updating satisfies various properties, including the famous rule of Thomas Bayes, and also a less well known but also important validity-increase property:

$$
\left.\omega\right|_{p} \vDash p \geq \omega \vDash p
$$

It states that after absorbing the evidence $p$ the validity of $p$ increases. This makes perfect good sense and is crucial in probabilistic learning.

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 $\left.\omega\right|_{p}$ involves both a distribution $\omega$ and a factor $p$. In presence of a channel we can first transform the state $\omega$ or the factor $p$ and then update. But we can also transform an updated state $\left.\omega\right|_{p}$. 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; it is commonly called causal reasoning. There is also backward inference, 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 6.7, but also for Bayesian networks, in Sections 6.4 and 6.5 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.
In Section 6.7 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 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.

At the end of this chapter, Section 6.8 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 $\operatorname{Pred}(X) \rightarrow[0,1]$ 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.

### 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). 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 $p \in \operatorname{Fact}(X)=\left(\mathbb{R}_{\geq 0}\right)^{X}$ be a factor, on the same space $X$.

1 If the validity $\omega \vDash p$ is non-zero, we define a new distribution $\left.\omega\right|_{p} \in \mathcal{D}(X)$ as normalised pointwise product of $\omega$ and $p$ :

$$
\begin{equation*}
\left.\omega\right|_{p}(x):=\frac{\omega(x) \cdot p(x)}{\omega \models p} \quad \text { i.e. }\left.\quad \omega\right|_{p}=\sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \vDash p}|x\rangle . \tag{6.1}
\end{equation*}
$$

This distribution $\left.\omega\right|_{p}$ may be pronounced as: $\omega$, given $p$.
2 The conditional expectation or conditional validity of an observable $q$ on $X$, given $p$ and $\omega$, is the validity:

$$
\left.\omega\right|_{p} \vDash q .
$$

3 For a channel $c: X \leadsto Y$ and a factor $q$ on $Y$ we define the updated channel $\left.c\right|_{q}: X \leadsto Y$ via pointwise updating:

$$
\left.c\right|_{q}(x):=\left.c(x)\right|_{q} .
$$

In writing $\left.c\right|_{q}$ we assume that validity $c(x) \vDash q=(c=\ll q)(x)$ is non-zero, for each $x \in X$.

Often, the distribution $\omega$ before updating is called the prior or the a priori distribution, whereas the updated distribution $\left.\omega\right|_{p}$ is called the posterior or the a posteriori distribution. The posterior incorporates the evidence given by the factor $p$. One may thus expect that in the updated state $\left.\omega\right|_{p}$ the factor $p$ is more true than in $\omega$. This is indeed the case, as will be shown in Theorem 6.1.5 below.
In Equation (6.1) we see that the validity $\omega \vDash p$ is used for normalisation. Computing this validity may be computationally expensive, when the distribution $\omega$ has a very large support. We shall describe how to sample an updated distribution in Subsection 6.6.1.

Notice that we define updating $\left.\omega\right|_{p}$ only for factor $p: X \rightarrow \mathbb{R}_{\geq 0}$, with nonnegative outcomes, and not for arbitrary observables $X \rightarrow \mathbb{R}$. The latter lead to negative probabilities, which are excluded in our setting.

The conditioning $\left.c\right|_{q}$ of a channel in item (3) is in fact a generalisation of the conditioning of a state $\left.\omega\right|_{p}$ in item (1), since the state $\omega \in \mathcal{D}(X)$ can be seen as a channel $\omega: \mathbf{1} \leadsto X$ 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 $\omega$ is this implicit distribution, then $P(E \mid D)$ corresponds to the conditional expectation expressed by the validity $\left.\omega\right|_{\mathbf{1}_{D}} \vDash \mathbf{1}_{E}$ of the sharp predicate $\mathbf{1}_{E}$, in the state $\omega$ updated with
the sharp predicate $\mathbf{1}_{D}$. Indeed,

$$
\begin{aligned}
\omega \mid \mathbf{1}_{D} \vDash \mathbf{1}_{E} & =\left.\sum_{x \in X} \omega\right|_{\mathbf{1}_{D}}(x) \cdot \mathbf{1}_{E}(x) \\
& \stackrel{6.1]}{=} \sum_{x \in E} \frac{\omega(x) \cdot \mathbf{1}_{D}(x)}{\omega \vDash \mathbf{1}_{D}} \\
& =\frac{\sum_{x \in E \cap D} \omega(x)}{\sum_{x \in D} \omega(x)}=\frac{P(E \cap D)}{P(D)}=P(E \mid D) .
\end{aligned}
$$

The formulation $\left.\omega\right|_{p}$ 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 [21, 37, 79]. It is what we use as default form.

## Example 6.1.2.

1 Let's take the numbers of a dice as sample space: pips $=\{1,2,3,4,5,6\}$, with a fair / uniform dice distribution dice $=$ unif $_{\text {pips }}=\frac{1}{6}|1\rangle+\frac{1}{6}|2\rangle+\frac{1}{6}|3\rangle+\frac{1}{6}|4\rangle+$ $\frac{1}{6}|5\rangle+\frac{1}{6}|6\rangle$. We consider the predicate evenish $\in \operatorname{Pred}($ pips $)=[0,1]^{\text {pips }}$ expressing that we are fairly certain of the outcome being even:

$$
\begin{array}{lll}
\operatorname{evenish}(1)=\frac{1}{5} & \text { evenish }(3)=\frac{1}{10} & \text { evenish }(5)=\frac{1}{10} \\
\operatorname{evenish}(2)=\frac{9}{10} & \text { evenish }(4)=\frac{9}{10} & \text { evenish }(6)=\frac{4}{5}
\end{array}
$$

We first compute the validity of evenish for our fair dice:

$$
\begin{aligned}
\text { dice } \vDash \text { evenish } & =\sum_{x \in \operatorname{pips}} \operatorname{dice}(x) \cdot \operatorname{evenish}(x) \\
& =\frac{1}{6} \cdot \frac{1}{5}+\frac{1}{6} \cdot \frac{9}{10}+\frac{1}{6} \cdot \frac{1}{10}+\frac{1}{6} \cdot \frac{9}{10}+\frac{1}{6} \cdot \frac{1}{10}+\frac{1}{6} \cdot \frac{4}{5} \\
& =\frac{2+9+1+9+1+8}{60}=\frac{1}{2} .
\end{aligned}
$$

If we take evenish as evidence, we can update our dice state and get:

$$
\begin{aligned}
\text { dice }\left.\right|_{\text {evenish }}= & \sum_{x \in \text { pips }} \frac{\operatorname{dice}(x) \cdot \operatorname{evenish}(x)}{\text { dice } \models \text { evenish }}|x\rangle \\
= & \frac{1 / 6 \cdot 1 / 5}{1 / 2}|1\rangle+\frac{1 / 6 \cdot 9 / 10}{1 / 2}|2\rangle+\frac{1 / 6 \cdot 1 / 10}{1 / 2}|3\rangle \\
& \quad+\frac{1 / 6 \cdot 9 / 10}{1 / 2}|4\rangle+\frac{1 / 6 \cdot 1 / 10}{1 / 2}|5\rangle+\frac{1 / 6 \cdot 4 / 5}{1 / 2}|6\rangle \\
= & \frac{1}{15}|1\rangle+\frac{3}{10}|2\rangle+\frac{1}{30}|3\rangle+\frac{3}{10}|4\rangle+\frac{1}{30}|5\rangle+\frac{4}{15}|6\rangle .
\end{aligned}
$$

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.

2 The following alarm example is due to Pearl [145]. It involves an 'alarm' set $A=\left\{a, a^{\perp}\right\}$ and a 'burglary' set $B=\left\{b, b^{\perp}\right\}$, with the following a priori joint distribution $\omega \in \mathcal{D}(A \times B)$.

$$
0.000095|a, b\rangle+0.009999\left|a, b^{\perp}\right\rangle+0.000005\left|a^{\perp}, b\right\rangle+0.989901\left|a^{\perp}, b^{\perp}\right\rangle
$$

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:

$$
\omega[0,1]=0.0001|b\rangle+0.9999\left|b^{\perp}\right\rangle .
$$

Someone reports that the alarm went off, but with only $80 \%$ certainty because of deafness. This can be described as a predicate $p: A \rightarrow[0,1]$ with $p(a)=0.8$ and $p\left(a^{\perp}\right)=0.2$. 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 $\omega$, since $p$ is a predicate on $A$ and $\omega$ is a (joint) distribution on the product set $A \times B$. This mismatch can be solved via weakening $p$ to $p \otimes \mathbf{1}=\pi_{1}=\ll p$, so that it becomes a predicate on $A \times B$. Then we can do the update of the joint distribution:

$$
\left.\omega\right|_{p \otimes \mathbf{1}} \in \mathcal{D}(A \times B)
$$

In order understand it in detail we first compute the validity:

$$
\omega \vDash p \otimes \mathbb{1}=\sum_{x \in A, y \in B} \omega(x, y) \cdot p(x)=0.206
$$

We can then compute the updated joint state as:

$$
\begin{aligned}
\left.\omega\right|_{p \otimes \mathbf{1}}= & \sum_{x \in A, y \in B} \frac{\omega(x, y) \cdot p(x)}{\omega \models p}|x, y\rangle \\
= & 0.0003688|a, b\rangle+0.03882\left|a, b^{\perp}\right\rangle \\
& \quad+0.000004853\left|a^{\perp}, b\right\rangle+0.9608\left|a^{\perp}, b^{\perp}\right\rangle .
\end{aligned}
$$

The resulting posterior burglary distribution - with the alarm evidence taken into account - is obtained by taking the second marginal of the updated distribution:

$$
\left(\left.\omega\right|_{p \otimes 1}\right)[0,1]=0.0004|b\rangle+0.9996\left|b^{\perp}\right\rangle .
$$

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 [97]) or 'crossover inference' happens precisely because the joint distribution $\omega$ is entwined, so that the different parts can influence each other.

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 6.1.3. Let $\omega$ be distribution on a sample space $X$, and let $p, q$ be factors on $X$.

1 The product rule holds for conditional validity:

$$
\left.\omega\right|_{p} \vDash q=\frac{\omega \vDash p \& q}{\omega \vDash p} .
$$

2 Bayes' rule holds:

$$
\left.\omega\right|_{p} \vDash q=\frac{\left(\left.\omega\right|_{q} \vDash p\right) \cdot(\omega \vDash q)}{\omega \vDash p} .
$$

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.

Proof. 1 We straightforwardly compute:

$$
\begin{aligned}
\left.\omega\right|_{p} \vDash q=\left.\sum_{x \in X} \omega\right|_{p}(x) \cdot q(x) & \stackrel{\text { 6.1] }}{-} \sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \vDash p} \cdot q(x) \\
& =\frac{\sum_{x \in X} \omega(x) \cdot p(x) \cdot q(x)}{\omega \vDash p} \\
& =\frac{\sum_{x \in X} \omega(x) \cdot(p \& q)(x)}{\omega \vDash p}=\frac{\omega \vDash p \& q}{\omega \vDash p} .
\end{aligned}
$$

2 This follows directly by using the previous item twice, in combination with the commutativity of conjunction $\&$, in:

$$
\left.\omega\right|_{p} \vDash q \stackrel{\text { 凹 }}{=} \frac{\omega \vDash p \& q}{\omega \vDash p}=\frac{\omega \vDash q \& p}{\omega \vDash p} \stackrel{\left(\left.\omega\right|_{q} \vDash p\right) \cdot(\omega \vDash q)}{\omega \vDash p} \text {. }
$$

Example 6.1.4. We instantiate Proposition 6.1.3 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:

$$
P(E \mid D)=\omega \left\lvert\, \mathbf{1}_{D} \vDash \mathbf{1}_{E}=\frac{\omega \vDash \mathbf{1}_{D} \& \mathbf{1}_{E}}{\omega \vDash \mathbf{1}_{D}}=\frac{\omega \vDash \mathbf{1}_{D \cap E}}{P(D)}=\frac{P(D \cap E)}{P(D)}\right.
$$

2 Bayes rule, in the general formulation of Proposition 6.1.3 (2) specialises to the well known inversion property of conditional probabilities:

$$
P(E \mid D)=\left.\omega\right|_{\mathbf{1}_{D}} \vDash \mathbf{1}_{E}=\frac{\left(\left.\omega\right|_{\mathbf{1}_{E}} \vDash \mathbf{1}_{D}\right) \cdot\left(\omega \vDash \mathbf{1}_{D}\right)}{\omega \vDash \mathbf{1}_{E}}=\frac{P(D \mid E) \cdot P(E)}{P(D)} .
$$

We have explained updating $\left.\omega\right|_{p}$ as incorporating the evidence $p$ into the distribution $\omega$. Thus, one expects $p$ to be 'more true' in $\left.\omega\right|_{p}$ than in $\omega$. The next result shows that this is indeed the case. It plays an important role in learning see Chapter ??.

Theorem 6.1.5 (Validity-increase). For a distribution $\omega$ and a factor $p$ on the same set, if the validity $\omega \vDash p$ is non-zero, one has:

$$
\left.\omega\right|_{p} \vDash p \geq \omega \vDash p
$$

Proof. We recall the inequality $\omega \vDash p \& p \geq(\omega \vDash p)^{2}$ from Corollary 5.1.4(1), or from Exercise 4.2 .8 (2). Then, by the product rule from Theorem 6.1.3 (1),

$$
\left.\omega\right|_{p} \vDash p=\frac{\omega \vDash p \& p}{\omega \vDash p} \geq \frac{(\omega \vDash p)^{2}}{\omega \vDash p}=\omega \vDash p .
$$

We add a few more basic facts about conditioning.
Lemma 6.1.6. Let $\omega$ be distribution on $X$, with factors $p, q \in \operatorname{Fact}(X)$.
1 Conditioning with truth has no effect:

$$
\left.\omega\right|_{1}=\omega
$$

2 Conditioning with a point predicate gives a point state: for $a \in X$,

$$
\left.\omega\right|_{\mathbf{1}_{a}}=1|a\rangle, \quad \text { assuming } \omega(a) \neq 0 .
$$

3 Iterated conditionings commute:

$$
\left.\left(\left.\omega\right|_{p}\right)\right|_{q}=\left.\omega\right|_{p \& q}=\left.\left(\left.\omega\right|_{q}\right)\right|_{p}
$$

4 Conditioning is stable under multiplication of the factor with a scalar $s>0$ :

$$
\left.\omega\right|_{s \cdot p}=\left.\omega\right|_{p}
$$

5 Conditioning can be done component-wise, for product states and factors:

$$
\left.(\sigma \otimes \tau)\right|_{(p \otimes q)}=\left(\left.\sigma\right|_{q}\right) \otimes\left(\left.\tau\right|_{p}\right)
$$

6 Marginalisation of a conditioning with a (similarly) weakened predicate is conditioning of the marginalised state:

$$
\left.\omega\right|_{1 \otimes q}[0,1]=\left.\omega[0,1]\right|_{q} .
$$

7 For a function $f: X \rightarrow Y$, used as deterministic channel,

$$
\left.(f \gg=\omega)\right|_{q}=f \gg=\left(\left.\omega\right|_{q \circ f}\right) .
$$

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.

Proof. 1 Trivial since $\omega \vDash \mathbf{1}=1$.
2 Assuming $\omega(a) \neq 0$ we get for each $x \in X$,

$$
\left.\omega\left|\mathbf{1}_{a}(x)=\frac{\omega(x) \cdot \mathbf{1}_{a}(x)}{\omega \vDash \mathbf{1}_{a}}=\frac{\omega(a) \cdot 1|a\rangle(x)}{\omega(a)}=1\right| a\right\rangle(x) .
$$

3 It suffices to prove:

$$
\begin{aligned}
\left.\left(\left.\omega\right|_{p}\right)\right|_{q}(x)=\frac{\left.\omega\right|_{p}(x) \cdot q(x)}{\left.\omega\right|_{p} \vDash q} & =\frac{\omega(x) \cdot p(x) / \omega \vDash p \cdot q(x)}{\omega \vDash p \& q / \omega \vDash p} \quad \text { by Proposition6.1.3 (1) } \\
& =\frac{\omega(x) \cdot(p \& q)(x)}{\omega \vDash p \& q}=\left.\omega\right|_{p \& q}(x) .
\end{aligned}
$$

4 First we have:

$$
\begin{aligned}
\omega \vDash s \cdot p=\sum_{x \in X} \omega(x) \cdot(s \cdot p)(x) & =\sum_{x \in X} \omega(x) \cdot s \cdot p(x) \\
& =s \cdot\left(\sum_{x \in X} \omega(x) \cdot p(x)\right)=s \cdot(\omega \vDash p) .
\end{aligned}
$$

Next:

$$
\left.\omega\right|_{s \cdot p}(x)=\frac{\omega(x) \cdot(s \cdot p)(x)}{\omega \models s \cdot p}=\frac{\omega(x) \cdot s \cdot p(x)}{s \cdot(\omega \models p)}=\frac{\omega(x) \cdot p(x)}{\omega \models p}=\left.\omega\right|_{p}(x)
$$

5 For states $\sigma \in \mathcal{D}(X), \tau \in \mathcal{D}(Y)$ and factors $p$ on $X$ and $q$ on $Y$ one has:

$$
\begin{aligned}
\left(\left.(\sigma \otimes \tau)\right|_{(p \otimes q)}\right)(x, y) & =\frac{(\sigma \otimes \tau)(x, y) \cdot(p \otimes q)(x, y)}{(\sigma \otimes \tau) \vDash(p \otimes q)} \\
& =\frac{\sigma(x) \cdot \tau(y) \cdot p(x) \cdot q(y)}{(\sigma \vDash p) \cdot(\tau \vDash q)} \quad \text { by Lemma4.2.9 } \\
& =\frac{\sigma(x) \cdot p(x)}{\sigma \vDash p} \cdot \frac{\tau(y) \cdot q(y)}{\tau \vDash q} \\
& =\left(\left.\sigma\right|_{p}\right)(x) \cdot\left(\left.\tau\right|_{q}\right)(y) \\
& =\left(\left(\left.\sigma\right|_{p}\right) \otimes\left(\left.\tau\right|_{q}\right)\right)(x, y) .
\end{aligned}
$$

6 Let $\omega \in \mathcal{D}(X \times Y)$ and $q$ be a factor on $Y$; then for an element $y \in Y$,

$$
\begin{aligned}
\left(\left.\omega\right|_{1 \otimes q}[0,1]\right)(y) & =\left.\sum_{x \in X} \omega\right|_{1 \otimes q}(x, y) \\
& =\sum_{x \in X} \frac{\omega(x, y) \cdot(\mathbf{1} \otimes q)(x, y)}{\omega \vDash \mathbf{1} \otimes q} \\
& \stackrel{(4.7)}{=} \frac{(\omega[0,1])(y) \cdot q(y)}{\omega[0,1] \vDash q}=\left(\left.\omega[0,1]\right|_{q}\right)(y) .
\end{aligned}
$$

7 For $f: X \rightarrow Y, \omega \in \mathcal{D}(X)$ and $q \in \operatorname{Fact}(Y)$,

$$
\begin{aligned}
\left(\left.(f \gg \omega)\right|_{q}\right)(y)=\left.\mathcal{D}(f)(\omega)\right|_{q}(y) & =\frac{\mathcal{D}(f)(\omega)(y) \cdot q(y)}{\mathcal{D}(f)(\omega) \vDash q} \\
& =\sum_{x \in f^{-1}(y)} \frac{\omega(x) \cdot q(y)}{\sum_{x \in f^{-1}(y)} \omega(x) \cdot p(y)} \\
& =\sum_{x \in f^{-1}(y)} \frac{\omega(x) \cdot q(f(x))}{\sum_{x} \omega(x) \cdot q(f(x))} \\
& =\left.\sum_{x \in f^{-1}(y)} \omega\right|_{q \circ f}(x) \\
& =\mathcal{D}(f)\left(\left.\omega\right|_{q \circ f}\right)(y)=\left(f \gg\left(\left.\omega\right|_{q \circ f}\right)\right)(y) .
\end{aligned}
$$

In the beginning of this section we have defined updating $\left.\omega\right|_{p}$ for a state $\omega \in \mathcal{D}(X)$ and a factor $p: X \rightarrow \mathbb{R}_{\geq 0}$. 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 6.1.6 (4) shows that updating with the factor $p$ is the same as updating with the predicate $\frac{1}{B} \cdot p$. A further fact is that when we restrict the factor $p$ to the (finite) support $\operatorname{supp}(\omega) \subseteq$ $X$ of the distribution at hand, then it is bounded, for instance with bound $B=$ $\max \{p(x) \mid x \in \operatorname{supp}(\omega)\}$. 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 conclude with another example.
Example 6.1.7. 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 $u n i f_{X}=\sum_{0 \leq i \leq K} \frac{1}{K+1}|i\rangle \in$ $\mathcal{D}(X)$, for both urns. There is a predicate red : $X \rightarrow[0,1]$, namely $\operatorname{red}(i)=\frac{i}{K}$.

It thus gives the likelihood of a red ball. It is not hard to see that a priori we have:

$$
\text { unif }_{X} \vDash \text { red }=\frac{1}{2} .
$$

We use the order $\geq$ on $X$ as a sharp predicate geq: $X \times X \rightarrow[0,1]$, where $\operatorname{geq}(x, y)=1$ iff $x \geq y$. We can then form the updated joint state:

$$
\left.\left(u n i f_{X} \otimes u n i f_{X}\right)\right|_{\text {geq }} \in \mathcal{D}(X \times X)
$$

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:

$$
\begin{aligned}
& \left(\left(u n i f_{X} \otimes{\left.u n i f_{X}\right)}^{g_{\text {geq }}}\right)[1,0] \vDash \text { red }=\frac{2}{3}\right. \\
& \left(\left.\left(u n i f_{X} \otimes \text { unif }_{X}\right)\right|_{\text {geq }}\right)[0,1] \vDash \text { red }=\frac{1}{3} .
\end{aligned}
$$

Drawing red from the first now clearly has a higher probability. Details of the verification are left as an exercise below.

## Exercises

6.1.1 Check that $\left.\omega\right|_{p}$ can be described as $\operatorname{Flrn}(\omega \bullet p)$, using the action $\bullet$ from Exercise 4.2.15
6.1.2 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.
1 Take as 'at least one girl' and 'at least one boy' predicates on $\{G, B\} \times\{G, B\}:$

$$
\begin{aligned}
g & :=\left(\mathbf{1}_{B} \otimes \mathbf{1}_{B}\right)^{\perp}=\left(\mathbf{1}_{G} \otimes \mathbf{1}\right) \otimes\left(\mathbf{1}_{B} \otimes \mathbf{1}_{G}\right) \\
b & :=\left(\mathbf{1}_{G} \otimes \mathbf{1}_{G}\right)^{\perp}=\left(\mathbf{1}_{B} \otimes \mathbf{1}\right) \otimes\left(\mathbf{1}_{G} \otimes \mathbf{1}_{B}\right) .
\end{aligned}
$$

Compute unif $\otimes$ unif $\vDash g$ and unif $\otimes$ unif $\vDash b$.
2 Check that unif $\otimes$ unif $\left.\right|_{b} \vDash g=\frac{2}{3}$.
(For a description and solution of this problem in a special library for probabilistic programming of the functional programming language Haskell, see [50].)
6.1.3 In the setting of Example 6.1.2 (1) define a new predicate oddish $=$ evenish ${ }^{\perp}=\mathbf{1}$ - evenish.

1 Compute dice $\left.\right|_{o d d i s h}$

2 Prove the equation below, involving a convex sum of states on the left-hand side.

$$
\left.(\text { dice } \vDash \text { evenish }) \cdot \omega\right|_{\text {evenish }}+\left.(\text { dice } \vDash \text { oddish }) \cdot \omega\right|_{\text {oddish }}=\text { dice } .
$$

6.1.4 Let $p: X \rightarrow \mathbb{R}_{\geq 0}$ be a non-zero factor, on a finite set $X$. Check that updating the uniform distribution on $X$ with $p$, as in:

$$
\left.\operatorname{unif}_{X}\right|_{p},
$$

is a way of turning the factor $p$ into a distribution on $X$ via normalisation.
6.1.5 Prove that:

$$
\omega \vDash p^{2} \leq\left.\omega\right|_{p} \vDash p^{2} .
$$

Hint: Use Bayes' law in combination with Theorem 6.1.5 and Corollary 5.1.4
6.1.6 Let $p_{1}, \ldots, p_{n} \in \operatorname{Pred}(X)$ be a test, i.e. an $n$-tuple of predicates on $X$ with $p_{1} \otimes \cdots \otimes p_{n}=\mathbf{1}$. Let $\omega \in \mathcal{D}(X)$ and $q \in \operatorname{Fact}(X)$.
1 Check that $1=\sum_{i}\left(\omega \vDash p_{i}\right)$.
2 Prove what is called the law of total probability:

$$
\begin{equation*}
\omega=\left.\sum_{1 \leq i \leq n}\left(\omega \mid=p_{i}\right) \cdot \omega\right|_{p_{i}} \tag{6.2}
\end{equation*}
$$

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 6.1.3.
(The expression on the right-hand side in 6.2) is used to turn a test into a 'denotation' function $\mathcal{D}(X) \rightarrow \mathcal{D}(\mathcal{D}(X))$ in [132, 133], namely as $\left.\omega \mapsto \sum_{i}\left(\omega \vDash p_{i}\right)|\omega|_{p_{i}}\right\rangle$. This proces is described more abstractly in terms of 'hypernormalisation' in [76].)
3 Show that:

$$
\omega \vDash q=\sum_{1 \leq i \leq n} \omega \vDash q \& p_{i} .
$$

4 Prove now:

$$
\left.\omega\right|_{q} \vDash p_{i}=\frac{\omega \vDash q \& p_{i}}{\sum_{j} \omega \vDash q \& p_{j}} .
$$

6.1.7 Show that conditioning a convex sum of states yields a convex sum of
conditioned states: for $\sigma, \tau \in \mathcal{D}(X), p \in \operatorname{Fact}(X)$ and $r, s \in[0,1]$ with $r+s=1$,

$$
\begin{aligned}
& \left.(r \cdot \sigma+s \cdot \tau)\right|_{p} \\
& =\left.\frac{r \cdot(\sigma \vDash p)}{r \cdot(\sigma \vDash p)+s \cdot(\tau \vDash p)} \cdot \sigma\right|_{p}+\left.\frac{s \cdot(\tau \vDash p)}{r \cdot(\sigma \vDash p)+s \cdot(\tau \vDash p)} \cdot \tau\right|_{p} \\
& =\left.\frac{r \cdot(\sigma \vDash p)}{(r \cdot \sigma+s \cdot \tau) \vDash p} \cdot \sigma\right|_{p}+\left.\frac{s \cdot(\tau \vDash p)}{(r \cdot \sigma+s \cdot \tau) \vDash p} \cdot \tau\right|_{p} .
\end{aligned}
$$

6.1.8 Consider $\omega \in \mathcal{D}(X)$ and $p \in \operatorname{Fact}(X)$ where $p$ is non-zero, at least on the support of $\omega$. Check that updating $\omega$ with $p$ can be undone via updating with $\frac{1}{p}$.
6.1.9 This exercise will demonstrate that conditioning may both create and remove entwinedness of distributions.

1 Write yes $=\mathbf{1}_{1}: 2 \rightarrow[0,1]$, where $2=\{0,1\}$, and no $=$ yes $^{\perp}=\mathbf{1}_{0}$. Prove that the following conditioning of a non-entwined state,

$$
\tau:=\left.(\text { flip } \otimes \text { flip })\right|_{(\text {yes } \otimes y e s) \otimes(n o \otimes n o)}
$$

is entwined.
2 Consider the state $\omega \in \mathcal{D}(2 \times 2 \times 2)$ given by:

$$
\begin{aligned}
\left.\omega=\frac{1}{18} \right\rvert\, & 111\rangle+\frac{1}{9}|110\rangle+\frac{2}{9}|101\rangle+\frac{1}{9}|100\rangle \\
& +\frac{1}{9}|011\rangle+\frac{2}{9}|010\rangle+\frac{1}{9}|001\rangle+\frac{1}{18}|000\rangle
\end{aligned}
$$

Prove that $\omega$ 's first and third component are entwined:

$$
\omega[1,0,1] \neq \omega[1,0,0] \otimes \omega[0,0,1]
$$

3 Now let $\rho$ be the following conditioning of $\omega$ :

$$
\rho:=\left.\omega\right|_{1 \otimes \mathrm{y} e s \otimes \mathbf{1}}
$$

Prove that $\rho$ '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.
6.1.10 Show that for $\omega \in \mathcal{D}(X)$ and $p_{1}, p_{2} \in \operatorname{Fact}(X)$ one has:

$$
\left.(\Delta \gg \omega)\right|_{p_{1} \otimes p_{2}}=\Delta \gg=\left(\left.\omega\right|_{p_{1} \& p_{2}}\right) .
$$

Note that this is a consequence of Lemma 6.1.6,7.
6.1.11 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 $f: X \rightarrow Y$, updating the associated deterministic channel $\langle f\rangle: X \mapsto Y$ has no effect:

$$
\left.\langle f\rangle\right|_{q}=\langle f\rangle .
$$

6.1.12 Let $c: Z \leadsto X$ and $d: Z \rightsquigarrow Y$ be two channels with a common domain $Z$, and with factors $p \in \operatorname{Fact}(X)$ and $q \in \operatorname{Fact}(Y)$ on their codomains. Prove that the update of a tuple channel is the tuple of the updates:

$$
\left.\langle c, d\rangle\right|_{p \otimes q}=\left\langle\left. c\right|_{p},\left.d\right|_{q}\right\rangle .
$$

Prove also that for $e: U \leadsto X$ and $f: V \mapsto Y$,

$$
\left.(e \otimes f)\right|_{p \otimes q}=\left(\left.e\right|_{p}\right) \otimes\left(\left.f\right|_{q}\right)
$$

6.1.13 In [97] the influence of a predicate $p$ on a state $\omega$ is measured via the total variation distance $d\left(\omega,\left.\omega\right|_{p}\right)$. This influence can be zero, for the truth predicate $p=\mathbf{1}$.

Consider the set $\{H, T\}$ with state $\operatorname{flip}(r)=r|H\rangle+(1-r)|T\rangle$, and with predicate $p=\mathbf{1}_{H}$. Prove that $d\left(f \operatorname{flp}(r),\left.f l i p(r)\right|_{p}\right) \rightarrow 1$ as $r \rightarrow 0$.
6.1.14 Consider the situation in Example 6.1.7 and prove consecutively:
$\left.1 \quad\left(u^{\prime 2} f_{X} \otimes u^{\prime} i f_{X}\right)\right|_{g e q}=\sum_{i, j \in X, i \geq j} \frac{2}{(K+1)(K+2)}|i, j\rangle ;$
2 the first and second marginals are:

$$
\begin{aligned}
& \left(\left(\text { unif }_{X} \otimes{\left.u n i f_{X}\right)}^{g_{g e q}}\right)[1,0]=\sum_{0 \leq i \leq K} \frac{2(i+1)}{(K+1)(K+2)}|i\rangle\right. \\
& \left(\left.\left(u n i f_{X} \otimes u n i f_{X}\right)\right|_{g e q}\right)[0,1]=\sum_{0 \leq i \leq K} \frac{2(K+1-i)}{(K+1)(K+2)}|i\rangle
\end{aligned}
$$

3 the validities of the predicate red in these marginals are:

$$
\begin{aligned}
& \left(\left(u n i f_{X} \otimes{\left.u n i f_{X}\right)}^{g_{g e q}}\right)[1,0] \vDash \text { red }=\frac{2}{3}\right. \\
& \left(\left.\left(u n i f_{X} \otimes u n i f_{X}\right)\right|_{g e q}\right)[0,1] \vDash \text { red }=\frac{1}{3} .
\end{aligned}
$$

Hint: Recall Proposition 1.2.6(1), and (2).

### 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 [96]. 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 state on the domain of a channel $c: X \rightarrow$ $Y$.

1 For a factor $p \in \operatorname{Fact}(X)$, we define forward inference as transformation along $c$ of the state $\omega$ updated with $p$, as in:

$$
c \gg=\left.\omega\right|_{p} \in \mathcal{D}(Y) .
$$

This is also called causal reasoning.
2 For a factor $q \in \operatorname{Fact}(Y)$, backward inference is updating of $\omega$ with the transformed factor:

$$
\left.\omega\right|_{c \ll q} \in \mathcal{D}(X) .
$$

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 7.7

In both cases the distribution $\omega$ is often called the prior distribution or simply the prior. Similarly, $c \gg=\left.\omega\right|_{p}$ and $\left.\omega\right|_{c \ll q}$ 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

Theorem 6.2.2. The validity of a factor $q$ in a predicted state $c \gg=\omega$ is increased when $\omega$ is replaced by $\left.\omega\right|_{c \leqslant<q}$, in:

$$
c \gg=\left(\left.\omega\right|_{c \approx \ll q}\right) \vDash q \geq c \gg=\omega \vDash q,
$$

for $\omega \in \mathcal{D}(X), c: X \rightarrow Y$ and $q \in \operatorname{Fact}(Y)$.

Proof. Via the back-and-forth transformation of Proposition 4.3 .3 and the validity increase of Theorem 6.1.5:

$$
\begin{aligned}
c \gg\left(\left.\omega\right|_{c=<q}\right) \vDash q & =\left.\omega\right|_{c \approx<q} \vDash c=\ll q \\
& \geq \omega \vDash c=\ll q \\
& =c \gg=\omega \vDash q .
\end{aligned}
$$

In the remainder of this section we illustrate the forward and backward inferences mechanisms in several 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, before reading the analysis and the solution.

Example 6.2.3. We start with the following question from [158, 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: unif $=\frac{1}{2}|H\rangle+\frac{1}{2}|T\rangle$. The above description implicitly contains a channel $c:\{H, T\} \leadsto\{W, B\}$, determined by the two urns:

$$
\begin{aligned}
c(H) & =F \operatorname{lrn}(2|W\rangle+7|B\rangle) & c(T) & =F \operatorname{lrn}(5|W\rangle+6|B\rangle) \\
& =\frac{2}{9}|W\rangle+\frac{7}{9}|B\rangle & & =\frac{5}{11}|W\rangle+\frac{6}{11}|B\rangle .
\end{aligned}
$$

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 $c=\ll \mathbf{1}_{W}$ on the sample space $\{H, T\}$. It is given by:

$$
\left(c=\ll \mathbf{1}_{W}\right)(H)=\sum_{x \in\{W, B\}} c(H)(x) \cdot \mathbf{1}_{W}(x)=c(H)(W)=\frac{2}{9} .
$$

Similarly we get $\left(c=\ll \mathbf{1}_{W}\right)(T)=c(T)(W)=\frac{5}{11}$.
The answer that we are interested in is obtained by updating the prior unif with the transformed evidence $c=\ll \mathbf{1}_{W}$, as given by unif $\left.\right|_{c \approx<} \mathbf{1}_{H}$. This is an instance of backward inference.

In order to obtain the answer, we first we compute the validity:

$$
\text { unif } \begin{aligned}
\models c=\ll \mathbf{1}_{W} & =\operatorname{unif}(H) \cdot\left(c=\ll \mathbf{1}_{W}\right)(H)+\operatorname{unif}(T) \cdot\left(c=\ll \mathbf{1}_{W}\right)(T) \\
& =\frac{1}{2} \cdot \frac{2}{9}+\frac{1}{2} \cdot \frac{5}{11}=\frac{1}{9}+\frac{5}{22}=\frac{67}{198} .
\end{aligned}
$$

Then:

$$
\text { unif }\left.\right|_{c \approx<1_{W}}=\frac{1 / 2 \cdot 2 / 9}{67 / 198}|H\rangle+\frac{1 / 2 \cdot 5 / 11}{67 / 198}|T\rangle=\frac{22}{67}|H\rangle+\frac{45}{67}|T\rangle .
$$

Thus, the conditional probability of heads is $\frac{22}{67}$. The same outcome is obtained in [158], of course, but there via an application of Bayes' rule.

Example 6.2.4. Consider the following classical question from [175].
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 $c:\{G, B\} \leadsto\{G, B\}$ given by:

$$
c(G)=\frac{4}{5}|G\rangle+\frac{1}{5}|B\rangle \quad c(B)=\frac{1}{5}|G\rangle+\frac{4}{5}|B\rangle .
$$

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:

$$
\left.\omega\right|_{c=<1_{B}}=\frac{17}{29}|G\rangle+\frac{12}{29}|B\rangle \approx 0.5862|G\rangle+0.4138|B\rangle .
$$

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 $\omega$. It is argued in [175] that humans find it difficult to take such base rates (or priors) into account. This phenomenon is called base rate neglect, see also [62].

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 $Y=\{1,2, \ldots, 10\}$.

$$
q=\frac{1}{10} \cdot \mathbf{1}_{1}+\frac{3}{10} \cdot \mathbf{1}_{2}+\frac{3}{10} \cdot \mathbf{1}_{3}+\frac{2}{10} \cdot \mathbf{1}_{4}+\frac{1}{10} \cdot \mathbf{1}_{5} .
$$

Using the original mood distribution $\omega=\frac{1}{8}|p\rangle+\frac{3}{8}|n\rangle+\frac{1}{2}|o\rangle$ and channel $c: X \leadsto Y$ from Example 2.4.3, we can compute the validity of this predicate $q$ in the predicted state $c \gg \omega$ as:

$$
c \gg \omega \models q=\omega \vDash c=\ll q=\frac{299}{4000}=0.07475 .
$$

One can check that the updated state $\omega^{\prime}=\left.\omega\right|_{c \approx<q}$ obtained via backward inference is:

$$
\omega^{\prime}=\frac{77}{299}|p\rangle+\frac{162}{299}|n\rangle+\frac{60}{299}|o\rangle \approx 0.2575|p\rangle+0.5418|n\rangle+0.2007|o\rangle .
$$

Interestingly, after updating, the teacher has more realistic view in the sense that the validity of the predicate $q$ has risen to $c \gg=\omega^{\prime} \vDash q=\frac{15577}{149500} \approx 0.1042$. 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 (1.28) 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 $b: M \leadsto B$ describing the blood pressure distribution for each medicine type. This channel is obtained by column-wise frequentist learning:

$$
b(0)=\frac{2}{3}|H\rangle+\frac{1}{3}|L\rangle \quad b(1)=\frac{7}{9}|H\rangle+\frac{2}{9}|L\rangle \quad b(2)=\frac{5}{8}|H\rangle+\frac{3}{8}|L\rangle .
$$

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 $b \gg=\omega$ is $\frac{7}{10}|H\rangle+\frac{3}{10}|L\rangle$. It is the distribution that is learnt from the totals column in Table 1.28). 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 $\mathbf{1}_{E}: B \rightarrow[0,1]$. Then:

$$
\left.\omega \vDash \mathbf{1}_{E}=\frac{9}{20}+\frac{2}{5}=\frac{17}{20} \quad \text { so } \quad \omega\left|\mathbf{1}_{E}=\frac{9 / 20}{17 / 20}\right| 1\right\rangle+\frac{2 / 5}{17 / 20}|2\rangle=\frac{9}{17}|1\rangle+\frac{8}{17}|2\rangle .
$$

Forward reasoning, precisely as in Definition 6.2.1 (1), gives:

$$
\begin{aligned}
b \gg\left(\left.\omega\right|_{\mathbf{1}_{E}}\right) & =b \gg=\left(\frac{9}{17}|1\rangle+\frac{8}{17}|2\rangle\right) \\
& =\left(\frac{9}{17} \cdot \frac{7}{9}+\frac{8}{17} \cdot \frac{5}{8}\right)|H\rangle+\left(\frac{9}{17} \cdot \frac{2}{9}+\frac{8}{17} \cdot \frac{3}{8}\right)|H\rangle \\
& =\frac{12}{17}|H\rangle+\frac{5}{17}|L\rangle .
\end{aligned}
$$

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.

$$
\begin{aligned}
\left(b=\ll \mathbf{1}_{H}\right)(x) & =\sum_{y \in B} b(x)(y) \cdot \mathbf{1}_{H}(y)=b(x)(H) \\
\omega \models b=<\mathbf{1}_{H} & =\sum_{x \in M} \omega(x) \cdot\left(b=\ll \mathbf{1}_{H}\right)(x)=\sum_{x \in M} \omega(x) \cdot b(x)(H) \\
& =\frac{3}{20} \cdot \frac{2}{3}+\frac{9}{20} \cdot \frac{7}{9}+\frac{2}{5} \cdot \frac{5}{8}=\frac{7}{10} \\
\left.\omega\right|_{b=\ll \mathbf{1}_{H}} & =\sum_{x \in M} \frac{\omega(x) \cdot\left(b=\ll \mathbf{1}_{H}\right)(x)}{\omega \vDash b=\ll \mathbf{1}_{H}}|x\rangle \\
& =\frac{3 / 20 \cdot 2 / 3}{7 / 10}|0\rangle+\frac{9 / 20 \cdot 7 / 9}{7 / 10}|1\rangle+\frac{2 / 5 \cdot 5 / 8}{7 / 10}|2\rangle \\
& =\frac{1}{7}|0\rangle+\frac{1}{2}|1\rangle+\frac{5}{14}|2\rangle \approx 0.1429|0\rangle+0.5|1\rangle+0.3571|2\rangle .
\end{aligned}
$$

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 $q: B \rightarrow[0,1]$ 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:

$$
\begin{aligned}
\left.\omega\right|_{b \ll q} & =\frac{39}{272}|0\rangle+\frac{135}{272}|1\rangle+\frac{49}{136}|2\rangle \\
& \approx 0.1434|0\rangle+0.4963|1\rangle+0.3603|2\rangle .
\end{aligned}
$$

This slightly differs from the outcome with sharp evidence.
Example 6.2.7. The following question comes from [174] §6.1.3] (and is also used in [140]).

One fish is contained within the confines of an opaque fishbowl. The fish is equally likely to be a piranha or a goldfish. A sushi lover throws a piranha into the fish bowl alongside the other fish. Then, immediately, before either fish can devour the other, one of the fish is blindly removed from the fishbowl. The fish that has been removed from the bowl turns out to be a piranha. What is the probability that the fish that was originally in the bowl by itself was a piranha?

Let's use the letters ' p ' and ' g ' for piranha and goldfish. We are looking at a situation with multiple fish in a bowl, where we cannot distinguish the order. Hence we describe the contents of the bowl as a (natural) multiset over $\{p, g\}$, that is, as an element of $\mathcal{N}(\{p, g\})$. The initial situation can then be described as a distribution $\omega \in \mathcal{D}(\mathcal{N}(\{p, g\}))$ with:

$$
\left.\left.\left.\left.\omega=\frac{1}{2}|1| p\right\rangle\right\rangle+\frac{1}{2}|1| g\right\rangle\right\rangle .
$$

Adding a piranha to the bowl involves a function $A: \mathcal{N}(\{p, g\}) \rightarrow \mathcal{N}(\{p, g\})$, such that $A(\varphi)=(\varphi(p)+1)|p\rangle+\varphi(g)|g\rangle$. It forms a deterministic channel.
We use a piranha predicate $P: \mathcal{N}(\{p, g\}) \rightarrow[0,1]$ that gives the likelihood $P(\varphi)$ of taking a piranha from a multiset / bowl $\varphi$. Thus:

$$
P(\varphi):=\operatorname{Flrn}(\varphi)(p) \stackrel{\sqrt{2.5}}{-} \frac{\varphi(p)}{\varphi(p)+\varphi(g)} .
$$

We have now collected all ingredients to answer the question via backward inference along the deterministic channel $A$. It involves the following steps.

$$
\begin{aligned}
(A=\ll)(\varphi) & =P(A(\varphi))=\frac{\varphi(p)+1}{\varphi(p)+1+\varphi(g)} \\
\omega \vDash A=\ll P & =\frac{1}{2} \cdot(A=<P)(1|p\rangle)+\frac{1}{2} \cdot(A=<P)(1|g\rangle) \\
& =\frac{1}{2} \cdot \frac{1+1}{1+1+0}+\frac{1}{2} \cdot \frac{0+1}{0+1+1}=\frac{1}{2}+\frac{1}{4}=\frac{3}{4} \\
\left.\omega\right|_{A \approx<P} & \left.\left.\left.\left.=\frac{1 / 2 \cdot 1}{3 / 4}|1| p\right\rangle\right\rangle+\frac{1 / 2 \cdot 1 / 2}{3 / 4}|1| g\right\rangle\right\rangle \\
& \left.\left.\left.\left.=\frac{2}{3}|1| p\right\rangle\right\rangle+\frac{1}{3}|1| g\right\rangle\right\rangle .
\end{aligned}
$$

Hence the answer to the question in the beginning of this example is: $\frac{2}{3}$ probability that the original fish is a piranha.

Example 6.2.8. In [160, §20.1] a situation is described with five different bags, numbered $1, \ldots, 5$, each containing its own mixture of cherry (C) and lime (L) candies. This situation can be described via a candy channel:


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 [160, §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 $c=\ll \mathbf{1}_{L}: B \rightarrow[0,1]$ given by:

$$
c=\ll \mathbf{1}_{L}=\bigotimes_{i} c(i)(L) \cdot \mathbf{1}_{i}=\frac{1}{4} \cdot \mathbf{1}_{2} \otimes \frac{1}{2} \cdot \mathbf{1}_{3} \oslash \frac{3}{4} \cdot \mathbf{1}_{4} \otimes 1 \cdot \mathbf{1}_{5} .
$$

The question is what we learn about the bag distribution after observing this predicate 10 consecutive times? This involves computing:

$$
\begin{aligned}
\left.\omega\right|_{c \approx<\mathbf{1}_{L}} & =\frac{1}{10}|2\rangle+\frac{2}{5}|3\rangle+\frac{3}{10}|4\rangle+\frac{1}{5}|5\rangle \\
\left.\left.\omega\right|_{c \approx<\mathbf{1}_{L}}\right|_{c \approx<\mathbf{1}_{L}} & =\left.\omega\right|_{\left(c \approx<\mathbf{1}_{L}\right) \&\left(c=<\mathbf{1}_{L}\right)}=\left.\omega\right|_{\left(c=<\mathbf{1}_{L}\right)^{2}} \\
& =\frac{1}{26}|2\rangle+\frac{4}{13}|3\rangle+\frac{9}{26}|4\rangle+\frac{4}{13}|5\rangle \\
& \approx 0.0385|2\rangle+0.308|3\rangle+0.346|4\rangle+0.308|5\rangle \\
\left.\left.\left.\omega\right|_{c \approx<\mathbf{1}_{L}}\right|_{c \ll \mathbf{1}_{L}}\right|_{c \approx<\mathbf{1}_{L}} & =\left.\omega\right|_{\left(c \approx<\mathbf{1}_{L}\right) \&\left(c=<\mathbf{1}_{L}\right) \&\left(c=<\mathbf{1}_{L}\right)}=\left.\omega\right|_{\left(c \approx<\mathbf{1}_{L}\right)^{3}} \\
& =\frac{1}{76}|2\rangle+\frac{4}{19}|3\rangle+\frac{27}{76}|4\rangle+\frac{8}{19}|5\rangle \\
& \approx 0.0132|2\rangle+0.211|3\rangle+0.355|4\rangle+0.421|5\rangle \ldots
\end{aligned}
$$

Figure 20.1 in [160] gives a plot of these distributions; it is reconstructed 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.

Example 6.2.9. 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 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


Figure 6.1 Posterior, updated bag distributions $\left.\omega\right|_{\left(c \approx \ll 1_{L}\right)^{n}}$ for $n=0,1, \ldots, 10$, aligned vertically, after $n$ candy draws that all happen to be lime.
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 $(T)$.

$$
C=\left\{c, c^{\perp}\right\} \quad D=\{0,1,2,3,4,5,6\} \quad T=\{p, n\} .
$$

The test probabilities are then captured via a test channel $t: C \times D \leadsto T$, given in the following 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 .

$$
\begin{aligned}
t\left(c^{\perp}, i\right) & =\frac{\frac{1}{100}|p\rangle+\frac{99}{100}|n\rangle}{t(c, i)}= \begin{cases}1|n\rangle & \text { if } i=0 \text { or } i=1 \\
\frac{2}{10}|p\rangle+\frac{8}{10}|n\rangle & \text { if } i=2 \\
\frac{3}{10}|p\rangle+\frac{7}{10}|n\rangle & \text { if } i=3 \\
\frac{4}{10}|p\rangle+\frac{6}{10}|n\rangle & \text { if } i=4 \\
\frac{6}{10}|p\rangle+\frac{4}{10}|n\rangle & \text { if } i=5 \\
\frac{8}{10}|p\rangle+\frac{2}{10}|n\rangle & \text { if } i=6 .\end{cases}
\end{aligned}
$$




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 $\left(r|c\rangle+(1-r)\left|c^{\perp}\right\rangle\right) \otimes \varphi_{2}$, where $r \in[0,1]$ is the prior Covid probability. Similarly for N2, P5, N5.

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:

$$
\sigma_{2}=\frac{1}{4}|1\rangle+\frac{1}{2}|2\rangle+\frac{1}{4}|3\rangle \quad \sigma_{5}=\frac{1}{4}|4\rangle+\frac{1}{2}|5\rangle+\frac{1}{4}|6\rangle .
$$

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 $\omega=u n i f_{C}=\frac{1}{2}|c\rangle+\frac{1}{2}\left|c^{\perp}\right\rangle$.

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 $t=\ll \mathbf{1}_{p}$ 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:

$$
\left(\left.\left(\omega \otimes \sigma_{2}\right)\right|_{t \ll \mathbf{1}_{p}}\right)[1,0] \in \mathcal{D}(C) .
$$

We shall go through the computation step-by-step. First,

$$
\begin{aligned}
t: \ll \mathbf{1}_{p}= & \sum_{x \in C, y \in D} t(x, y)(p) \cdot \mathbf{1}_{(x, y)} \\
= & \frac{2}{10} \cdot \mathbf{1}_{(c, 2)}+\frac{3}{10} \cdot \mathbf{1}_{(c, 3)}+\frac{4}{10} \cdot \mathbf{1}_{(c, 4)}+\frac{6}{10} \cdot \mathbf{1}_{(c, 5)}+\frac{8}{10} \cdot \mathbf{1}_{(c, 6)} \\
& +\frac{1}{100} \cdot \mathbf{1}_{\left(c^{\perp}, 0\right)}+\frac{1}{100} \cdot \mathbf{1}_{\left(c^{\perp}, 1\right)}+\frac{1}{100} \cdot \mathbf{1}_{\left(c^{\perp}, 2\right)}+\frac{1}{100} \cdot \mathbf{1}_{\left(c^{\perp}, 3\right)} \\
& +\frac{1}{100} \cdot \mathbf{1}_{\left(c^{\perp}, 4\right)}+\frac{1}{100} \cdot \mathbf{1}_{\left(c^{\perp}, 5\right)}+\frac{1}{100} \cdot \mathbf{1}_{\left(c^{\perp}, 6\right)} .
\end{aligned}
$$

Then:

$$
\begin{aligned}
\omega \otimes \sigma_{2} \vDash t=\ll \mathbf{1}_{p} & =\sum_{x \in C, y \in D} \omega(x) \cdot \sigma_{2}(y) \cdot\left(t=\ll \mathbf{1}_{p}\right)(x, y) \\
& =\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{10}+\frac{1}{2} \cdot \frac{1}{4} \cdot \frac{3}{10}+\frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{100}+\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{100}+\frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{100} \\
& =\frac{40+30+1+2+1}{800}=\frac{37}{400} .
\end{aligned}
$$

But then:

$$
\begin{aligned}
\left.\left(\omega \otimes \sigma_{2}\right)\right|_{t=<\mathbf{1}_{p}}= & \frac{1 / 2 \cdot 1 / 2 \cdot 2 / 10}{37 / 400}|c, 2\rangle+\frac{1 / 2 \cdot 1 / 4 \cdot 3 / 10}{37 / 400}|c, 3\rangle+\frac{1 / 2 \cdot 1 / 4 \cdot 1 / 100}{37 / 400}\left|c^{\perp}, 1\right\rangle \\
& \quad+\frac{1 / 2 \cdot 1 / 2 \cdot 1 / 100}{37 / 400}\left|c^{\perp}, 2\right\rangle+\frac{1 / 2 \cdot 1 / 4 \cdot 1 / 100}{37 / 400}\left|c^{\perp}, 3\right\rangle \\
= & \frac{20}{37}|c, 2\rangle+\frac{15}{37}|c, 3\rangle+\frac{1}{74}\left|c^{\perp}, 1\right\rangle+\frac{1}{37}\left|c^{\perp}, 2\right\rangle+\frac{1}{74}\left|c^{\perp}, 3\right\rangle .
\end{aligned}
$$

Finally:

$$
\left(\left.\left(\omega \otimes \sigma_{2}\right)\right|_{t \ll 1_{p}}\right)[1,0]=\frac{35}{37}|c\rangle+\frac{2}{37}\left|c^{\perp}\right\rangle \approx 0.946|c\rangle+0.054\left|c^{\perp}\right\rangle .
$$

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:

$$
\left(\left.\left(\omega \otimes \sigma_{2}\right)\right|_{t \ll \mathbf{1}_{n}}\right)[1,0]=\frac{165}{363}|c\rangle+\frac{198}{363}\left|c^{\perp}\right\rangle \approx 0.455|c\rangle+0.545\left|c^{\perp}\right\rangle .
$$

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:

$$
\begin{aligned}
& \left(\left.\left(\omega \otimes \sigma_{5}\right)\right|_{t \ll \mathbf{1}_{p}}\right)[1,0]=\frac{60}{61}|c\rangle+\frac{1}{61}\left|c^{\perp}\right\rangle \approx 0.984|c\rangle+0.016\left|c^{\perp}\right\rangle \\
& \left(\left.\left(\omega \otimes \sigma_{5}\right)\right|_{t \approx<\mathbf{1}_{n}}\right)[1,0]=\frac{40}{139}|c\rangle+\frac{99}{139}\left|c^{\perp}\right\rangle \approx 0.288|c\rangle+0.712\left|c^{\perp}\right\rangle .
\end{aligned}
$$

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 $\omega=r|c\rangle+$ $(1-r)\left|c^{\perp}\right\rangle$ 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 6.2.10. 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 $F_{F}$ is:

$$
F=\{20,30,40, \ldots, 300\} \quad \text { with } \quad \text { unif }_{F}=\sum_{x \in F} \frac{1}{29}|x\rangle .
$$

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 (22). 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 $c: F \rightsquigarrow\{0,1, \ldots, K\}$.

$$
c(i):=b n[K]\left(\frac{K}{i}\right) \stackrel{2.1}{=} \sum_{0 \leq k \leq K}\binom{K}{k} \cdot\left(\frac{K}{i}\right)^{k} \cdot\left(\frac{i-K}{i}\right)^{K-k}|k\rangle .
$$

Once this is set up, we construct a posterior state by updating the prior with


Figure 6.3 The posterior fish number distribution after catching 5 marked fish, in multinomial mode, see Example 6.2.10
the information that five marked fish have been found. The latter is expressed as point predicate $\mathbf{1}_{5} \in \operatorname{Pred}(\{0, \ldots, K\})$ 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:

$$
\text { unif }\left._{F}\right|_{c=\ll \mathbf{1}_{5}}=\sum_{i \in F} \frac{(20 / i)^{5} \cdot(i-20 / i)^{15}}{\sum_{j}(20 / j)^{5} \cdot(j-20 / j)^{15}}|i\rangle .
$$

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 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 6.2.11. 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 [159, §4.8.3, Ex. 8h]).

Let's write $M=\left\{m, m^{\perp}\right\}$ 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 $\kappa=5|m\rangle+15\left|m^{\perp}\right\rangle$. 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 $20|m\rangle+(i-20)\left|m^{\perp}\right\rangle$ with $i \in F$.

We now use a channel:

$$
\{20,30, \ldots, 300\}=F \longrightarrow \stackrel{d}{\longrightarrow} \mathcal{N}[K](M) \cong\{0,1, \ldots, K\},
$$

given by:

$$
d(i):=\operatorname{hg}[K]\left(20|m\rangle+(i-20)\left|m^{\perp}\right\rangle\right) .
$$

The updated pond distribution is then:

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 already seen in Proposition 3.5.4 (1), 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 6.2.12. 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

[^8]

Figure 6.4 The posterior fish number distribution after catching 5 marked fish, in hypergeometric mode, see Example 6.2 .11
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:

$$
M i P=60 \quad \text { where } M i P \text { stands for: minimal production. }
$$

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:

$$
M a P=250
$$

Our aim is thus to derive a production distribution on the interval production space $P=[M i P, M a P]=[60,250] \subseteq \mathbb{N}$. We start from a uniform prior unif $f_{P} \in$ $\mathcal{D}(P)$.
The observation space $O$, containing the serial numbers of tanks that can be found is then the interval $O=[0, \mathrm{MaP}]$. 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


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 .12
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 $O=[0, M a P]$, for each production number $n \in P=[M i P, M a P]$. We define, for $i \in O$,

$$
q_{n}(i):= \begin{cases}n-i / n & \text { when } i<n \\ 0 & \text { when } i \geq n .\end{cases}
$$

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 $t: P \leadsto O$, from production to observation. We use the above predicates $q_{n}$ to update the uniform distribution on $O$, as in:

$$
t(n):=\text { unif }\left._{O}\right|_{q_{n}} .
$$

Figure 6.5 describes the observation distributions $t(\mathrm{MiP})$ for minimal production on the left. The distribution $t(\mathrm{MaP})$ for maximal production is on the right. Non-zero probabilities exist only for $i<M i P=60$ on the left, in decreasing order. When production is maximal, all numbers $i<M a P=250$ occur in the support of the distribution $t(\mathrm{MaP})$.

We now define a posterior distribution $\omega \in \mathcal{D}(P)$ via backward reasoning, using the discovered serial numbers as point predicates on $P$. Thus we take:

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 .12

## 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 $D=\left\{d, d^{+}\right\}$; 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 $c: D \leadsto T$;
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 Postive Predictive Value (PPV) and Negative Predictive

Value (NPV) of medical tests. Let's assume we have a disease prevalence (prior) given by $\omega=p|d\rangle+(1-p)\left|d^{\perp}\right\rangle$ with parameter $p \in[0,1]$ and a channel channel $c:\left\{d, d^{+}\right\} \rightsquigarrow\{p, n\}$ with sensitivity and specificity parameters $s e, s p \in[0,1]$ in:

$$
\begin{array}{lc}
\text { sensitivity: } & c(d) \\
\text { specificity: } & c\left(d^{\perp}\right) \\
\hline & =(1-s p)|p\rangle+(1-s e)|n\rangle \\
\text { sp| }|n\rangle .
\end{array}
$$

Check that:

$$
P P V:=\left.\omega\right|_{c \approx \ll 1_{p}}(d)=\frac{p \cdot s e}{p \cdot s e+(1-p) \cdot(1-s p)} .
$$

This is commonly expressed in medical textbooks as:
$P P V=\frac{\text { prevalence } \cdot \text { sensitivity }}{\text { prevalence } \cdot \text { sensitivity }+(1-\text { prevalence }) \cdot(1-\text { specificity })}$.
Check similarly that:

$$
N P V:=\left.\omega\right|_{c \lll 1_{n}}\left(d^{\perp}\right)=\frac{(1-p) \cdot s p}{p \cdot(1-s e)+(1-p) \cdot s p} .
$$

As an aside, the (positive) likelihood ratio $L R$ is the fraction:

$$
L R:=\frac{c(d)(p)}{c\left(d^{\perp}\right)(p)}=\frac{s e}{1-s p} .
$$

6.2.3 Give a channel-based analysis and answer to the following question from [158, 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 [9, §1.2]. We have two sets: $E=\left\{H, H^{\perp}\right\}$ about eating Hamburgers (or not), and $D=\left\{K, K^{\perp}\right\}$ 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:

$$
\omega=\frac{1}{2}|H\rangle+\frac{1}{2}\left|H^{\perp}\right\rangle \quad \text { and } \quad \sigma=\frac{1}{100,000}|K\rangle+\frac{99,999}{100,000}\left|K^{\perp}\right\rangle .
$$

1 Suppose that we know that $90 \%$ of the people who have KreuzfeldJacob disease eat Hamburgers. Use this additional information to define a channel $c: D \leadsto E$ with $c \gg=\sigma=\omega$.
2 Compute the probability of getting Kreuzfeld-Jacob for someone eating hamburgers (via backward inference).
6.2.6 Consider in the context of Example 6.2.9 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 $\omega$. Show that the posterior 'days' distribution is:

$$
\begin{aligned}
\left(\left.\left(\omega \otimes \sigma_{2}\right)\right|_{t \ll 1_{n}}\right)[0,1] & =\frac{199}{726}|1\rangle+\frac{179}{363}|2\rangle+\frac{169}{726}|3\rangle \\
& \approx 0.274|1\rangle+0.493|2\rangle+0.233|3\rangle .
\end{aligned}
$$

Explain why there a (small) shift 'forward', making the earlier days more likely in this posterior - with respect to the prior $\sigma_{2}$.
6.2.7 Consider the following challenge, copied from [168].
(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 $W=\{w d$, we $\}$ be a set with elements $w d$ for weekday and we for weekend, with prior distribution $\frac{5}{7}|w d\rangle+\frac{2}{7}|w e\rangle$. Use the Poisson distribution to define a channel bus: $W \rightarrow \mathcal{D}_{\infty}(\mathbb{N})$ 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 inference. One recurring topic is the crossover influence through updating of joint states. The abstract results of this section will be illustrated again 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 6.3, 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 $c: X \leadsto Y$ and $d: X \leadsto Z$. Define a joint distribution $\omega \in \mathcal{D}(Y \times Z)$ as:

$$
\begin{equation*}
\omega:=\langle c, d\rangle\rangle=\sigma= \tag{6.3}
\end{equation*}
$$



Then, for a factor $q \in \operatorname{Fact}(Y)$,

$$
\begin{equation*}
\left(\left.\omega\right|_{q \otimes 1}\right)[0,1]=d \gg=\left.\sigma\right|_{c \approx<q} . \tag{6.4}
\end{equation*}
$$

The right-hand-side of this equation (6.4) involves a forward inference of a backward inference. What happens can be described at an intuitive level using the string diagram in (6.3): the evidence $q$ is first pull backward (down) along channel $c$, the resulting factor $c=\ll q$ is used to update the distribution $\sigma$, and then the result is pushed forward (up) along channel $d$.

The left-hand-side of the equation (6.4) can also be written in terms of projections $\pi_{1}: Y \times Z \rightarrow Y$ and $\pi_{2}: Y \times Z \rightarrow Z$, namely as: $\pi_{2} \gg=\left.\omega\right|_{\pi_{1}=<q}$. The formulation with the projections will be generalised to an inference query in 6.16

Proof. We first note that:

$$
\begin{aligned}
\omega \vDash q \otimes \mathbf{1} & =\langle c, d\rangle \gg \sigma \vDash q \otimes \mathbf{1} & & \\
& =\sigma \models\langle c, d\rangle=\ll(q \otimes \mathbf{1}) & & \text { by Proposition4.3.3 } \\
& =\sigma \vDash(c=\ll q) \&(d=\ll \mathbf{1}) & & \text { by Lemma4.3.2] } \\
& =\sigma \vDash(c=\ll q) \& \mathbf{1} & & \\
& =\sigma \models c=\ll q . & &
\end{aligned}
$$

Then, for an element $z \in Z$,

$$
\begin{aligned}
\left(\left.\omega\right|_{q \otimes 1}\right)[0,1](z) & =\sum_{y \in Y} \frac{\omega(y, z) \cdot(q \otimes \mathbf{1})(y, z)}{\omega \vDash q \otimes \mathbf{1}} \\
& =\sum_{y \in Y} \sum_{x \in X} \frac{\sigma(x) \cdot c(x)(y) \cdot d(x)(z) \cdot q(y)}{\sigma \vDash c=\ll q} \\
& =\sum_{x \in X} \frac{\sigma(x) \cdot(c=\ll q)(x) \cdot d(x)(z)}{\sigma \vDash c=\ll q} \\
& =\left.\sum_{x \in X} \sigma\right|_{c=<q}(x) \cdot d(x)(z)=\left(d \gg=\left.\sigma\right|_{c=<q}\right)(z)
\end{aligned}
$$

At several earlier places in this book we have encountered string-diagram-
matic equations of the form (6.5) 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 $c: X \leadsto Y$ and $d: Y \leadsto X$ for which the following equation between graphs holds.


As a consequence, via marginalisation, $\sigma=d \gg=\tau$ and $\tau=c \gg=\sigma$.
For a factor $q \in \operatorname{Fact}(Y)$, backward inference along $c$ and forward inference along d coincide:

$$
\begin{equation*}
\left.\sigma\right|_{c \approx<q}=d \gg=\left.\tau\right|_{q} . \tag{6.6}
\end{equation*}
$$

In particular, when $q$ is a point predicate $\mathbf{1}_{y}$ for $y \in Y$ we get:

$$
\begin{equation*}
\left.\sigma\right|_{c=\ll \mathbf{1}_{y}}=d(y) . \tag{6.7}
\end{equation*}
$$

Thus, the channel $d$ is completely determined by $\sigma$ and $c$, under the condition that the validity $\sigma \vDash c \equiv<\mathbf{1}_{y}=c »=\sigma \vDash \mathbf{1}_{y}=(c \geqslant=\sigma)(y)$ is non-zero, for each $y \in Y$, so that the update $\left.\sigma\right|_{c=<\mathbf{1}_{y}}$ is well-defined. Equivalently, $c \geqslant=\sigma$ must have full support.

In the next chapter we shall see that the formula $\left.\sigma\right|_{c=\ll \mathbf{1}_{y}}$ defines a reversed 'dagger' channel $c_{\sigma}^{\dagger}: Y \rightsquigarrow X$ applied to $y$.

Proof. We show that an arbitrary predicate $p$ gets the same validity in both distributions, see Remark 4.2.11 We use Proposition 4.3 .3 several times.

$$
\begin{aligned}
& \left.\sigma\right|_{c: \ll q} \vDash p=\frac{\sigma \vDash(c=\ll q) \& p}{\sigma \vDash c=\ll q} \\
& =\frac{\sigma \vDash\langle c, i d\rangle=\ll(q \otimes p)}{c \gg=\sigma \vDash q} \\
& =\frac{\langle c, i d\rangle \gg=\sigma \vDash q \otimes p}{\pi_{1} \gg=(\langle c, i d\rangle \gg \sigma) \vDash q} \\
& =\frac{\langle i d, d\rangle \gg=\tau \vDash q \otimes p}{\pi_{1} \gg=(\langle i d, d\rangle \gg \tau) \vDash q} \quad \text { by assumption } \\
& =\frac{\tau \vDash q \&(d=\ll p)}{\tau \vDash q} \quad \text { again by Lemma4.3.2 (7) } \\
& =\left.\tau\right|_{q} \vDash d=\ll p \quad \text { by Theorem 6.1.3 (1) } \\
& =\left.d \gg \tau\right|_{q} \vDash p . \\
& \text { by Theorem 6.1.3 (1) } \\
& \text { by Lemma 4.3.2 (7) } \\
& \text { by Theorem 6.1.3 (1) }
\end{aligned}
$$

In the special case when $q$ is a point predicate $\mathbf{1}_{y}$ we get:

$$
\begin{aligned}
\left.\sigma\right|_{c=<\mathbf{1}_{y}} \stackrel{\sqrt[6.6]{=}}{=} d \gg=\left.\tau\right|_{\mathbf{1}_{y}} & =d \gg=1|y\rangle \quad \text { by Lemma 6.1.6 (2) } \\
& =d(y) .
\end{aligned}
$$

This result can be applied in several situations that we have seen before, involving an equation between graphs, of the form (6.5).

## Examples 6.3.3.

1 The archetypal equation of the form (6.5) is the one in Theorem 3.3.1, relating accumulation and arrangement (3.14). 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 $\vec{x} \in X^{K}$,

$$
\left.\operatorname{iid}[K](\omega)\right|_{\operatorname{acc}=<\mathbf{1}_{\varphi}}=\left.\operatorname{arr}(\varphi) \quad \operatorname{mn}[K](\omega)\right|_{\operatorname{arr}=<\mathbf{1}_{\vec{x}}}=\langle\operatorname{acc}\rangle(\vec{x}) .
$$

In these examples the prior distributions iid $[K](\omega)$ and $m n[K](\omega)$ do not play a role in the outcome. This phenomenon will be described in terms of 'sufficient statistic' in Section 7.6
2 In Exercise 3.2.12 we have seen that the draw-store-delete channel DSD can be obtained via updating, namely as: $\left.\omega \otimes m n[K](\omega)\right|_{\text {mcons }_{\nu}}=\operatorname{DSD}(\psi)$, for an arbitrary distribution $\omega$. It is an instance of Equation (6.7), since:

- the predicate $\mathrm{mcons}_{\psi}$ used in Exercise 3.2.12 can also be written as a transformation of a point predicate, namely as mcons $=\ll \mathbf{1}_{\psi}$;
- there is a graph equations diagram (3.13), that fits the pattern (6.5) in Theorem 6.3.2
3 In Theorem 3.4.4 we have seen the graph equations diagram 3.25). By applying Equation (6.7) we see how hypergeometric(-store) distributions can be obtained from multinomial ones via updating. For an urn $v \in \mathcal{N}[L+K](X)$,

$$
h g s[K](v)=\left.m n[K](\omega) \otimes m n[L](\omega)\right|_{\text {sum } \leqslant \ll 1_{v}} .
$$

By taking the first marginals on both sides we get the hypergeometric distribution as marginal of an update:

$$
h g[K](v)=\operatorname{hgs}[K](v)[1,0]=\left(\left.m n[K](\omega) \otimes m n[L](\omega)\right|_{\text {sum }=\ll 1_{v}}\right)[1,0] .
$$

4 In Theorem 3.3.8 we have seen Diagram (3.21) from which we can extra an equation, for an arbitrary distribution $\omega \in \mathcal{D}(X)$ with full support and for each $\lambda \in \mathbb{R}_{>0}$,

$$
m n[K](\omega)=\left.\left(\bigotimes_{x \in X} \operatorname{pois}[\lambda \cdot \omega(x)]\right)\right|_{\text {sum }=\ll 1_{K}}
$$

Formally, we have to add a frequencies isomorphism Freq: $\mathbb{N}^{X} \stackrel{\cong}{\Rightarrow} \mathcal{N}(X)$ on the right hand side, so that the types precisely match. This map Freq turns a tuple $t: X \rightarrow \mathbb{N}$ 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. [159, §6.4].

Recall from Definition6.1.1 that we use updating not only for distributions, written as $\left.\omega\right|_{p}$, but also for channels $c$, via elementwise updating: $\left.c\right|_{q}(x):=$ $\left.c(x)\right|_{q}$. This is used in the next result.

Theorem 6.3.4. Let $c: X \leadsto Y$ be a channel with a state $\omega \in \mathcal{D}(X)$ on its domain and a factor $q \in \operatorname{Fact}(Y)$ on its codomain. Then:

$$
\begin{equation*}
\left.(c \gg=\omega)\right|_{q}=\left.c\right|_{q} \gg=\left.\omega\right|_{c \approx<q} . \tag{6.8}
\end{equation*}
$$

Proof. For each $y \in Y$,

$$
\begin{aligned}
& \left.(c \gg=\omega)\right|_{q}(y)=\frac{(c »=\omega)(y) \cdot q(y)}{c \gg=\omega \vDash q} \\
& =\sum_{x \in X} \frac{\omega(x) \cdot c(x)(y) \cdot q(y)}{\omega \models c=\ll q} \\
& =\sum_{x \in X} \frac{\omega(x) \cdot(c=\ll q)(x)}{\omega \models c=\ll q} \cdot \frac{c(x)(y) \cdot q(y)}{(c=\ll q)(x)} \\
& =\left.\sum_{x \in X} \omega\right|_{c \approx<q}(x) \cdot \frac{c(x)(y) \cdot q(y)}{c(x) \models q} \\
& =\left.\left.\sum_{x \in X} \omega\right|_{c \lll q}(x) \cdot c(x)\right|_{q}(y) \\
& =\left.\left.\sum_{x \in X} \omega\right|_{c \approx<q}(x) \cdot c\right|_{q}(x)(y)=\left(\left.c\right|_{q} \gg=\left.\omega\right|_{c=<q}\right)(y) .
\end{aligned}
$$

This result has a number of useful consequences.
Corollary 6.3.5. For appropriately typed channels, states, and factors:
$\left.1(d \odot c)\right|_{q}=\left.\left.d\right|_{q} \odot c\right|_{d=<q}$, a form of chain rule;
$\left.2(\langle c, d\rangle \gg \omega)\right|_{p \otimes q}=\left\langle\left. c\right|_{p},\left.d\right|_{q}\right\rangle \gg=\left.\omega\right|_{(c \approx \ll p) \&(d=<q)} ;$
$\left.3((e \otimes f) \gg=\omega)\right|_{p \otimes q}=\left(\left.\left.e\right|_{p} \otimes f\right|_{q}\right) \gg=\left.\omega\right|_{(e=\ll p) \otimes(f \approx<q)}$.
Proof. 1 Since:

$$
\begin{aligned}
&\left.(d \odot c)\right|_{q}(x)=\left.(d \odot c)(x)\right|_{q}=\left.(d \gg=c(x))\right|_{q} \\
&\left.\left.\stackrel{|6.8|}{=} d\right|_{q} \gg c(x)\right|_{d \ll q} \\
&=\left.d\right|_{q} \gg=\left(\left.c\right|_{d \approx \ll q}(x)\right)=\left(\left.\left.d\right|_{q} \odot c\right|_{d \approx<q}\right)(x) .
\end{aligned}
$$

2 By Exercise 6.1.12 and Lemma 4.3.2 (7):

$$
\begin{aligned}
\left.(\langle c, d\rangle \gg \omega)\right|_{p \otimes q} & \stackrel{\sqrt{6.8}}{=}\left(\left.\langle c, d\rangle\right|_{p \otimes q} \gg=\left.\omega\right|_{\langle c, d\rangle ;<(p \otimes q)}\right. \\
& =\left\langle\left. c\right|_{p},\left.d\right|_{q}\right\rangle \gg=\left.\omega\right|_{(c \approx<p) \&(d \approx q)} .
\end{aligned}
$$

3 Similarly:

$$
\begin{aligned}
&\left.((e \otimes f) \gg \omega)\right|_{p \otimes q} \stackrel{\sqrt{6.8}}{=}\left(\left.(e \otimes f)\right|_{p \otimes q} \gg=\left.\omega\right|_{(e \otimes f)=<(p \otimes q)}\right. \\
&=\left.\left(\left.\left.e\right|_{p} \otimes f\right|_{q}\right) \gg \omega\right|_{(e=<p) \otimes(f=\langle q)} .
\end{aligned}
$$

## Exercises

6.3.1 Consider a joint distribution $\tau \in \mathcal{D}(X \times Y)$ together with two channel $c: X \rightsquigarrow U$ and $d: Y \rightsquigarrow V$, and define $\omega \in \mathcal{D}(U \times V)$ as:


Prove for a factor $q$ on $U$ that:

$$
\left(\left.\omega\right|_{q \otimes 1}\right)[0,1]=d \gg=\left(\left.\tau\right|_{(c=\langle q) \otimes 1}[1,0]\right) .
$$

6.3.2 Prove the two equations in Example 6.3.3 (1) directly, via the definition of updating.
6.3.3 Apply 6.7) to diagram 3.18.
6.3.4 Let $X$ be a finite set and $v \in \mathcal{N}(X)$ 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:

$$
\left.p l[K+N](v)\right|_{h g[K]=<\mathbf{1}_{\varphi}}=\mathcal{D}(-+\varphi)(p l[N](v+\varphi)) .
$$

6.3.5 $1 \quad$ 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 $\omega \in \mathcal{D}(X)$,

$$
\begin{aligned}
& \left.\left(m n\left[K_{1}\right](\omega) \otimes \cdots \otimes m n\left[K_{N}\right](\omega)\right)\right|_{\text {sum }=\ll \mathbf{1}_{\psi}} \\
& =\sum_{\varphi_{1} \leq K_{1} \psi, \ldots, \varphi_{N} \leq K_{N} \psi, \Sigma_{i} \varphi_{i}=\psi} \frac{\binom{\psi}{\varphi_{1}, \ldots, \varphi_{N}}}{\binom{K}{K_{1}, \ldots, K_{N}}}\left|\varphi_{1}, \ldots, \varphi_{N}\right\rangle .
\end{aligned}
$$

(Exercise 1.8 .5 guarantees that the probabilities in this sum over multisets add up to one.)
2 Describe the diagram of the form (6.5) that gives rise to the equation in the previous item.
6.3.6 Let $c: X \rightarrow Y$ be a channel, with state $\omega \in \mathcal{D}(X \times Z)$.

1 Prove that for a factor $p \in \operatorname{Fact}(Z)$,

$$
\left.((c \otimes i d) \gg=\omega)\right|_{\otimes p}=(c \otimes i d) \gg=\left.\omega\right|_{1_{\otimes p}} .
$$

2 Show also that for $q \in \operatorname{Fact}(Y \times Z)$,

$$
\left(\left.((c \otimes i d) \gg=\omega)\right|_{q}\right)[0,1]=\left(\left.\omega\right|_{(c \otimes i d) \ll q}\right)[0,1] .
$$

6.3.7 Let $c: X \leadsto Y$ be a channel with a distribution $\omega \in \mathcal{D}(X)$ and with two factors $p \in \operatorname{Fact}(X), q \in \operatorname{Fact}(Y)$. Check that:

$$
\left.c\right|_{q} \gg=\left(\left.\omega\right|_{p \&(c \approx<q)}\right)=\left.\left(c \gg=\left.\omega\right|_{p}\right)\right|_{q} .
$$

Check that Theorem6.3.4 is a special case, for $p=\mathbf{1}$.
6.3.8 Let $p \in \operatorname{Fact}(X)$ and $q \in \operatorname{Fact}(Y)$ be two factors on spaces $X, Y$.

1 Show that for two channels $c: Z \leadsto X$ and $d: Z \leadsto Y$ with a distribution $\sigma \in \mathcal{D}(Z)$ on their (common) domain, one has:

$$
\begin{aligned}
& \left.(\langle c, d\rangle \gg \sigma)\right|_{p \otimes q}[1,0]=\left.\left(c \gg=\left.\sigma\right|_{d=\ll q}\right)\right|_{p} \\
& \left.(\langle c, d\rangle \gg \sigma)\right|_{p \otimes q}[0,1]=\left.\left(d \gg=\left.\sigma\right|_{c \lll p}\right)\right|_{q}
\end{aligned}
$$

2 For channels $e: U \leadsto X$ and $d: V \nrightarrow Y$ with a joint state $\omega \in$ $\mathcal{D}(U \times V)$ one has:

$$
\begin{aligned}
& \left.((e \otimes f) \gg=\omega)\right|_{p \otimes q}[1,0]=\left.\left(e \gg=\left(\left.\omega\right|_{1 \otimes(f=<q)}[1,0]\right)\right)\right|_{p} \\
& \left.((e \otimes f) \gg=\omega)\right|_{p \otimes q}[0,1]=\left.\left(f \gg=\left(\left.\omega\right|_{(e=\langle p) \otimes 1}[0,1]\right)\right)\right|_{q} .
\end{aligned}
$$

### 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 [121]. 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.


Figure 6.7 The string diagram of the Asia Bayesian network, with node abbreviations: bronc $=$ bronchitis, dysp $=$ dyspnea, lung $=$ lung cancer, tub $=$ tuberculosis. The wires all have 2-element (yes/no) sets of the form $A=\left\{a, a^{\perp}\right\}$. These sets are used in the string diagram as types, annotating the wires.

Figure 6.9 reformulates these probability tables as distributions and channels, so that channel-based reasoning techniques can be used (as introduced in [96] for Bayesian networks). Specifically, this Figure 6.9 introduces distributions smoke $\in \mathcal{D}(S)$ and asia $\in \mathcal{D}(A)$ and channels lung: $S \leadsto L$, tub: $A \rightsquigarrow$ $T$, bronc: $S \leadsto B$, xray: $E \hookrightarrow X$, dysp: $B \times E \rightsquigarrow D$, either: $L \times T \rightsquigarrow E$.

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 $\mathbf{1}_{b^{\perp}}=\left(\mathbf{1}_{b}\right)^{\perp}: B \rightarrow[0,1]$ on the set $B=\left\{b, b^{\perp}\right\}$ 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 bronc: $S \rightarrow B$, and obtain a predicate bronc $=\ll \mathbf{1}_{b^{\perp}}$ on $S$. The latter can be used to update the smoking distribution on $S$. Subsequently, we can push the updated distribution forward

smoke $\quad P$ (bronc)

| $s$ | 0.6 |
| :---: | :--- |
| $s^{\perp}$ | 0.3 |


| bronc | either | $P$ (dysp) |
| :---: | :---: | :---: |
| $b$ | $e$ | 0.9 |
| $b$ | $e^{\perp}$ | 0.7 |
| $b^{\perp}$ | $e$ | 0.8 |
| $b^{\perp}$ | $e^{\perp}$ | 0.1 |

$$
\frac{P(\text { asia })}{0.01}
$$

| asia | $P($ tub $)$ |
| :---: | :---: |
| $a$ | 0.05 |
| $a^{\perp}$ | 0.01 |

either $\quad P$ (xray)

| $e$ | 0.98 |
| :---: | :---: |
| $e^{\perp}$ | 0.05 |


| lung | tub | $P$ (either) |
| :---: | :---: | :---: |
| $\ell$ | $t$ | 1 |
| $\ell$ | $t^{\perp}$ | 1 |
| $\ell^{\perp}$ | $t$ | 1 |
| $\ell^{\perp}$ | $t^{\perp}$ | 0 |

Figure 6.8 The conditional probability tables of the Asia Bayesian network, copied from [121 Table 1].

$$
\begin{aligned}
\operatorname{smoke} & =0.5|s\rangle+0.5\left|s^{\perp}\right\rangle & \text { asia } & =0.01|a\rangle+0.99\left|a^{\perp}\right\rangle \\
\operatorname{lung}(s) & =0.1|\ell\rangle+0.5\left|\ell^{\perp}\right\rangle & \operatorname{tub}(a) & =0.05|t\rangle+0.95\left|t^{\perp}\right\rangle \\
\operatorname{lung}\left(s^{\perp}\right) & =0.01|\ell\rangle+0.99\left|\ell^{\perp}\right\rangle & \operatorname{tub}\left(a^{\perp}\right) & =0.01|t\rangle+0.99\left|t^{\perp}\right\rangle \\
\operatorname{bronc}(s) & =0.6|b\rangle+0.4\left|b^{\perp}\right\rangle & \operatorname{xray}(e) & =0.98|x\rangle+0.02\left|x^{\perp}\right\rangle \\
\operatorname{bronc}\left(s^{\perp}\right) & =0.3|b\rangle+0.7\left|b^{\perp}\right\rangle & \operatorname{xray}\left(e^{\perp}\right) & =0.05|x\rangle+0.95\left|x^{\perp}\right\rangle \\
\operatorname{dysp}(b, e) & =0.9|d\rangle+0.1\left|d^{\perp}\right\rangle & \text { either }(\ell, t) & =1|e\rangle \\
\operatorname{dysp}\left(b, e^{\perp}\right) & =0.7|d\rangle+0.3\left|d^{\perp}\right\rangle & \text { either }\left(\ell, t^{\perp}\right) & =1|e\rangle \\
\operatorname{dysp}\left(b^{\perp}, e\right) & =0.8|d\rangle+0.2\left|d^{\perp}\right\rangle & \text { either }\left(\ell^{\perp}, t\right) & =1|e\rangle \\
\operatorname{dysp}\left(b^{\perp}, e^{\perp}\right) & =0.1|d\rangle+0.9\left|d^{\perp}\right\rangle & \text { either }\left(\ell^{\perp}, t^{\perp}\right) & =1\left|e^{\perp}\right\rangle
\end{aligned}
$$

Figure 6.9 The conditional probability tables from Figure 6.8. reformulated as distributions and channels.
along the lung channel lung: $S \leadsto L$ 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:

$$
\begin{equation*}
\text { lung }>=\left(\text { smoke }\left.\right|_{\text {bronc }=\left\langle\mathbf{1}_{b^{\perp}}\right.}\right)=0.0427|\ell\rangle+0.9573\left|\ell^{\perp}\right\rangle . \tag{6.9}
\end{equation*}
$$

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 $\mathbf{1}_{x} \in \operatorname{Pred}(X)$ for $X=\left\{x, x^{\perp}\right\}$.

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 lung $\otimes t u b: S \times A \rightarrow$ $L \times T$. The resulting predicate on $S \times A$ can be used to update the product state smoke $\otimes$ asia. Then we can take the first marginal to obtain the desired outcome. Thus we compute:

$$
\begin{align*}
& \left(\left.(\text { smoke } \otimes \text { asia })\right|_{(\text {lung } \otimes \text { tub })=\ll\left(\text { either }=\ll\left(\text { xray } \lll \mathbf{1}_{x}\right)\right)}\right)[1,0] \\
& =\left(\left.(\text { smoke } \otimes \text { asia })\right|_{(\text {xray } \odot \text { either } \odot(\text { lung } \otimes t u b))=\ll \mathbf{1}_{x}}\right)[1,0]  \tag{6.10}\\
& =0.6878|s\rangle+0.3122\left|s^{\perp}\right\rangle
\end{align*}
$$

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:

$$
\begin{aligned}
\sigma & :=(\text { bronc } \otimes \text { lung } \otimes \text { id }) \gg=((\Delta \otimes \text { tub }) \gg=(\text { smoke } \otimes \text { asia }) \\
& =(\langle\text { bronc, lung }\rangle \otimes \text { tub }) \gg(\text { smoke } \otimes \text { asia }) \\
& =(\langle\text { bronc, lung }\rangle>=\text { smoke }) \otimes(\text { tub } \gg=\text { asia }) .
\end{aligned}
$$

Recall that we write $\Delta$ for the copy channel, in this expression of type $S \leadsto$ $S \times S$.

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:

$$
p:=\left(\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1}_{t}\right) \&\left((\text { id } \otimes \text { either })=\ll\left(d y s=\ll \mathbf{1}_{d}\right)\right) .
$$

The result that we are after is now obtained via updating and marginalisation:

$$
\begin{equation*}
\left.\sigma\right|_{p}[0,1,0]=0.0558|\ell\rangle+0.9442\left|\ell^{\perp}\right\rangle \tag{6.11}
\end{equation*}
$$

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 $\sigma$, 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 $S \times L \times T$,

$$
\begin{aligned}
\tau & :=(\text { id } \otimes \text { lung } \otimes \text { id }) \gg=((\Delta \otimes \text { tub }) \gg=(\text { smoke } \otimes \text { asia })) \\
& =(\langle\text { id }, \text { lung }\rangle \otimes \text { tub }) \gg(\text { smoke } \otimes \text { asia }) \\
& =(\langle\text { id }, \text { lung }\rangle \gg=\text { smoke }) \otimes(\text { tub } \gg=\text { asia }) .
\end{aligned}
$$

The bronc channel is now used for predicate transformation in the predicate:

$$
q:=\left(\mathbf{1} \otimes \mathbf{1} \otimes \mathbf{1}_{t}\right) \&\left((\text { bronc } \otimes \text { either })=\ll\left(d y s=\ll \mathbf{1}_{d}\right)\right) .
$$

The same updated lung cancer distribution is now obtained as:

$$
\begin{equation*}
\left.\tau\right|_{q}[0,1,0]=0.0558|\ell\rangle+0.9442\left|\ell^{\perp}\right\rangle . \tag{6.12}
\end{equation*}
$$

The reason why the outcomes 6.11 and 6.12 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.
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 $\alpha$ 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 $\alpha$, the evidence $\mathbf{1}_{b^{\perp}}$ is a predicate on the set $B$, which is the second product


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 $\alpha$ on the product space $S \times B \times L \times D \times E \times X \times T \times A$.
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 $\pi_{2}=\ll \mathbf{1}_{b^{\perp}}$ on the entire product, via the second projection $\pi_{2}: S \times B \times L \times D \times E \times X \times T \times A \rightarrow B$. Similarly, we use the third projection $\pi_{3}$ for marginalisation, to get as outcome:

$$
\begin{align*}
\pi_{3} \gg=\alpha{\mid \pi_{2} \approx<\mathbf{1}_{b \perp}} & =\pi_{3} \gg=\left.\alpha\right|_{\mathbf{1} \otimes \mathbf{1}_{b \perp} \otimes \mathbf{1} \otimes 1 \otimes 1 \otimes 1 \otimes 1 \otimes 1}  \tag{6.13}\\
& =0.0427|\ell\rangle+0.9573\left|\ell^{\perp}\right\rangle .
\end{align*}
$$

The fact that we get the same outcome as in the earlier combination 6.9 of backward and forward inference follows essentially from Theorem 6.3.1

Similarly, the smoking distribution 6.10 of given a positive xray can be obtained via the joint distribution $\alpha$ as:

$$
\begin{align*}
\pi_{1} \gg=\left.\alpha\right|_{\pi_{6}=\left\langle\mathbf{1}_{x}\right.} & =\pi_{1} \gg=\left.\alpha\right|_{\mathbf{1} \otimes 1 \otimes 1 \otimes 1 \otimes 1 \otimes \mathbf{1}_{x} \otimes 1 \otimes 1} \\
& =0.6878|s\rangle+0.3122\left|s^{\perp}\right\rangle . \tag{6.14}
\end{align*}
$$

The lung cancer distributions (6.11) and 6.12) given dyspnoea and tuberculosis is:

$$
\begin{align*}
\pi_{3} \gg=\left.\alpha\right|_{\left(\pi_{4}=\left\langle\mathbf{1}_{d}\right) \&\left(\pi_{7} *<\mathbf{1}_{l}\right)\right.} & =\pi_{3} \gg=\left.\alpha\right|_{\mathbf{1}} \otimes 1 \otimes 1 \otimes \mathbf{1}_{d} \otimes 1 \otimes 1 \otimes 1 \otimes 1, \otimes 1  \tag{6.15}\\
& =0.0558|\ell\rangle+0.9442\left|\ell^{\perp}\right\rangle .
\end{align*}
$$

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 $\omega \in \mathcal{D}\left(X_{1} \times \cdots \times X_{n}\right)$ with evidence $p \in \operatorname{Fact}\left(X_{i}\right)$ and output in $X_{j}$ as distribution:

$$
\begin{equation*}
\pi_{j} \gg=\left.\omega\right|_{\pi_{i} \ll p} \in \mathcal{D}\left(X_{j}\right) . \tag{6.16}
\end{equation*}
$$

It combines forward and backward inference along projection channels. When the joint state $\omega$ is 'factored' in terms of channels - as illustrated in the string diagram in Figure 6.10 - such inference query 6.16 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 $\frac{4349}{5200}|d\rangle+\frac{851}{5200}\left|d^{\perp}\right\rangle$.
6.4.2 We continue with the wetness network.

1 Add copiers and wires to the string diagram in Figure 2.4 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 $\omega \in$ $\mathcal{D}(A \times B \times D \times C \times E)$.

$$
\begin{aligned}
& \omega=(i d \otimes i d \otimes w g \otimes i d \otimes s r) \\
& \quad \odot\left(i d \otimes \Delta_{2} \otimes \Delta_{3}\right) \\
& \quad \bullet(i d \otimes s p \otimes s r) \odot \Delta_{3} \odot w i .
\end{aligned}
$$

3 Explicitly, this joint distribution $\omega$ is:

```
\(\frac{399}{6250}|a, b, d, c, e\rangle+\frac{171}{6250}\left|a, b, d, c, e^{\perp}\right\rangle+\frac{27}{1250}\left|a, b, d, c^{\perp}, e^{\perp}\right\rangle\)
\(+\frac{21}{6250}\left|a, b, d^{\perp}, c, e\right\rangle+\frac{9}{6250}\left|a, b, d^{\perp}, c, e^{\perp}\right\rangle+\frac{3}{1250}\left|a, b, d^{\perp}, c^{\perp}, e^{\perp}\right\rangle\)
\(+\frac{672}{3125}\left|a, b^{\perp}, d, c, e\right\rangle+\frac{288}{3125}\left|a, b^{\perp}, d, c, e^{\perp}\right\rangle+\frac{168}{3125}\left|a, b^{\perp}, d^{\perp}, c, e\right\rangle\)
\(+\frac{72}{3125}\left|a, b^{\perp}, d^{\perp}, c, e^{\perp}\right\rangle+\frac{12}{125}\left|a, b^{\perp}, d^{\perp}, c^{\perp}, e^{\perp}\right\rangle+\frac{399}{20000}\left|a^{\perp}, b, d, c, e\right\rangle\)
\(+\frac{171}{20000}\left|a^{\perp}, b, d, c, e^{\perp}\right\rangle+\frac{243}{1000}\left|a^{\perp}, b, d, c^{\perp}, e^{\perp}\right\rangle+\frac{21}{20000}\left|a^{\perp}, b, d^{\perp}, c, e\right\rangle\)
\(+\frac{9}{20000}\left|a^{\perp}, b, d^{\perp}, c, e^{\perp}\right\rangle+\frac{27}{1000}\left|a^{\perp}, b, d^{\perp}, c^{\perp}, e^{\perp}\right\rangle+\frac{7}{1250}\left|a^{\perp}, b^{\perp}, d, c, e\right\rangle\)
\(+\frac{3}{1250}\left|a^{\perp}, b^{\perp}, d, c, e^{\perp}\right\rangle+\frac{7}{5000}\left|a^{\perp}, b^{\perp}, d^{\perp}, c, e\right\rangle+\frac{3}{5000}\left|a^{\perp}, b^{\perp}, d^{\perp}, c, e^{\perp}\right\rangle\)
\(+\frac{9}{100}\left|a^{\perp}, b^{\perp}, d^{\perp}, c^{\perp}, e^{\perp}\right\rangle\)
```

Express the two inferences in the previous exercise as suitable marginalisations of this joint distribution $\omega$, updated with suitably weakened predicates, as in (6.13), (6.14), 6.15).

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 6.11) and in 6.12 can be explained via Exercises 6.3.6 (2) and 4.3.9

### 6.5 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.
Let us start with a graphical description of an example of what is called a
hidden Markov model.


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 6.17), 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, how likely is it to see 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, but also in robotics and in biological sequences. 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 ?? and ??. Markov models are examples of probabilistic automata. Such automata will be studied separately, in Chapter ??.

Definition 6.5.1. A Markov model (or a Markov chain) is given by a set $X$ of 'internal positions', typically finite, with a 'transition' channel $t: X \leadsto X$ 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 $e: X \leadsto Y$, where $Y$ is a set of 'outputs'.

In the above illustration (6.17) we have as sets of positions and outputs:

$$
X=\{\text { Cloudy, Sunny, Rainy }\} \quad Y=\{\text { Stay-in, Go-out }\},
$$

with transition channel $t: X \leadsto X$,

$$
\begin{aligned}
t(\text { Cloudy }) & \left.\left.\left.\left.=\frac{1}{2} \right\rvert\, \text { Cloudy }\right\rangle \left.+\frac{1}{5} \right\rvert\, \text { Sunny }\right\rangle \left.+\frac{3}{10} \right\rvert\, \text { Rainy }\right\rangle \\
t \text { (Sunny }) & \left.\left.\left.\left.=\frac{3}{20} \right\rvert\, \text { Cloudy }\right\rangle \left.+\frac{4}{5} \right\rvert\, \text { Sunny }\right\rangle \left.+\frac{1}{20} \right\rvert\, \text { Rainy }\right\rangle \\
t(\text { Rainy }) & \left.\left.\left.\left.=\frac{1}{5} \right\rvert\, \text { Cloudy }\right\rangle \left.+\frac{1}{5} \right\rvert\, \text { Sunny }\right\rangle \left.+\frac{3}{5} \right\rvert\, \text { Rainy }\right\rangle,
\end{aligned}
$$

and emission channel $e: X \leadsto Y$,

$$
\begin{aligned}
e(\text { Cloudy }) & \left.\left.\left.=\frac{1}{2} \right\rvert\, \text { Stay-in }\right\rangle \left.+\frac{1}{2} \right\rvert\, \text { Go-out }\right\rangle \\
e(\text { Sunny }) & \left.\left.\left.=\frac{1}{5} \right\rvert\, \text { Stay-in }\right\rangle \left.+\frac{4}{5} \right\rvert\, \text { Go-out }\right\rangle \\
e(\text { Rainy }) & \left.\left.\left.=\frac{9}{10} \right\rvert\, \text { Stay-in }\right\rangle \left.+\frac{1}{10} \right\rvert\, \text { Go-out }\right\rangle .
\end{aligned}
$$

An initial state is missing in the picture 6.17.
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 $t: X \leadsto X$ is an endochannel on $X$, that is, a channel from the space $X$ to itself. As a function, it is of the form $t: X \rightarrow \mathcal{D}(X)$; it is an instance of a coalgebra, that is, a map of the form $A \rightarrow F(A)$ for a functor $F$, see Section ?? 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 $\sigma$ and transition channel $t$ one can form successor states via state transformation:

```
\(\sigma\)
```



```
\(t^{n} \gg=\sigma\)
```

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 6.5.6 below.
One interesting problem in the area of Markov chains is to find a 'stationary' state $\sigma_{\infty}$ with $t \gg \sigma_{\infty}=\sigma_{\infty}$, see Exercise 6.5 .2 for an illustration, and also Exercise 2.4.18 for a sufficient condition.
 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.

### 6.5.1 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 $1 \stackrel{\sigma}{\rightarrow} X \xrightarrow{t} X \xrightarrow{e} Y$. For each $n \in \mathbb{N}$ a channel $\langle e, t\rangle_{n}: X \mapsto$ $Y^{n} \times X$ is defined in the following manner:

$$
\begin{align*}
\langle e, t\rangle_{0} & :=\left(X \xrightarrow{\text { id }} X \cong \mathbf{1} \times X \cong Y^{0} \times X\right) \\
\langle e, t\rangle_{n+1} & :=\left(X \xrightarrow{\langle e, t\rangle_{n}} Y^{n} \times X \xrightarrow{\text { idn} \otimes\langle e, t\rangle} Y^{n} \times(Y \times X) \cong Y^{n+1} \times X\right) . \tag{6.18}
\end{align*}
$$

We recall that the tuple $\langle e, t\rangle$ of channels is $(e \otimes t) \odot \Delta$, see Definition 2.4.4 3. With these tuples we can form a joint state $\left.\langle e, t\rangle_{n}\right\rangle>\sigma \in \mathcal{D}\left(Y^{n} \times X\right)$. As a string
diagram it looks as follows.


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 $\mathbf{1}_{y}: Y \rightarrow[0,1]$, for $y \in Y$, but, as mentioned, here we allow more general observables $Y \rightarrow \mathbb{R}$.
Definition 6.5.2. Let $\mathcal{H}=(1 \stackrel{\sigma}{\gtrdot} X \xrightarrow{t} X \xrightarrow{e} Y)$ be a hidden Markov model and let $\vec{p}=p_{1}, \ldots, p_{n}$ be a list of observables on $Y$. The validity $\mathcal{H} \vDash \vec{p}$ of this sequence $\vec{p}$ in the model $\mathcal{H}$ is defined via the tuples (6.18) as:

$$
\begin{align*}
& \mathcal{H} \vDash \vec{p}:=\left.\left(\langle e, t\rangle_{n}\right\rangle=\sigma\right)[1, \ldots, 1,0] \vDash p_{1} \otimes \cdots \otimes p_{n} \\
&\left.\stackrel{4.7]}{=}\langle e, t\rangle_{n}\right\rangle=\sigma \vDash p_{1} \otimes \cdots \otimes p_{n} \otimes \mathbf{1} \\
& \stackrel{4.11]}{=} \sigma \vDash\langle e, t\rangle_{n}=\ll\left(p_{1} \otimes \cdots \otimes p_{n} \otimes \mathbf{1}\right) . \tag{6.20}
\end{align*}
$$

The marginalisation mask $[1, \ldots, 1,0]$ contains $n$ times the number 1. It ensures that the $X$ outcome in (6.19) 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. [15, 158] or [108, App. A]. An alternative algorithm is described in Exercise 6.5.4.

Proposition 6.5.3. The HMM-validity 6.20 can be computed as:

$$
\begin{align*}
& \mathcal{H} \vDash \vec{p}=\sigma \vDash\left(e=\ll p_{1}\right) \& \\
& t=\ll\left(\left(e=\ll p_{2}\right) \&\right.  \tag{6.21}\\
& t=\ll\left(\left(e=\ll p_{3}\right) \& \cdots\right. \\
&\left.\left.t=\ll\left(e=\ll p_{n}\right) \cdots\right)\right) .
\end{align*}
$$

This validity can be calculated recursively in forward manner as:

$$
\sigma \vDash \alpha(\vec{p}) \quad \text { where } \quad\left\{\begin{aligned}
\alpha([q]) & =e=\ll q \\
\alpha\left(\left[q_{1}\right]+\vec{q}\right) & =\left(e=\ll q_{1}\right) \&(t=\ll \alpha(\vec{q})) .
\end{aligned}\right.
$$

Alternatively, this validity can be calculated recursively in backward manner as:

$$
\sigma \vDash \beta(\vec{p}, \mathbf{1}) \text { where }\left\{\begin{aligned}
\beta([q]) & =q \\
\beta\left(\vec{q}+\left[q_{n}, q_{n+1}\right]\right) & =\beta\left(\vec{q}+\left[\left(e=\ll q_{n}\right) \&\left(t=\ll q_{n+1}\right)\right]\right) .
\end{aligned}\right.
$$

Proof. We first prove, by induction on $n \geq 1$ that for observables $p_{i}$ on $Y$ and $q$ on $X$ one has:

$$
\begin{align*}
& \langle e, t\rangle_{n}=\ll\left(p_{1} \otimes \cdots \otimes p_{n} \otimes q\right) \\
& =\left(e=\ll p_{1}\right) \& t=\ll\left(\left(e=\ll p_{2}\right) \& t=\ll\left(\cdots t=\ll\left(\left(e=\ll p_{n}\right) \& t=<q\right) \cdots\right)\right) \tag{*}
\end{align*}
$$

The base case $n=1$ is easy by Lemma 4.3.2, 77:

$$
\langle e, t\rangle_{1}=\ll\left(p_{1} \otimes q\right)=\left(e=\ll p_{1}\right) \&(t=\ll q)
$$

For the induction step we reason as follows.

$$
\begin{aligned}
& \langle e, t\rangle_{n+1}=\ll\left(p_{1} \otimes \cdots \otimes p_{n} \otimes p_{n+1} \otimes q\right) \\
& =\langle e, t\rangle_{n}=<\left(\left(i d^{n} \otimes\langle t, e\rangle\right)=\ll\left(p_{1} \otimes \cdots \otimes p_{n} \otimes p_{n+1} \otimes q\right)\right) \\
& =\langle e, t\rangle_{n}=\ll\left(p_{1} \otimes \cdots \otimes p_{n} \otimes\left(\langle e, t\rangle=\ll\left(p_{n+1} \otimes q\right)\right)\right. \\
& =\langle e, t\rangle_{n}=\ll\left(p_{1} \otimes \cdots \otimes p_{n} \otimes\left(\left(e=\ll p_{n+1}\right) \&(t=\ll q)\right)\right) \quad \text { as just shown } \\
& \stackrel{(\mathrm{IH})}{=}\left(e=\lll p_{1}\right) \& t=\ll\left(\left(e=\ll p_{2}\right) \& t=\ll(\cdots\right. \\
& \left.\left.\quad t=\ll\left(\left(e \lll p_{n}\right) \& t=\ll\left(\left(e=\ll p_{n+1}\right) \&(t=\ll q)\right)\right) \cdots\right)\right) .
\end{aligned}
$$

We can now prove Equation (6.21):

```
\(\mathcal{H} \vDash \vec{p}\)
\(\stackrel{6.20}{=} \sigma \models\langle e, t\rangle_{n}=\ll\left(p_{1} \otimes \cdots \otimes p_{n} \otimes \mathbf{1}\right)\)
    \(\stackrel{(*)}{=}\left(e=\ll p_{1}\right) \& t=\ll\left(\left(e=\ll p_{2}\right) \& t=\ll\left(\cdots t=\ll\left(\left(e=\lll p_{n}\right) \& t=\ll \mathbf{1}\right) \cdots\right)\right)\)
    \(=\left(e=\ll p_{1}\right) \& t=\ll\left(\left(e=\ll p_{2}\right) \& t=\ll\left(\cdots t=\ll\left(e=\ll p_{n}\right) \cdots\right)\right)\).
```


### 6.5.2 Filtering

Given a sequence $\vec{p}$ of factors, one can compute their validity $\mathcal{H} \vDash \vec{p}$ in a HMM $\mathcal{H}$, as described above. But we 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 6.5.4. Let $\mathcal{H}=(\mathbf{1} \stackrel{\sigma}{\gtrdot} X \stackrel{t}{\mapsto} X \stackrel{e}{\mapsto} Y)$ be a hidden Markov model and let $\vec{p}=p_{1}, \ldots, p_{n}$ be a list of factors on $Y$. It gives rise to the filtered sequence of states $\sigma_{1}, \sigma_{1}, \ldots, \sigma_{n+1} \in \mathcal{D}(X)$ following the observe-update-proceed principle:

$$
\sigma_{1}:=\sigma \quad \text { and } \quad \sigma_{i+1}:=t \gg=\left.\sigma_{i}\right|_{e \ll} p_{i} .
$$

In the terminology of Definition 6.2.1, the definition of the state $\sigma_{i+1}$ involves both forward and backward inference. Below we show that the final state $\sigma_{n+1}$ in the filtered sequence can also be obtained via crossover inference on a joint state, obtained via the tuple channels (6.18). This fact gives a theoretical justification, but is of little practical relevance - since joint states quickly become too big to handle.

Proposition 6.5.5. In the context of Definition 6.5.4

$$
\sigma_{n+1}=\left.\left(\langle e, t\rangle_{n} \gg=\sigma\right)\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes 1}[0, \ldots, 0,1]
$$

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.

$$
\begin{array}{ll}
\left.(\langle e, t\rangle \gg=\sigma)\right|_{p_{1} \otimes 1}[0,1] & \\
=\pi_{2} \gg\left(\left\langle\left. e\right|_{p_{1}}, t\right\rangle \gg=\left.\sigma\right|_{\langle e, t\rangle\left\langle\ll\left(p_{1} \otimes 1\right)\right.}\right) & \\
\text { by Corollary 6.3.5 (2) } \\
=\left.t \gg \sigma\right|_{e=\left\langle p_{1}\right.} & \\
=\sigma_{2} . & \text { by Lemma4.3.2 (7) } \\
&
\end{array}
$$

The induction step requires a bit more work:

$$
\begin{aligned}
& \left.\left(\langle e, t\rangle_{n+1}\right\rangle=\sigma\right)\left.\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes p_{n+1} \otimes 1}[0, \ldots, 0,0,1] \\
& \left.=\pi_{n+2} \gg=\left.\left(\left(\left(i d^{n} \otimes\langle e, t\rangle\right)\right\rangle=\left(\langle e, t\rangle_{n} \gg=\sigma\right)\right)\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes p_{n+1} \otimes \mathbf{1}}\right) \\
& \stackrel{6.87}{-} \pi_{n+2} \gg=\left(\left.\left(i d^{n} \otimes\langle e, t\rangle\right)\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes p_{n+1} \otimes 1}\right. \\
& \left.\left.\gg=\left(\langle e, t\rangle_{n}\right\rangle=\sigma\right)\left.\right|_{\left(d^{n} \otimes\langle e, t\rangle\right)<\left\langle\left(p_{1} \otimes \cdots \otimes p_{n} \otimes p_{n+1} \otimes \mathbf{1}\right)\right.}\right) \\
& \left.=\left(\pi_{n+2} \gg=\left(i d^{n} \otimes\left\langle\left. e\right|_{p_{n+1}}, t\right\rangle\right)\right) \gg=\left.\left(\left(\langle e, t\rangle_{n}\right\rangle=\sigma\right)\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes\left(\left(e=\left\langle\left\langle p_{n+1}\right) \&(t-<1)\right)\right.\right.}\right) \\
& =\left(t \odot \pi_{n+1}\right) \gg=\left(\left.\left(\langle e, t\rangle_{n} \gg \sigma\right)\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes\left(e=\ll p_{n+1}\right)}\right) \\
& =t \gg=\left(\left(\left.\left.\left(\langle e, t\rangle_{n} \gg=\sigma\right)\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes 1}\right|_{1 \otimes \cdots \otimes 1 \otimes\left(e \lll p_{n+1}\right)}[0, \ldots, 0,1]\right)\right. \\
& \text { by Lemma 6.1.6(3) and Lemma 4.2.10 (1) } \\
& =t \gg=\left(\left(\left.\left.\left(\langle e, t\rangle_{n} \gg=\sigma\right)\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes 1}[0, \ldots, 0,1]\right|_{e \approx\left\langle p_{n+1}\right.}\right)\right. \\
& \text { by Lemma 6.1.6 } 6 \\
& \stackrel{(\mathrm{IH})}{=} t \geqslant=\left(\left.\sigma_{n+1}\right|_{e=<p_{n+1}}\right) \\
& =\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 6.1.3

Corollary 6.5.6. Still in the context of Definition 6.5.4. 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:

$$
\sigma_{n+1} \vDash q=\frac{\left.\langle e, t\rangle_{n}\right\rangle=\sigma \vDash p_{1} \otimes \cdots \otimes p_{n} \otimes q}{\mathcal{H} \vDash \vec{p}} .
$$

Proof. This follows from Bayes' rule, in Proposition 6.1.3 (1):

$$
\begin{aligned}
\sigma_{n+1} \vDash q & =\left.\left(\langle e, t\rangle_{n} \gg=\sigma\right)\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes \mathbf{1}}[0, \ldots, 0,1] \vDash q \\
& =\left.\left(\langle e, t\rangle_{n} \gg \sigma\right)\right|_{p_{1} \otimes \cdots \otimes p_{n} \otimes \mathbf{1}} \vDash \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes q \\
& =\frac{\left(\langle e, t\rangle_{n} \gg \sigma\right) \vDash\left(p_{1} \otimes \cdots \otimes p_{n} \otimes \mathbf{1}\right) \&(\mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes q)}{\left(\langle e, t\rangle_{n} \gg \sigma\right) \vDash p_{1} \otimes \cdots \otimes p_{n} \otimes \mathbf{1}} \\
& =\frac{\left(\langle e, t\rangle_{n} \gg \sigma\right) \vDash p_{1} \otimes \cdots \otimes p_{n} \otimes q}{\mathcal{H} \vDash \vec{p}} .
\end{aligned}
$$

The last equation uses Lemma 4.2.10 1 and Definition 6.5 .2

### 6.5.3 Metropolis-Hastings

A question with practical relevance is the following. Suppose we have a nonempty multiset $\varphi \in \mathcal{M}(X)$, so with real-valued multiplicities, how can we sample from its normalisation $\operatorname{Flrn}(\varphi)=\frac{1}{\|\varphi\|} \cdot \varphi \in \mathcal{D}(X)$, without computing the size $\|\varphi\|$ ? When $\varphi$ 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 $\varphi$ into a Markov chain $c: X \leadsto X$ for which $\operatorname{Flrn}(\varphi)$ is a stationary state, forming a stationary state: $c »=\operatorname{Flrn}(\varphi)=\operatorname{Flrn}(\varphi)$, 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 $x_{n+1} \leftarrow c\left(x_{n}\right)$. One obtains a sequence $x_{0}, \ldots, x_{N}$ whose accumulation $\operatorname{acc}\left(x_{0}, \ldots, x_{N}\right) \in \mathcal{D}(X)$ approaches $\operatorname{Flrn}(\varphi)$, as $N$ increases, see e.g. [22, 173] for more information.

We sketch how to obain the channel $c: X \leadsto X$, from a multiset $\varphi \in \mathcal{M}(X)$. We assume that $\varphi$ has full support, i.e. that $X=\operatorname{supp}(\varphi)$ 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 $r: X \leadsto X$.

One first defines a function $\alpha: X \rightarrow \mathcal{M}(X)$ as:

$$
\alpha(x):=\sum_{y \in X} \min \left(1, \frac{\varphi(y)}{\varphi(x)}\right)|y\rangle .
$$

The channel $c: X \leadsto X$ is then defined as:

$$
c(x)(y):= \begin{cases}\frac{\alpha(x)(y)}{|X|} & \text { if } x \neq y \\ 1-\sum\left\{\left.\frac{\alpha(x)(z)}{|X|} \right\rvert\, z \in X, z \neq x\right\} & \text { if } x=y\end{cases}
$$

The second clause ensures that each $c(x)$ is a distribution,
We need to show is that $c \gg F \operatorname{lrn}(\varphi)=F \operatorname{lrn}(\varphi)$. 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:

$$
\begin{aligned}
& \alpha(u)=1|u\rangle+1|v\rangle+1|w\rangle \\
& \alpha(v)=\frac{\varphi(u)}{\varphi(v)}|u\rangle+1|v\rangle+\frac{\varphi(w)}{\varphi(v)}|w\rangle \\
& \alpha(w)=\frac{\varphi(u)}{\varphi(w)}|u\rangle+1|v\rangle+1|w\rangle .
\end{aligned}
$$

And:

$$
\begin{aligned}
& c(u)=\frac{1}{3}|u\rangle+\frac{1}{3}|v\rangle+\frac{1}{3}|w\rangle \\
& c(v)=\frac{\varphi(u)}{3 \varphi(v)}|u\rangle+\left(1-\frac{\varphi(u)}{3 \varphi(v)}-\frac{\varphi(w)}{3 \varphi(v)}\right)|v\rangle+\frac{\varphi(w)}{3 \varphi(v)}|w\rangle \\
& c(w)=\frac{\varphi(u)}{3 \varphi(w)}|u\rangle+\frac{1}{3}|v\rangle+\left(1-\frac{\varphi(u)}{3 \varphi(w)}-\frac{1}{3}\right)|w\rangle .
\end{aligned}
$$

We check that $\operatorname{Flrn}(\varphi)$ is a stationary state.

$$
\begin{aligned}
(c \gg \operatorname{Flrn}(\varphi))(u) & =\frac{1}{\|\varphi\|} \cdot \sum_{x \in X} \varphi(x) \cdot c(x)(u) \\
& =\frac{1}{\|\varphi\|} \cdot\left(\varphi(u) \cdot \frac{1}{3}+\varphi(v) \cdot \frac{\varphi(u)}{3 \varphi(v)}+\varphi(w) \cdot \frac{\varphi(u)}{3 \varphi(w)}\right) \\
& =\frac{1}{\|\varphi\|} \cdot \varphi(u)=\operatorname{Flrn}(\varphi)(u) .
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& (c \gg=F \operatorname{lrn}(\varphi))(v) \\
& =\frac{1}{\|\varphi\|} \cdot\left(\varphi(u) \cdot \frac{1}{3}+\varphi(v) \cdot\left(1-\frac{\varphi(u)}{3 \varphi(v)}-\frac{\varphi(w)}{3 \varphi(v)}\right)+\varphi(w) \cdot \frac{1}{3}\right) \\
& =\frac{1}{\|\varphi\|} \cdot \varphi(v)=F \operatorname{lrn}(\varphi)(v) \\
& (c \gg=\operatorname{lrn}(\varphi))(w) \\
& =\frac{1}{\|\varphi\|} \cdot\left(\varphi(u) \cdot \frac{1}{3}+\varphi(v) \cdot \frac{\varphi(w)}{3 \varphi(v)}+\varphi(w) \cdot\left(1-\frac{\varphi(u)}{3 \varphi(w)}-\frac{1}{3}\right)\right) \\
& =\frac{1}{\|\varphi\|} \cdot \varphi(w)=\operatorname{Flrn}(\varphi)(w) .
\end{aligned}
$$

The proof may be generalised to arbitrary multisets, but the many case distinctions easily obfuscate what is going on.

## Exercises

6.5.1 Consider the HMM example 6.17 with initial state $\sigma=1 \mid$ Cloudy $\rangle$.

1 Compute successive states $t^{n} \gg=\sigma$ for $n=0,1,2,3$.
2 Compute successive observations $e \gg\left(t^{n} \gg=\sigma\right)$ 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 6.5.4, with these same three (point) observations yields as final outcome:

$$
\begin{aligned}
& \left.\left.\left.\left.\frac{1867}{6696} \right\rvert\, \text { Cloudy }\right\rangle \left.+\frac{347}{1395} \right\rvert\, \text { Sunny }\right\rangle \left.+\frac{15817}{33480} \right\rvert\, \text { Rainy }\right\rangle \\
& \approx 0.279 \mid \text { Cloudy }\rangle+0.249 \mid \text { Sunny }\rangle+0.472 \mid \text { Rainy }\rangle .
\end{aligned}
$$

6.5.2 Consider the transition channel $t$ associated with the HMM example 6.17. Check that in order to find a stationary state $\sigma_{\infty}=x \mid$ Cloudy $\rangle+$ $y \mid$ Sunny $\rangle+z \mid$ Rainy $\rangle$ one has to solve the equations:

$$
\begin{aligned}
& x=\frac{1}{2} x+\frac{3}{20} y+\frac{1}{5} z \\
& y=\frac{1}{5} x+\frac{4}{5} y+\frac{1}{5} z \\
& z=\frac{3}{10} x+\frac{1}{20} y+\frac{3}{5} z
\end{aligned}
$$

Deduce that $\left.\sigma_{\infty}=\frac{1}{4} \right\rvert\,$ Cloudy $\rangle \left.+\frac{1}{2} \right\rvert\,$ Sunny $\rangle \left.+\frac{1}{4} \right\rvert\,$ Rainy $\rangle$ and doublecheck that $t \gg \sigma_{\infty}=\sigma_{\infty}$.
6.5.3 (The set-up of this exercise is copied from machine learning lecture notes of Doina Precup.) Consider a 5-state hallway of the form:


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 $t: X \mapsto X$ and $e: X \mapsto Y$ for a robot in this hallway are given by:

$$
\begin{array}{ll}
t(1)=\frac{3}{4}|1\rangle+\frac{1}{4}|2\rangle & e(1)=1|3\rangle \\
t(2)=\frac{1}{4}|1\rangle+\frac{1}{2}|2\rangle+\frac{1}{4}|3\rangle & e(2)=1|2\rangle \\
t(3)=\frac{1}{4}|2\rangle+\frac{1}{2}|3\rangle+\frac{1}{4}|4\rangle & e(3)=1|2\rangle \\
t(4)=\frac{1}{4}|3\rangle+\frac{1}{2}|3\rangle+\frac{1}{4}|5\rangle & e(4)=1|2\rangle \\
t(5)=\frac{1}{4}|4\rangle+\frac{3}{4}|5\rangle & e(5)=1|3\rangle .
\end{array}
$$

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 $\left[\mathbf{1}_{2}, \mathbf{1}_{2}, \mathbf{1}_{3}, \mathbf{1}_{2}, \mathbf{1}_{3}, \mathbf{1}_{3}\right]$.
1 Check that $(\sigma, t, e) \vDash \alpha=\frac{3}{512}$.
2 Next we filter with the sequence $\alpha$. Show that it leads succesively to the following states $\sigma_{i}$ as in Definition 6.5.4

$$
\begin{aligned}
\sigma_{1}:=\sigma & =1|3\rangle \\
\sigma_{2} & =\frac{1}{4}|2\rangle+\frac{1}{2}|3\rangle+\frac{1}{4}|4\rangle \\
\sigma_{3} & =\frac{1}{16}|1\rangle+\frac{1}{4}|2\rangle+\frac{3}{8}|3\rangle+\frac{1}{4}|4\rangle+\frac{1}{16}|5\rangle \\
\sigma_{4} & =\frac{3}{8}|1\rangle+\frac{1}{8}|2\rangle+\frac{1}{8}|4\rangle+\frac{3}{8}|5\rangle \\
\sigma_{5} & =\frac{1}{8}|1\rangle+\frac{1}{4}|2\rangle+\frac{1}{4}|3\rangle+\frac{1}{4}|4\rangle+\frac{1}{8}|5\rangle \\
\sigma_{6} & =\frac{3}{8}|1\rangle+\frac{1}{8}|2\rangle+\frac{1}{8}|4\rangle+\frac{3}{8}|5\rangle \\
\sigma_{7} & =\frac{3}{8}|1\rangle+\frac{1}{8}|2\rangle+\frac{1}{8}|4\rangle+\frac{3}{8}|5\rangle .
\end{aligned}
$$

6.5.4 Apply Bayes' rule to the validity formulation 6.21 in order to prove the correctness of the following HMM validity algorithm.

$$
\begin{aligned}
& (\sigma, t, e) \vDash[]:=1 \\
& (\sigma, t, e) \vDash\left[p_{1}\right]+\vec{p}:=\left(\sigma \models e=\ll p_{1}\right) \cdot\left(\left(\left.t \gg \sigma\right|_{e=\ll p_{1}}, t, e\right) \vDash \vec{p}\right) .
\end{aligned}
$$

(Notice the connection with filtering from Definition 6.5.4)
6.5.5 A random walk is a Markov model $d: \mathbb{Z} \leadsto \mathbb{Z}$ given by $d(n)=r \mid n-$ $1\rangle+(1-r)|n+1\rangle$ 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 $d \gg=\sigma, d^{2} \gg=\sigma, d^{3} \gg=\sigma, \ldots$ Which pattern emerges?
2 Prove that for $K \in \mathbb{N}$,

$$
\begin{aligned}
d^{K} \gg \sigma & =\sum_{0 \leq k \leq K} b n[K](1-r)(k)|2 k-K\rangle \\
& =\sum_{0 \leq k \leq K}\binom{K}{k} \cdot(1-r)^{k} \cdot r^{K-k}|2 k-K\rangle
\end{aligned}
$$

6.5.6 A Markov chain $X \leadsto X$ 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 [158, 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 $R=\left\{r, r^{\perp}\right\}$ for the state space of rain and no-rain outcomes, and capture the above probabilities via a channel $c: R \times R \mapsto R$.
2 Turn this channel $c$ into a Markov chain $\left\langle\pi_{2}, c\right\rangle: R \times R \leadsto R \times R$, 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 $\left\langle\pi_{2}, c\right\rangle$ both as a function and as a string diagram.
3 Generalise this approach to a history of length $N>1$ : turn a channel $X^{N} \leadsto X$ into a Markov model $X^{N} \rightsquigarrow X^{N}$, where the relevant history is incorporated into the sample space.
6.5.7 Use the approach of the prevous exercise to turn a hidden Markov model into a Markov model.
6.5.8 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.
6.5.9 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 6.5.3 to obtain a channel $c: X \leadsto X$ with $c »=\operatorname{Flrn}(\varphi)=\operatorname{Flrn}(\varphi)$.

### 6.6 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 4.4.3 that $\bar{p}^{\bullet}: \mathcal{N}(X) \rightarrow \mathbb{R}$ is the multiplicative extension of an observable $p: X \rightarrow \mathbb{R}$, see, defined by $\bar{p}^{\bullet}(\varphi)=\prod_{x \in X} p(x)^{\varphi(x)}$. Proposition 4.4.4 (2) we already saw that the validity $m n[K](\omega) \vDash \bar{p}^{\bullet}$ in a multinomial distribution equals the $K$-power $(\omega \vDash p)^{K}$ of the validity in the original distribution. We now extend this result to updating.

Theorem 6.6.1. Let $\omega \in \mathcal{D}(X)$ be given. For a factor (or predicate) $p: X \rightarrow$ $\mathbb{R}_{\geq 0}$ and a number $K \in \mathbb{N}$, one has the following results about updating with the multiplicative extension $\bar{p}^{\bullet}: \mathcal{N}(X) \rightarrow \mathbb{R}_{\geq 0}$

1 By updating a multinomial distribution with the multiplicative extension $\bar{p}^{\boldsymbol{*}}$ one still has a multinomial distribution, with an updated urn:

$$
\left.\operatorname{mn}[K](\omega)\right|_{\vec{p}} \cdot=\operatorname{mn}[K]\left(\left.\omega\right|_{p}\right) .
$$

2 Similarly, for the parallel multinomial law pml from Section 3.6.

$$
\left.\left.\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right)\right|_{\vec{p}^{\bullet}}=\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right|_{p}\right\rangle\right) .
$$

3 Updating a Poisson point process does not only effect the distribution $\omega$ but also the rate $\lambda$ :

$$
\left.\operatorname{Pmn}[\lambda](\omega)\right|_{p^{*}}=\operatorname{Pmn}[\lambda \cdot(\omega \vDash p)]\left(\left.\omega\right|_{p}\right) .
$$

Proof. 1 Via the validity formula of Proposition 4.4.4(2), for $\varphi \in \mathcal{N}[K](X)$,

$$
\begin{aligned}
\left.m n[K](\omega)\right|_{\bar{p}}(\varphi) & =\frac{m n[K](\omega)(\varphi) \cdot \bar{p}^{\bullet}(\varphi)}{m n[K](\omega) \vDash \bar{p}^{\bullet}} \\
& =(\varphi) \cdot \frac{\left(\prod_{x} \omega(x)^{\varphi(x)}\right) \cdot\left(\prod_{x} p(x)^{\varphi(x)}\right)}{(\omega \vDash p)^{K}} \\
& =(\varphi) \cdot \prod_{x}\left(\frac{\omega(x) \cdot p(x)}{\omega \models p}\right)^{\varphi(x)} \quad \text { since }\|\varphi\|=K \\
& =\left.(\varphi) \cdot \prod_{x} \omega\right|_{p}(x)^{\varphi(x)} \\
& =m n[K]\left(\left.\omega\right|_{p}\right)(\varphi) .
\end{aligned}
$$

2 Similarly, via Proposition 4.4.7.

$$
\begin{aligned}
& \left.\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right)\right|_{\bar{p}} \\
& =\sum_{\varphi_{\in \mathcal{N}[K](X)}} \frac{p m l\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right)(\varphi) \cdot \bar{p}^{\bullet}(\varphi)}{p m l\left(\sum_{i} n_{i}\left|\omega_{i}\right\rangle\right) \vDash \bar{p}^{\bullet}}|\varphi\rangle \quad \text { where } K=\sum_{i} n_{i} \\
& \stackrel{3.33}{=} \sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)} \frac{\prod_{i} m n\left[n_{i}\right]\left(\omega_{i}\right)\left(\varphi_{i}\right) \cdot \bar{p}^{\bullet}\left(\varphi_{i}\right)}{\prod_{i}\left(\omega_{i} \vDash p\right)^{n_{i}}}\left|\sum_{i} \varphi_{i}\right\rangle \\
& =\sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)} \prod_{i}\left(\varphi_{i}\right) \cdot \frac{\prod_{x}(\omega(x) \cdot p(x))^{\varphi_{i}(x)}}{\left(\omega_{i} \vDash p\right)^{n_{i}}}\left|\sum_{i} \varphi_{i}\right\rangle \\
& =\sum_{i, \varphi_{i} \in N\left[n_{i}\right](X)} \prod_{i}\left(\varphi_{i}\right) \cdot \prod_{x}\left(\frac{\omega(x) \cdot p(x)}{\omega_{i} \vDash p}\right)^{\varphi_{i}(x)}\left|\sum_{i} \varphi_{i}\right\rangle \\
& =\left.\sum_{i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right](X)} \prod_{i}\left(\varphi_{i}\right) \cdot \prod_{x} \omega_{i}\right|_{p}(x)^{\varphi_{i}(x)}\left|\sum_{i} \varphi_{i}\right\rangle \\
& =\sum_{\left.i, \varphi_{i} \in \mathcal{N}\left[n_{i}\right]\right](X)} \prod_{i} m n\left[n_{i}\right]\left(\left.\omega_{i}\right|_{p}\right)\left(\varphi_{i}\right)\left|\sum_{i} \varphi_{i}\right\rangle \\
& \left.=\operatorname{pml}\left(\sum_{i} n_{i}\left|\omega_{i}\right|_{p}\right\rangle\right) .
\end{aligned}
$$

3 Using Proposition 4.4.4(2) and (4) together with the item (1), we get for an arbitrary multiset $\varphi \in \mathcal{N}(X)$, say with of size $K=\|\varphi\|$,

$$
\begin{aligned}
\left.\operatorname{Pmn}[\lambda](\omega)\right|_{\bar{p}^{\bullet}}(\varphi) & =\frac{\operatorname{Pmn}[\lambda](\omega)(\varphi) \cdot \bar{p}^{\bullet}(\varphi)}{\operatorname{Pmn}[\lambda](\omega) \vDash \bar{p}^{\bullet}} \\
& \stackrel{\text { 3.522 }}{=} \frac{e^{-\lambda} \cdot \lambda^{K} \cdot \operatorname{mn}[K](\varphi) \cdot \bar{p}^{\bullet}(\varphi)}{K!\cdot e^{-\lambda \cdot\left(\omega \vDash p^{\perp}\right)}} \\
& =\frac{e^{-\lambda \cdot\left(1-\left(\omega \vDash p^{\perp}\right)\right)} \cdot \lambda^{K} \cdot(\omega \vDash p)^{K}}{K!} \cdot \frac{\operatorname{mn}[K](\varphi) \cdot \bar{p}^{\bullet}(\varphi)}{\operatorname{mn}[K](\omega) \vDash \bar{p}^{\bullet}} \\
& =\frac{e^{-\lambda \cdot(\omega \vDash p) \cdot(\lambda \cdot(\omega \vDash p))^{K}}}{K!} \cdot \operatorname{mn}[K]\left(\left.\omega\right|_{p}\right)(\varphi) \\
& \stackrel{\frac{3.52]}{=}}{ } \operatorname{Pmn}[\lambda \cdot(\omega \vDash p)]\left(\left.\omega\right|_{p}\right)(\varphi) .
\end{aligned}
$$

Example 6.6.2. We continue Example 3.9.2, 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 $p:\{a, b, c, d\} \rightarrow[0,1]$ given by $p(a)=p(b)=1$ and $p(c)=p(d)=0$. The validity $\omega \vDash p$ is $\omega(a)+\omega(b)=\frac{1}{8}+\frac{1}{6}=\frac{7}{24}$, and the updated distribution $\rho$ is:

$$
\rho:=\left.\omega\right|_{p}=\frac{1 / 8}{7 / 24}|a\rangle+\frac{1 / 6}{7 / 24}|b\rangle=\frac{3}{7}|a\rangle+\frac{4}{7}|b\rangle .
$$

The adapted rate for these two diseases is $\lambda \cdot(\omega \vDash p)=5 \cdot \frac{7}{24}=\frac{35}{24}$. Thus the arrival probabilities, for multisets of these two diseases $a, b$ only, are:

$$
\begin{aligned}
& \left.\operatorname{Pmn}[5](\omega)\right|_{\bar{p}^{*}}=\operatorname{Pmn}\left[5 \cdot \frac{7}{24}\right]\left(\left.\omega\right|_{p}\right)=\operatorname{Pmn}\left[\frac{35}{24}\right](\rho) \\
& =\operatorname{pois}\left[\frac{35}{24}\right](0) \cdot \operatorname{mn}[0](\rho)+\operatorname{pois}\left[\frac{35}{24}\right](1) \cdot \operatorname{mn}[1](\rho) \\
& \quad+\operatorname{pois}\left[\frac{35}{24}\right](2) \cdot \operatorname{mn}[2](\rho)+\cdots \\
& \left.\left.\left.=e^{-35 / 24} \cdot\left(1|\mathbf{0}\rangle+\frac{5}{8}|1| a\right\rangle\right\rangle+\frac{5}{6}|1| b\right\rangle\right\rangle \\
& \left.\left.\left.\left.\left.\left.\left.\quad+\frac{25}{128}|2| a\right\rangle\right\rangle+\frac{25}{48}|1| a\right\rangle+1|b\rangle\right\rangle+\frac{25}{72}|2| b\right\rangle\right\rangle+\cdots\right) \\
& =0.233|\mathbf{0}\rangle+0.145|1| a\rangle\rangle+0.194|1| b\rangle\rangle \\
& \quad+0.0454|2| a\rangle\rangle+0.121|1| a\rangle+1|b\rangle\rangle+0.0808|2| b\rangle\rangle+\cdots
\end{aligned}
$$

We now restrict to multiplicative extensions $\overline{\mathbf{1}_{E}}$ of sharp predicates, which are sharp themselves, see Exercise 4.4.3 The action $\varphi \bullet p$ of a factor $p$ on a multiset $\varphi$ from Exercise 4.2 .15 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 6.6.3. Let $v \in \mathcal{N}(X)$ be an urn and $E \subseteq X$ an event / subset, forming a sharp indicator predicate $\mathbf{1}_{E}: X \rightarrow[0,1]$.
l If $\left\|v \bullet \mathbf{1}_{E}\right\| \geq K$,

$$
\left.\operatorname{hg}[K](v)\right|_{\overrightarrow{\mathbf{1}_{E}}}=\operatorname{hg}[K]\left(v \bullet \mathbf{1}_{E}\right) .
$$

2 Similarly, if $v \bullet \mathbf{1}_{E}$ is non-empty,

$$
\left.p l[K](v)\right|_{\overline{1}_{E}} \cdot=p l[K]\left(v \bullet \mathbf{1}_{E}\right) .
$$

Proof. 1 We first recall from Exercise 4.4.3 that $\overline{\mathbf{1}_{E}}{ }^{\bullet}(\varphi)=1$ iff $\operatorname{supp}(\varphi) \subseteq E$, for $\varphi \in \mathcal{N}(X)$. Next, let's write $L=\|v\|$ for the size of the urn and $L_{E}=\| v \bullet$ $\mathbf{1}_{E} \|$ for the size of the restricted urn - where it is assumed that $L_{E} \geq K$.

Then:

$$
\begin{aligned}
& \operatorname{hg}[K](v) \vDash \overline{\mathbf{1}_{E}} \cdot=\sum_{\varphi \leq K^{v}} h g[K](v)(\varphi) \cdot \overline{\mathbf{1}_{E}}{ }^{\bullet}(\varphi) \\
& =\sum_{\varphi \leq{ }_{K} v, \text { supp }(\varphi) \subseteq E} h g[K](\nu)(\varphi) \\
& =\sum_{\varphi \leq K v \boldsymbol{\bullet}_{E}} h g[K](v)(\varphi) \\
& =\frac{\sum_{\varphi \leq_{K} v \bullet \boldsymbol{1}_{E}}\binom{v \bullet 1_{E}}{\varphi}}{\binom{L}{K}}=\frac{\binom{L_{E}}{K}}{\binom{L}{K}} \quad \text { by Lemma 1.8.2 }
\end{aligned}
$$

As a result:

$$
\begin{aligned}
\left.h g[K](v)\right|_{\overline{\mathbf{1}_{E}}} \cdot(\varphi) & =\sum_{\varphi \leq K^{v}} \frac{h g[K](v)(\varphi) \cdot \overline{\mathbf{1}_{E}} \cdot}{h g[K](v) \models \overline{\mathbf{1}_{E}}}|\varphi\rangle \\
& =\sum_{\varphi \leq_{K} v \bullet_{1}} \frac{\prod_{x}\binom{v(x)}{\varphi(x)}}{\binom{L}{K}} \cdot \frac{\binom{L}{K}}{\binom{L_{E}}{K}}|\varphi\rangle \\
& =\sum_{\varphi \leq_{K} v \bullet_{E}} \frac{\binom{v \bullet \mathbf{1}_{E}}{\varphi}}{\binom{L_{E}}{K}}|\varphi\rangle=\operatorname{hg}[K]\left(v \bullet \mathbf{1}_{E}\right) .
\end{aligned}
$$

2 Similarly.

### 6.6.1 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 $\left.\omega\right|_{\mathbf{1}_{E}}$, for a subset $E \subseteq X$, and on the right to $\left.\omega\right|_{p}$, for a fuzzy predicate $p: X \rightarrow[0,1]$.

```
x}\leftarrow
if x in E:
    return x
```

```
x}\leftarrow
y\leftarrowflip (p(y))
if }y==1
    return x
```

The mechanism on the left is rejection sampling, where all the samples from $\omega$ 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 $\operatorname{flip}(r)=$ $r|1\rangle+(1-r)|0\rangle$ 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 $X \rightarrow \mathbf{2}$, where $\mathbf{2}=\{0,1\}$. We use the tuple of functions $\left\langle i d, \mathbf{1}_{E}\right\rangle: X \rightarrow X \times \mathbf{2}$ to transform
the multinomial distribution $m n[K](\omega) \in \mathcal{D}(\mathcal{N}[K](X))$ to a distribution in $\mathcal{D}\left(\mathcal{N}[K](X \times \mathbf{2})\right.$ ). The second projection $\pi_{2}: X \times \mathbf{2} \rightarrow \mathbf{2}$ can be used as a sharp predicate. Its multiplicative extension $\overline{\pi_{2}}{ }^{\bullet}: \mathcal{N}[K](X \times \mathbf{2}) \rightarrow \mathbf{2}$ 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 6.22, we use:

$$
\begin{array}{ll}
\mathcal{D N}\left(\pi_{1}\right)\left(\left.\mathcal{D N}\left(\left\langle i d, \mathbf{1}_{E}\right\rangle\right)(\operatorname{mn}[K](\omega))\right|_{\bar{\pi}_{2}} \cdot\right) & \\
=\mathcal{D N}\left(\pi_{1}\right)\left(\left.\operatorname{mn}[K]\left(\mathcal{D}\left(\left\langle i d, \mathbf{1}_{E}\right\rangle\right)(\omega)\right)\right|_{\bar{\pi}_{2}} \cdot\right) & \\
=\mathcal{D N}\left(\pi_{1}\right)\left(\operatorname{mn}[K]\left(\left.\mathcal{D}\left(\left\langle i d, \mathbf{1}_{E}\right\rangle\right)(\omega)\right|_{\pi_{2}}\right)\right) & \\
=\mathcal{D N}\left(\pi_{1}\right)\left(\operatorname{mn}[K]\left(\mathcal{D}\left(\left\langle i d, \mathbf{1}_{E}\right\rangle\right)\left(\omega| |_{\pi_{2} \circ\left\langle i d, \mathbf{1}_{E}\right\rangle}\right\rangle\right)\right) & \\
\text { by Lemeoremality of multinomial } \\
=m n[K]\left(\mathcal{D}\left(\pi_{1}\right)\left(\mathcal{D}\left(\left\langle i d, 1, \mathbf{1}_{E}\right\rangle\right)\left(\left.\omega\right|_{\mathbf{1}_{E}}\right)\right)\right) & \\
=\operatorname{mn}[K]\left(\left.\omega\right|_{\mathbf{1}_{E}}\right) . &
\end{array}
$$

Hence, the program produces samples from the updated distribution $\left.\omega\right|_{1_{E}}$, as intended.
We turn to the update on the right in 6.22, with a fuzzy predicate $p: X \rightarrow$ $[0,1]$. We identify this function $p$ with the composite flip $\circ p$, giving a channel $p: X \rightsquigarrow \mathbf{2}$, see Exercise 4.3.6, as used above. We can form a tuple $\langle i d, p\rangle: X \rightsquigarrow$ $X \times \mathbf{2}$, 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 3.7.8. We thus take as semantics of the above program on right in 6.22):

```
\(\mathcal{D N}\left(\pi_{1}\right)\left(\left.(\mathcal{N}[K](\langle i d, p\rangle) \geqslant=m n[K](\omega))\right|_{\pi_{2}} \cdot\right)\)
    \(=\mathcal{D N}\left(\pi_{1}\right)\left(\left.m n[K](\langle i d, p\rangle \gg \omega)\right|_{\pi_{2}} \cdot\right)\) by channel naturality (3.43)
    \(=\mathcal{D N}\left(\pi_{1}\right)\left(\operatorname{mn}[K]\left(\left.(\langle i d, p\rangle \gg \omega)\right|_{\pi_{2}}\right)\right)\) by Theorem 6.6.1 (1)
    \(\stackrel{(*)}{=} \mathcal{D N}\left(\pi_{1}\right)\left(\operatorname{mn}[K]\left(\left.\sum_{x} \omega\right|_{p}(x)|x, 1\rangle\right)\right)\)
    \(=m n[K]\left(\mathcal{D}\left(\pi_{1}\right)\left(\left.\sum_{x} \omega\right|_{p}(x)|x, 1\rangle\right)\right) \quad\) by naturality of multinomial
    \(=m n[K]\left(\left.\omega\right|_{p}\right)\).
```

For the marked equation $\stackrel{\left({ }^{*}\right)}{=}$ we first note that the channel $\langle i d, p\rangle: X \leadsto X \times \mathbf{2}$ is of the form:

$$
\langle i d, p\rangle(x)=\sum_{x \in X} p(x)|x, 1\rangle+(1-p(x))|x, 0\rangle .
$$

Hence:

$$
\begin{aligned}
\langle i d, p\rangle \gg \omega \models \pi_{2} & =\sum_{x \in X} \omega(x) \cdot p(x) \cdot \pi_{2}(x, 1)+\omega(x) \cdot\left(1-p(x) \cdot \pi_{2}(x, 0)\right. \\
& =\sum_{x \in X} \omega(x) \cdot p(x)=\omega \models p .
\end{aligned}
$$

And thus:

$$
\left.(\langle i d, p\rangle \gg \omega)\right|_{\pi_{2}}=\sum_{x \in X} \frac{\omega(x) \cdot p(x)}{\omega \vDash p}|x, 1\rangle=\left.\sum_{x \in X} \omega\right|_{p}(x)|x, 1\rangle .
$$

An alternative to sample from $\left.\omega\right|_{p}$ is the Metropolis-Hastings method, see Subsection 6.5.3.

## Exercises

6.6.1 Prove Theorem 6.6.3 (2) yourself.
6.6.2 The question below is adapted from [159, §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 $\varphi$, that $\left(\varphi \cdot \mathbf{1}_{E}\right)(x)=\varphi(x) \cdot \mathbf{1}_{E}(x)$. Define sharp predicate $E_{L}: \mathcal{N}[K](X) \rightarrow[0,1]$ by:

$$
E_{L}(\varphi):= \begin{cases}1 & \text { if } \sum_{x \notin E} \varphi(x)=L \\ 0 & \text { otherwise } .\end{cases}
$$

Thus, $E_{L}(\varphi)=1 \mathrm{iff}\left\|\varphi \bullet \mathbf{1}_{\neg E}\right\|=L$.
Prove that:

$$
\mathcal{D}\left(-\bullet \mathbf{1}_{E}\right)\left(\left.m n[K](\omega)\right|_{E_{L}}\right)=m n[K-L]\left(\left.\omega\right|_{\mathbf{1}_{E}}\right) .
$$

6.6.3 Fix a multiset $\psi \in \mathcal{N}[K+1](X)$ of size $K+1$ and define the sharp predicate mcons $_{\psi}: X \times \mathcal{N}[K](X) \rightarrow[0,1]$ as:

$$
\operatorname{mcons}_{\psi}(x, \varphi):= \begin{cases}1 & \text { if } 1|x\rangle+\varphi=\psi \\ 0 & \text { otherwise }\end{cases}
$$

Show that for any distribution $\omega \in \mathcal{D}(X)$,

$$
\left.\omega \otimes m n[K](\omega)\right|_{m c o n s s_{\psi}}=\operatorname{DSD}(\psi),
$$

where the draw-store-delete channel $D S D$ is from Exercise 3.2.12. Later on we shall see that this equation follows from a general result, see Example 6.3.3 (2).
6.6.4 We return to the multiplicative extension $\bar{p}^{\bullet}: \mathcal{L}(X) \rightarrow \mathbb{R}_{\geq 0}$ to lists, of a factor $p: X \rightarrow \mathbb{R}_{\geq 0}$, from Exercise 4.4.7

1 Show that:

$$
\left.\operatorname{iid}[K](\omega)\right|_{\bar{p}^{*}}=\operatorname{iid}[K]\left(\left.\omega\right|_{p}\right)
$$

2 Prove next for Poisson-iid that:

$$
\left.\operatorname{Piid}[\lambda](\omega)\right|_{\vec{p}^{*}}=\operatorname{Piid}[\lambda \cdot(\omega \vDash p)](\omega)
$$

### 6.7 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 $\left.\omega\right|_{p}$ making $p$ more true, see Theorem 6.1.5. 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 of heads and tails. What can we then infer about $r$ ? By following the above approach we seek a distribution $\omega$ 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 6.7.1. 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:

$$
\begin{aligned}
{[a, b]_{N} } & :=\left\{\left.a+\left(i+\frac{1}{2}\right) s \right\rvert\, 0 \leq i<N\right\} \quad \text { where } \quad s:=\frac{b-a}{N} \\
& =\left\{a+\frac{1}{2} s, a+\frac{3}{2} s, \ldots, a+\frac{2 N-1}{2} s\right\} \\
& =\left\{a+\frac{1}{2} s, a+\frac{3}{2} s, \ldots, b-\frac{1}{2} s\right\} .
\end{aligned}
$$

2 Let $f: S \rightarrow \mathbb{R}_{\geq 0}$ be a function, defined on a finite subset $S \subseteq \mathbb{R}$. We write $\operatorname{Disc}(f, S) \in \mathcal{D}(S)$ for the discrete distribution defined as:

$$
\operatorname{Disc}(f, S):=\sum_{x \in S} \frac{f(x)}{t}|x\rangle \quad \text { where } \quad t:=\sum_{x \in S} f(x)
$$

Often we combine the notations from these two items and use discretised states of the form $\operatorname{Disc}\left(f,[a, b]_{N}\right)$.

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:

$$
[1,2]_{3}=\left\{1+\frac{1}{2} \cdot \frac{1}{3}, 1+\frac{3}{2} \cdot \frac{1}{3}, 1+\frac{5}{2} \cdot \frac{1}{3}\right\}=\left\{1+\frac{1}{6}, 1+\frac{1}{2}, 1+\frac{5}{6}\right\} .
$$

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 $f_{[a, b]_{N}}$, like before, or alternatively as $\operatorname{Disc}\left(\mathbf{1},[a, b]_{N}\right)$, where $\mathbf{1}$ is the constant-one function.

Example 6.7.2. 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 (2.14). This gives:

$$
[T, H, H, H, T, T, H, H] \longmapsto 5|H\rangle+3|T\rangle \longmapsto \frac{5}{8}|H\rangle+\frac{3}{8}|T\rangle
$$

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 $\operatorname{flip}(b)=b|H\rangle+(1-b)|T\rangle$. Thus we can see flip as a
channel flip: $[0,1] \rightsquigarrow\{H, T\}$. 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:

$$
\begin{aligned}
\text { unif }:=\operatorname{Disc}\left(\mathbf{1},[0,1]_{N}\right) & =\sum_{x \in[0,1]_{N}} \frac{1}{N}|x\rangle \\
& =\sum_{0 \leq i<N} \frac{1}{N}\left|\frac{2 i+1}{2 N}\right\rangle \\
& =\frac{1}{N}\left|\frac{1}{2 N}\right\rangle+\frac{1}{N}\left|\frac{3}{2 N}\right\rangle+\cdots+\frac{1}{N}\left|\frac{2 N-1}{2 N}\right\rangle .
\end{aligned}
$$

We use the flip operation as a channel, restricted to the discretised space:

$$
[0,1]_{N} \xrightarrow{\text { fip }}\{H, T\} \quad \text { given by } \quad \text { flip }(b)=b|H\rangle+(1-b)|T\rangle .
$$

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 6.7.2 below, that:

$$
\text { flip } \gg \text { unif } \vDash \mathbf{1}_{H}=\text { flip } \gg \text { unif } \models \mathbf{1}_{T}=\frac{1}{2} .
$$

Predicate transformation along flip yields two predicates on $[0,1]_{N}$ given by:

$$
\left(\text { flip }=\ll \mathbf{1}_{H}\right)(r)=r \quad \text { and } \quad\left(\text { flip }=\ll \mathbf{1}_{T}\right)(r)=1-r .
$$

Given the above sequence of head/tail observations [ $T, H, H, H, T, T, H, H$ ], we perform successive predicate transformations flip $=\ll \mathbf{1}_{(-)}$and update the prior (uniform) state accordingly. This gives, via Lemma 6.1.6(3),

```
\(\left.u n i f\right|_{\text {fip }=\ll 1_{T}}\)
```



```
\(\left.\left.\left.u n i f\right|_{f i p=\ll \mathbf{1}_{T}}\right|_{f i p=\ll \mathbf{1}_{H}}\right|_{f i p=<\mathbf{1}_{H}}=\operatorname{unif}_{\left(f f i p=<\mathbf{1}_{T}\right) \&\left(f l i p=<\mathbf{1}_{H}\right) \&\left(f l i p=<\mathbf{1}_{H}\right)}\)
    \(=\left.u^{\prime}\right|_{\left(f i p=\ll \mathbf{1}_{H}\right)^{2} \&\left(f i p=<\mathbf{1}_{T}\right)}\)
    \(\vdots\)
unif \(\left.\right|_{\left(f i p=<\mathbf{1}_{H}\right)^{5} \&\left(f f i p \equiv<\mathbf{1}_{T}\right)^{3}}\)
```

As we already know from Lemma 6.1.6(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 $\left(\text { flip }=\ll \mathbf{1}_{H}\right)^{5} \&\left(\text { flip }=\ll \mathbf{1}_{T}\right)^{3}$ in the last line above.

An overview of the distributions arising from these succesive updates is given in Figure 6.11. 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 6.11

After these eight updates, let's write $\rho:=\operatorname{unif}_{\left(\text {fip } \equiv \ll 1_{H}\right)^{5} \&\left(f l i p \equiv 1_{T}\right)^{3}}$ for the resulting distribution. We now ask three questions.

1 Where does $\rho$ reach its highest value, and what is it? The answers are given by the singleton set of maximum values:

$$
\operatorname{argmax}(\rho)=\left\{\frac{5}{8}\right\} \quad \text { with } \quad \rho\left(\frac{5}{8}\right) \approx 0.025347
$$

2 What is the predicted coin distribution? The outcome, with truncated multiplicities, is:

$$
\text { flip } \gg \rho=0.6|H\rangle+0.4|T\rangle \text {. }
$$

3 What is the expected value of $\rho$ ? It is:

$$
\operatorname{mean}(\rho)=0.6=\rho \vDash \text { flip }=\ll \mathbf{1}_{H}
$$

For mathematical reason $\square^{2}$ the exact outcome is 0.6 . However, we have used approximation via discretisation. The value computed with this discretisation is 0.599999985316273 . 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 6.11 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 Section ??. 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 $\operatorname{Beta}_{N}: \mathbb{N}_{>0} \times \mathbb{N}_{>0} \rightarrow[0,1]_{N}$ as normalisation:

$$
\begin{align*}
\operatorname{Beta}_{N}(a, b) & :=\operatorname{Flrn}\left(\sum_{r \in[0,1]_{N}} r^{a-1} \cdot(1-r)^{b-1}|r\rangle\right) \\
& =\sum_{r \in[0,1]_{N}} \frac{r_{N}{ }^{a-1} \cdot(1-r)^{b-1}}{\sum_{s \in[0,1]_{N}} s^{a-1} \cdot(1-s)^{b-1}}|r\rangle . \tag{6.23}
\end{align*}
$$

Then we have the following results.

[^9]

Figure 6.11 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 corresponding continuous Beta-distribution. It shows how closely the discrete approximations match the continuous probability density function.

Proposition 6.7.3. Let $N \in \mathbb{N}$ be the the discretisation parameter, used in the chopped subspace $[0,1]_{N} \subseteq[0,1]$, with uniform distribution unif $\in \mathcal{D}\left([0,1]_{N}\right)$, and let $a, b \in \mathbb{N}_{>0}$.

1 The discretised Beta distribution $\operatorname{Beta}_{N}(a, b)$ satisfies:

$$
\operatorname{Beta}_{N}(a, b)=\text { unif }\left.\right|_{\left(f i p=<\mathbf{1}_{H}\right)^{a-1} \&\left(f i p \approx<\mathbf{1}_{T}\right)^{b-1}} .
$$

2 The class of discretised Beta distributions is closed under backward inference along flip:

$$
\begin{aligned}
\left.\operatorname{Beta}_{N}(a, b)\right|_{\text {flip } \ll \mathbf{1}_{H}} & =\operatorname{Beta}_{N}(a+1, b) \\
\left.\operatorname{Beta}_{N}(a, b)\right|_{f l i p=<\mathbf{1}_{T}} & =\operatorname{Beta}_{N}(a, b+1) .
\end{aligned}
$$

More generally, for $n, m \in \mathbb{N}$ one has:

$$
\left.\operatorname{Beta}_{N}(a, b)\right|_{\left(f i p \approx<\mathbf{1}_{H}\right)^{n} \&\left(f i p \approx<\mathbf{1}_{T}\right)^{m}}=\operatorname{Beta}_{N}(a+n, b+m) .
$$

This second item shows how backward inference of a prior $\operatorname{Beta}_{N}(a, b)$ along a flip can be expressed via an update of the 'hyperparameters' $a, b$.

Proof. 1 For $r \in[0,1]_{N}$ we have:

$$
\begin{aligned}
\text { unif }\left.\right|_{\left(\text {flip }<\mathbf{1}_{H}\right)^{a-1} \&\left(f i p=<\mathbf{1}_{T}\right)^{b-1}}(r) & =\frac{\text { unif }(r) \cdot\left(\text { flip }=\ll \mathbf{1}_{H}\right)^{a-1}(r) \cdot\left(\text { flip }=\ll \mathbf{1}_{T}\right)^{b-1}(r)}{\text { unif } \models\left(\text { flip }=\ll \mathbf{1}_{H}\right)^{a-1} \&\left(\text { flip }=\ll \mathbf{1}_{T}\right)^{b-1}} \\
& =\frac{1 /_{N} \cdot r^{a-1} \cdot(1-r)^{b-1}}{\sum_{s \in[0,1]_{N}} 1 / N \cdot s^{a-1} \cdot(1-s)^{b-1}} \\
& =\operatorname{Beta}_{N}(a, b) .
\end{aligned}
$$

2 We use this result and Lemma 6.1.6 (3) to prove the general statement:

$$
\begin{aligned}
& \left.\operatorname{Beta}_{N}(a, b)\right|_{\left(f i p=\ll \mathbf{1}_{H}\right)^{n}} \&\left(\text { ffip }=<\mathbf{1}_{T}\right)^{n} \\
& =\text { unif }\left.\left.\right|_{\left(f l i p=\mathbf{1}_{H}\right)^{a-1} \&\left(f i p=\ll \mathbf{1}_{T}\right)^{b-1}}\right|_{\left(f i p=\ll \mathbf{1}_{H}\right)^{n} \&\left(f i p=\ll \mathbf{1}_{T}\right)^{m}} \\
& =\text { unif }\left.\right|_{\left(f i p=<\mathbf{1}_{H}\right)^{a-1} \&\left(f f i p=<\mathbf{1}_{T}\right)^{b-1} \&\left(f i l p=<\mathbf{1}_{H}\right)^{n} \&\left(f f i p=<\mathbf{1}_{T}\right)^{m}} \\
& =\text { unif }\left.\right|_{\left(f l i p=<\mathbf{1}_{H}\right)^{a+n-1} \&\left(f l i p=<\mathbf{1}_{T}\right)^{b+m-1}} \\
& =\operatorname{Beta}_{N}(a+n, b+m) \text {. }
\end{aligned}
$$

### 6.7.1 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 that we write $\mathcal{N}[K](X)$ for the set of natural multisets - with natural numbers as multiplicities - of size $K$. We recall from Theorem4.5.9 (1) that for $K>0$ we collect 'fractional' distributions in a subset:

$$
\mathcal{D}[K](X):=\{\operatorname{Flrn}(\varphi) \mid \varphi \in \mathcal{N}[K](X)\}=\left\{\left.\frac{1}{K} \cdot \varphi \right\rvert\, \varphi \in \mathcal{N}[K](X)\right\} .
$$

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 the multisets:

$$
5|a\rangle, 4|a\rangle+1|b\rangle, 3|a\rangle+2|b\rangle, 2|a\rangle+3|b\rangle, 1|a\rangle+4|b\rangle, 5|b\rangle
$$

Hence, $\mathcal{D}[5](\{a, b\})$ contains the distributions:

$$
1|a\rangle, \frac{4}{5}|a\rangle+\frac{1}{5}|b\rangle, \frac{3}{5}|a\rangle+\frac{2}{5}|b\rangle, \frac{2}{5}|a\rangle+\frac{3}{5}|b\rangle, \frac{1}{5}|a\rangle+\frac{4}{5}|b\rangle, 1|b\rangle .
$$

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 4.5.9 (1). Figure 4.4 gives a visual representation.

Example 6.7.4. Consider an election ${ }^{3}$ 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 $v_{K}$ is the uniform distribution over these preference distributions:

$$
v_{K}:=\operatorname{unif}_{\mathcal{D}[K](X)}=\sum_{\varphi \in \mathcal{N}[K](X)} \frac{1}{\left(\binom{3}{K}\right)}|\operatorname{Flrn}(\varphi)\rangle \in \mathcal{D}(\mathcal{D}[K](X)) .
$$

We use the following sharp predicates on $\mathcal{D}[K](X)$, called: aw for $a$ wins, $b w$ for $b$ wins, $c w$ for $c$ wins, and $n w$ for no-one wins. They are defined on a fractional distribution $\sigma \in \mathcal{D}[K](X)$ as:

$$
\begin{aligned}
& \operatorname{aw}(\sigma)=1 \Longleftrightarrow \sigma(a)>\frac{1}{2} \\
& \operatorname{bw}(\sigma)=1 \Longleftrightarrow \sigma(b)>\frac{1}{2} \\
& c w(\sigma)=1 \Longleftrightarrow \sigma(c)>\frac{1}{2} \\
& n w(\sigma)=1 \Longleftrightarrow \sigma(a) \leq \frac{1}{2} \text { and } \sigma(b) \leq \frac{1}{2} \text { and } \sigma(c) \leq \frac{1}{2} .
\end{aligned}
$$

The prior validities:

$$
v_{K} \vDash a w \quad v_{K} \vDash b w \quad v_{K} \vDash c w \quad v_{K} \vDash n w
$$

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 $i: \mathcal{D}[K](X) \hookrightarrow X$. Like in the above coin scenario, we perform successive backward inferences, using
${ }^{3}$ This illustration is copied and adapted from Sam Staton's tutorial at the Mathematical Foundations of Programming Semantics conference, 2021.
the above poll numbers, to obtain a posterior $\rho_{K}$ obtained via updates:

$$
\rho_{K}:=\left.\left.\left.v_{K}\right|_{\left(i \lll \mathbf{1}_{a}\right)^{52}}\right|_{\left(i \lll \mathbf{1}_{b}\right)^{28}}\right|_{\left(i \ll \mathbf{1}_{c}\right)^{20}}=\left.v_{K}\right|_{\left(i \lll \mathbf{1}_{a}\right)^{52} \&\left(i \approx<\mathbf{1}_{b}\right)^{28} \&\left(i \approx<\mathbf{1}_{c}\right)^{20} .} .
$$

The posterior validity $\rho_{K} \vDash$ aw takes the following values, for several values of the discretisation parameter $K$.

|  | $K=100$ | $K=500$ | $K=1000$ |
| :---: | :---: | :---: | :---: |
| 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

6.7.1 Recall the $N$-element set $[a, b]_{N}$ from Definition 6.7.1 (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?
6.7.2 Consider coin parameter learning in Example 6.7.2.

1 Show that the predicition in the prior (uniform) state unif on [0,1] ${ }_{N}$ gives a fair coin, i.e.

$$
\text { flip } \gg=\text { unif }=\frac{1}{2}|1\rangle+\frac{1}{2}|0\rangle \text {. }
$$

This equality is independent of $N$.
2 Prove that, also independently of $N$,

$$
\text { unif } \vDash \text { flip }=\ll \mathbf{1}_{H}=\frac{1}{2} \text {. }
$$

3 Show next that:

$$
\left.u n i f\right|_{f i p=\ll \mathbf{1}_{H}}=\sum_{x \in[0,1]_{N}} \frac{2 x}{N}|x\rangle .
$$

4 Use Proposition 1.2.6 (2) to prove that:

$$
\left(\text { flip } \gg=\left(\text { unif }\left.\right|_{f i p=\ll \mathbf{1}_{H}}\right)\right)(H)=\frac{2}{3}-\frac{1}{6 N^{2}} .
$$

Conclude that flip $\gg\left(\right.$ unif $\left._{\text {flip }^{=}\left\langle\mathbf{1}_{H}\right.}\right)$ approaches $\frac{2}{3}|H\rangle+\frac{1}{3}|T\rangle$ as $N$ goes to infinity.
6.7.3 Check that the probabilities $($ flip $\gg \rho)(H)$ and mean $(\rho)$ are the same, in Example 6.7.2.
6.7.4 Check that $\operatorname{Beta}_{N}(1,1)$ in 6.23 is the uniform distribution unif on $[0,1]_{N}$.
6.7.5 Consider the discretised beta channel $\operatorname{Beta}_{N}: \mathbb{N} \times \mathbb{N} \rightarrow \mathcal{D}(X)$ in the following diagram:

where the maps add and infer are given by:

$$
\begin{aligned}
\operatorname{add}(a, b, n|H\rangle+m|T\rangle) & =(a+n, b+m) \\
\operatorname{infer}(\omega, n|H\rangle+m|T\rangle) & =\left.\omega\right|_{\left(\text {ffip }=\ll \mathbf{1}_{H}\right)^{n} \&\left(f l i p=\ll \mathbf{1}_{T}\right)^{m} .} .
\end{aligned}
$$

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 Also infer: $\mathcal{D}\left([0,1]_{N}\right) \times \mathcal{N}(\{H, T\}) \rightarrow \mathcal{D}\left([0,1]_{N}\right)$ is a monoid action.
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 [87].)
6.7.6 We accept, without proof, the following equation. For $a, b \in \mathbb{N}$,

$$
\begin{equation*}
\lim _{N \rightarrow \infty} \frac{1}{N} \cdot \sum_{r \in[0,1]_{N}} r^{a-1} \cdot(1-r)^{b-1}=\frac{(a-1)!\cdot(b-1)!}{(a+b-1)!} . \tag{*}
\end{equation*}
$$

Use equation $(*)$ to prove that the binary Pólya distribution can be obtained from the binomial, via the limit of the discretised Beta distribution (6.23):

$$
\begin{aligned}
& \lim _{N \rightarrow \infty}\left(b n[K] \gg \operatorname{Beta}_{N}(a, b)\right)(i) \\
& =\operatorname{pl}[K](a|H\rangle+b|T\rangle)(i|H\rangle+(K-i)|T\rangle) .
\end{aligned}
$$

Later, in Exercise ?? we shall see a proper continuous version of this result.
6.7.7 Redo the coin bias learning from the beginning of this section in the style of Example 6.7.4, via discretisation $\mathcal{D}[K](\mathbf{2})$ of states in $\mathcal{D}(\mathbf{2}) \cong$ [ 0,1 ].

### 6.8 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 $\operatorname{Pred}(X)=$ $[0,1]^{X}$ 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 $p \boxtimes q$, and scalar multiplication $r \cdot p$, see Subsection 4.2.3.

The next result is a Riesz-style representation result, representing distributions on $X$ via a double dual $[0,1]^{[0,1]^{X}}$.

## Theorem 6.8.1.

1 Let X be a finite set. There is then a 'representation' isomorphism,

$$
\begin{equation*}
\mathcal{D}(X) \xrightarrow[\cong]{\underline{\imath}} \operatorname{EMod}(\operatorname{Pred}(X),[0,1]) \tag{6.24}
\end{equation*}
$$

given by validity: $\mathcal{V}(\omega)(p):=\omega \vDash p$. It gives a bijective correspondence between distributions in $\mathcal{D}(X)$ and 'predicate evaluations' Pred $(X) \rightarrow[0,1]$ that preserve the effect module structure.
2 This isomorphism 6.24) is natural in $X$, with respect to channels $c: X \leadsto Y$; they give a commuting diagram of the following form:


This result uses the notation $\operatorname{EMod}(\operatorname{Pred}(X),[0,1])$ for the 'hom' set of homomorphisms of effect modules $\operatorname{Pred}(X) \rightarrow[0,1]$. The mapping $X \mapsto$ $\operatorname{EMod}(\operatorname{Pred}(X),[0,1])$ forms a monad on Sets, known as the expectation monad, see [89].

Proof. 1 We have already seen that $\mathcal{V}(\omega): \operatorname{Pred}(X) \rightarrow[0,1]$ preserves the
effect module structure, see Lemma 4.2.6 In order to show that it is an isomorphism, let $h: \operatorname{Pred}(X) \rightarrow[0,1]$ be a homomorphism of effect modules. It gives rise to a distribution, via point predicates $\mathbf{1}_{x} \in \operatorname{Pred}(X)$, for $x \in X$.

$$
\mathcal{V}^{-1}(h):=\sum_{x \in X} h\left(\mathbf{1}_{x}\right)|x\rangle \in \mathcal{D}(X)
$$

The probabilities $h\left(\mathbf{1}_{x}\right)$ add up to one since $h$ is a map of effect modules:

$$
\sum_{x \in X} h\left(\mathbf{1}_{x}\right)=h\left(\emptyset_{x \in X} \mathbf{1}_{x}\right)=h(\mathbf{1})=1
$$

These $\mathcal{V}$ and $\mathcal{V}^{-1}$ are each other's inverses:

$$
\begin{aligned}
\left(\mathcal{V}^{-1} \circ \mathcal{V}\right)(\omega)=\sum_{x \in X} \mathcal{V}(\omega)\left(\mathbf{1}_{x}\right)|x\rangle & =\sum_{x \in X}\left(\omega \vDash \mathbf{1}_{x}\right)|x\rangle \\
& =\sum_{x \in X} \omega(x)|x\rangle=\omega \\
\left(\mathcal{V} \circ \mathcal{V}^{-1}\right)(h)(p)=\left(\sum_{x \in X} h\left(\mathbf{1}_{x}\right)|x\rangle\right) \vDash p & =\sum_{x \in X} h\left(\mathbf{1}_{x}\right) \cdot p(x) \\
& =h\left(\emptyset_{x \in X} p(x) \cdot \mathbf{1}_{x}\right)=h(p)
\end{aligned}
$$

In the last line we use that $h$ is a homomorphism of effect modules and that the predicate $p$ has a normal form $\mathbb{Q}_{x} p(x) \cdot \mathbf{1}_{x}$, see Lemma 4.2.3 (2).
2 Naturality for the representation isomorphism $\mathcal{V}$ in 6.24 follows from Proposition4.3.3 for a channel $c: X \mapsto Y$,

$$
\begin{aligned}
((h \mapsto h(c=\ll-)) \circ \mathcal{V})(\omega)(q) & =\mathcal{V}(\omega)(c=\ll q) \\
& =\omega \vDash c=\ll q \\
& =c \gg \omega \vDash q \\
& =\mathcal{V}(c \gg=\omega)(q) \\
& =(\mathcal{V} \circ(c \gg=-))(\omega)(q) .
\end{aligned}
$$

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 6.24. The right-hand-side fits the Bayesian view, which focuses on assigning probabilities to belief functions (predicates), see [42], in this situation via special functions $\operatorname{Pred}(X) \rightarrow[0,1]$ 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 6.8.6, 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 (6.24). 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 6.8.2. Fix finite sets $X, Y$.

1 Let $h: \operatorname{Pred}(X \times Y) \rightarrow[0,1]$ be a map of effect modules. Define first and second marginals $h[1,0]: \operatorname{Pred}(X) \rightarrow[0,1]$ and $h[0,1]: \operatorname{Pred}(Y) \rightarrow[0,1]$ of $h$ as:

$$
h[1,0](p):=h(p \otimes \mathbf{1}) \quad \text { and } \quad h[0,1](q):=h(\mathbf{1} \otimes q) .
$$

Then: $\mathcal{V}^{-1}(h[1,0])=\mathcal{V}^{-1}(h)[1,0]$ and $\mathcal{V}^{-1}(h[0,1])=\mathcal{V}^{-1}(h)[0,1]$.
2 Let $h: \operatorname{Pred}(X) \rightarrow[0,1]$ and $k: \operatorname{Pred}(Y) \rightarrow[0,1]$ be maps of effect modules. We define their tensor product $h \otimes k: \operatorname{Pred}(X \times Y) \rightarrow[0,1]$ as:

$$
\begin{aligned}
(h \otimes k)(r) & :=h(x \mapsto k(r(x,-))) \\
& =k(y \mapsto h(r(-, y))) .
\end{aligned}
$$

Then: $\mathcal{V}^{-1}(h \otimes k)=\mathcal{V}^{-1}(h) \otimes \mathcal{V}^{-1}(k)$.
3 Let $h: \operatorname{Pred}(X) \rightarrow[0,1]$ be a map of effect modules and let $p \in \operatorname{Pred}(X)$ be a predicate with $h(p) \neq 0$. We define an update $\left.h\right|_{p}: \operatorname{Pred}(X) \rightarrow[0,1]$ as:

$$
\left.h\right|_{p}(q):=\frac{h(p \& q)}{h(p)}
$$

Then: $\mathcal{V}^{-1}\left(\left.h\right|_{p}\right)=\left.\mathcal{V}^{-1}(h)\right|_{p}$.
Proof. 1 We consider the first marginal only. It is easy to see that the map $h[1,0]: \operatorname{Pred}(X) \rightarrow[0,1]$, as defined above, preserves the effect module
structure. We get an equality of distributions since for $x \in X$,

$$
\begin{aligned}
\mathcal{V}^{-1}(h[1,0])(x)=h[1,0]\left(\mathbf{1}_{x}\right) & =h\left(\mathbf{1}_{x} \otimes \mathbf{1}\right) \\
& =h\left(\mathbf{1}_{x} \otimes \bigotimes_{y \in Y} \mathbf{1}_{y}\right) \\
& =\sum_{y \in Y} h\left(\mathbf{1}_{x} \otimes \mathbf{1}_{y}\right) \\
& =\sum_{y \in Y} h\left(\mathbf{1}_{(x, y)}\right) \\
& =\sum_{y \in Y} \mathcal{V}^{-1}(h)(x, y)=\mathcal{V}^{-1}(h)[1,0](x) .
\end{aligned}
$$

2 For elements $u \in X$ and $v \in Y$ we have:

$$
\begin{aligned}
\mathcal{V}^{-1}(h \otimes k)(u, v)=(h \otimes k)\left(\mathbf{1}_{(u, v)}\right) & =h\left(x \mapsto k\left(\mathbf{1}_{(u, v)}(x,-)\right)\right) \\
& =h\left(x \mapsto k\left(\mathbf{1}_{u}(x) \cdot \mathbf{1}_{v}\right)\right) \\
& =h\left(x \mapsto \mathbf{1}_{u}(x) \cdot k\left(\mathbf{1}_{v}\right)\right) \\
& =h\left(k\left(\mathbf{1}_{v}\right) \cdot \mathbf{1}_{u}\right) \\
& =k\left(\mathbf{1}_{v}\right) \cdot h\left(\mathbf{1}_{u}\right) \\
& =\mathcal{V}^{-1}(h)(u) \cdot \mathcal{V}^{-1}(k)(v) \\
& =\left(\mathcal{V}^{-1}(h) \otimes \mathcal{V}^{-1}(k)\right)(u, v) .
\end{aligned}
$$

The same can be shown for the other formulation of $h \otimes k$ in item (2).
3 Again, $\left.h\right|_{p}: \operatorname{Pred}(X) \rightarrow[0,1]$ is a map of effect modules. Next, for $x \in X$,

$$
\begin{aligned}
\mathcal{V}^{-1}\left(\left.h\right|_{p}\right)(x)=\left.h\right|_{p}\left(\mathbf{1}_{x}\right) & =\frac{h\left(p \otimes \mathbf{1}_{x}\right)}{h(p)} \\
& =\frac{\mathcal{V}^{-1}(h) \vDash p \otimes \mathbf{1}_{x}}{\mathcal{V}^{-1}(h) \vDash p} \\
& =\left.\mathcal{V}^{-1}(h)\right|_{p} \vDash \mathbf{1}_{x} \quad \text { by Bayes, see Theorem 6.1.3 2 } \\
& =\left.\mathcal{V}^{-1}(h)\right|_{p}(x) .
\end{aligned}
$$

The above results make it possible to do probability theory in a distributionfree style, with only predicates and predicate transformers (effect module maps). We illustrate how this works using an earlier example.

Example 6.8.3. We illustrate the purely Bayesian approach, on the right-handside of the isomorphism 6.24, 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 $\Omega:=\mathcal{V}(\omega): \operatorname{Pred}(M) \rightarrow$ $[0,1]$, where $M=\{0,1,2\}$ is the set of medicines. Thus:

$$
\Omega(p):=\mathcal{V}(\omega)(p)=\frac{3}{20} \cdot p(0)+\frac{9}{20} \cdot p(1)+\frac{2}{5} \cdot p(2) .
$$

It can be updated with the sharp predicate $\mathbf{1}_{E} \in \operatorname{Pred}(M)$, where $E=\{1,2\} \subseteq$ $M$ gives a restriction to medicines 1,2 . Via the update formula of Proposition 6.8.2 (3) we get a new predicate evaluation $\left.\Omega\right|_{1_{E}}: \operatorname{Pred}(M) \rightarrow[0,1]$, namely:

$$
\left.\Omega\right|_{\mathbf{1}_{E}}(p)=\frac{\Omega\left(p \& \mathbf{1}_{E}\right)}{\Omega\left(\mathbf{1}_{E}\right)}=\frac{9 / 20 \cdot p(1)+2 / 5 \cdot p(2)}{9 / 20+2 / 5}=\frac{9}{17} \cdot p(1)+\frac{8}{17} \cdot p(2) .
$$

It corresponds to the updated distribution $\left.\omega\right|_{\mathbf{1}_{E}}=\frac{9}{17}|1\rangle+\frac{8}{17}|2\rangle$ that we obtained in Example 6.2.6
Next we recall from Example 6.2 .6 the channel $b: M \leadsto B$, 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 4.3.10 that such a channel bijectively corresponds to a predicate transformer operation $\beta=b=<(-): \operatorname{Pred}(B) \rightarrow \operatorname{Pred}(M)$, preserving the effect module structure of predicates. This map $\beta$ puts channels 'on the Bayesian' side, in the world of predicates. Explicitly, for $q \in \operatorname{Pred}(B)$, it is defined as:

$$
\begin{aligned}
\beta(q)(0):=(b=\ll q)(0) & =b(0)(H) \cdot q(H)+b(0)(L) \cdot q(L) \\
& =\frac{2}{3} \cdot q(H)+\frac{1}{3} \cdot q(L) .
\end{aligned}
$$

Similarly, $\beta(q)(1):=\frac{7}{9} \cdot q(H)+\frac{2}{9} \cdot q(L)$ and $\beta(q)(2):=\frac{5}{8} \cdot q(H)+\frac{3}{8} \cdot q(L)$.
We are interested in learning the new predicate evaluation (on $M$ ) when we have evidence of a high blood pressure. We then update $\Omega$ with $\beta\left(\mathbf{1}_{H}\right)$, giving as new predicate evaluation:

$$
\begin{aligned}
\left.\Omega\right|_{\beta\left(\mathbf{1}_{H}\right)}(p) & =\frac{\Omega\left(p \& \beta\left(\mathbf{1}_{H}\right)\right)}{\Omega\left(\beta\left(\mathbf{1}_{H}\right)\right)} \\
& =\frac{3 / 20 \cdot \beta\left(\mathbf{1}_{H}\right)(0) \cdot p(0)+9 / 20 \cdot \beta\left(\mathbf{1}_{H}\right)(1) \cdot p(1)+2 / 5 \cdot \beta\left(\mathbf{1}_{H}\right)(2) \cdot p(2)}{3 / 20 \cdot \beta\left(\mathbf{1}_{H}\right)(0)+9 / 20 \cdot \beta\left(\mathbf{1}_{H}\right)(1)+2 / 5 \cdot \beta\left(\mathbf{1}_{H}\right)(2)} \\
& =\frac{3 / 20 \cdot 2 / 3 \cdot p(0)+9 / 20 \cdot 7 / 9 \cdot p(1)+2 / 5 \cdot 5 / 8 \cdot p(2)}{3 / 20 \cdot 2 / 3+9 / 20 \cdot 7 / 9+2 / 5 \cdot 5 / 8} \\
& =\frac{1}{7} \cdot p(0)+\frac{1}{2} \cdot p(1)+\frac{5}{14} \cdot p(2) .
\end{aligned}
$$

This predicate evaluation corresponds, via the isomorphism (6.24), to the updated distribution $\left.\omega\right|_{b=\ll 1_{H}}=\frac{1}{7}|0\rangle+\frac{1}{2}|1\rangle+\frac{5}{14}|2\rangle$ computed in Example 6.2.6.

We finish this section with an extension of the correspondence 6.24 between distributions in $\mathcal{D}(X)$ and predicate evaluations $\operatorname{Pred}(X) \rightarrow[0,1]$ 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 $\infty$ joins and $\infty$-continuity for what are commonly called $\omega$-joins and $\omega$-continuity. Since we often use the Greek letter $\omega$ for states / distributions, we choose to avoid using it also for the first infinite ordinal $\omega$ in these expressions. The use
of the sign $\infty$ 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 [153, 2], since they provide least fixed points of continuous functions, see Exercise 6.8.4 Instead of requiring joins of countable chains, one can also use joins of directed sets.

Definition 6.8.4. 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 $\infty$-cpo, or, in words, a countably complete partial order, if each ascending chain $\left(a_{n}\right)$ in $A$ has a least upperbound (join) $\bigvee_{n} a_{n} \in A$.
3 A monotone function $f: A \rightarrow B$ between two $\infty$-complete posets $A, B$ is called $\infty$-continuous if it preserves joins of ascending chains, that is, if $f\left(\bigvee_{n} a_{n}\right)=\bigvee_{n} f\left(a_{n}\right)$, for each ascending chain $\left(a_{n}\right)$ in $A$.
4 We shall write $\infty$-Cpo for the category of $\infty$-cpo's with $\infty$-continuous (and, implicitly, monotone) functions between them.

These countable joins can be combined with effect algebra / module structure, see Definition 4.2.1 Later on, in Section ?? we shall see how these effect algebras with joins capture the order-theoretic essentials of so-called $\sigma$ algebras of measurable subsets.

## Definition 6.8.5.

1 An $\infty$-effect algebra is an effect algebra which is $\infty$-complete as a poset, using the order described in Exercise 4.2.16. From the same exercise we know that the sum $\boxtimes$ automatically preserves joins in each argument.

We shall write $\infty$-EA for the category with $\infty$-effect algebras as objects, and with $\infty$-continuous effect algebra maps as morphisms.
2 Similarly, an $\infty$-effect module is an effect module $E$ which is $\infty$-complete as poset, and whose scalar multiplication $[0,1] \times E \rightarrow E$ is $\infty$-continuous in each argument.

The category $\infty$-EMod contains such $\infty$-effect modules, with $\infty$-continous effect module maps between them.

In the last point we implicitly use that the unit interval $[0,1]$ is an $\infty$-cpo. This is obvious.

For each set $X$, the powerset $\mathcal{P}(X)$ is an effect algebra with arbitrary joins, so in particular $\infty$-joins. The unit interval $[0,1]$ is an $\infty$-effect module, and more generally, each set of predicates $\operatorname{Pred}(X)=[0,1]^{X}$ is an $\infty$-effect module, via
pointwise joins. Later on, in continuous probability, we shall see measurable spaces whose sets of measurable subsets form examples of $\infty$-effect algebras.

We recall from Exercise 4.2.17 that the sum operation $\otimes$ of an $\infty$-effect algebra is continuous in both arguments separately. Explicitly: $\bigvee_{n} x \otimes y_{n}=$ $x \otimes \bigvee_{n} y_{n}$. Further, if there are joins $\bigvee$, there are also meets $\Lambda$ via orthosupplement.

Theorem 6.8.6. For each countable set $X$ there is a representation isomorphism,

$$
\begin{equation*}
\mathcal{D}_{\infty}(X) \xrightarrow[\cong]{\cong} \infty-\mathbf{E M o d}(\operatorname{Pred}(X),[0,1]) \tag{6.25}
\end{equation*}
$$

also given by validity: $\mathcal{V}(\omega)(p):=\omega \vDash p$. This isomorphism is natural in $X$, like in Theorem 6.8.1 (2).

Proof. Much of this works as in the proof of Theorem 6.8.1. except that we have to prove that $\mathcal{V}(\omega): \operatorname{Pred}(X) \rightarrow[0,1]$ is $\infty$-continuous, now that we have $\omega \in \mathcal{D}_{\infty}(X)$. So let $\left(p_{n}\right)$ be an ascending chain of predicates, with pointwise join $p=\bigvee_{n} p_{n}$. Then:

$$
\begin{aligned}
\mathcal{V}(\omega)\left(\bigvee_{n} p_{n}\right)=\omega \vDash \bigvee_{n} p_{n} & =\sum_{x \in X} \omega(x) \cdot\left(\bigvee_{n} p_{n}\right)(x) \\
& =\sum_{x \in X} \omega(x) \cdot\left(\bigvee_{n} p_{n}(x)\right) \\
& =\sum_{x \in X} \bigvee_{n} \omega(x) \cdot p_{n}(x) \\
& \stackrel{(*)}{=} \bigvee_{n} \sum_{x \in X} \omega(x) \cdot p_{n}(x)=\bigvee_{n} \mathcal{V}(\omega)\left(p_{n}\right) .
\end{aligned}
$$

The direction $(\geq)$ of the marked equation $\stackrel{(*)}{=}$ holds by monotonicity. For ( $\leq$ ) we reason as follows. Since $X$ is countable, we can write it as $X=\left\{x_{1}, x_{2}, \ldots\right\}$. For each $N \in \mathbb{N}$ we can use that finite sums preserve $\infty$-joins in:

$$
\sum_{k \leq N} \bigvee_{n} \omega\left(x_{k}\right) \cdot p_{n}\left(x_{k}\right)=\bigvee_{n} \sum_{k \leq N} \omega\left(x_{k}\right) \cdot p_{n}\left(x_{k}\right) \leq \bigvee_{n} \sum_{x \in X} \omega(x) \cdot p_{n}(x)
$$

Hence:

$$
\sum_{x \in X} \bigvee_{n} \omega(x) \cdot p_{n}(x)=\lim _{N \rightarrow \infty} \sum_{k \leq N} \bigvee_{n} \omega\left(x_{k}\right) \cdot p_{n}\left(x_{k}\right) \leq \bigvee_{n} \sum_{x \in X} \omega(x) \cdot p_{n}(x)
$$

The inverse of $\mathcal{V}$ in 6.25) is defined as countable formal sum, using that $X$ is countable: for a homomorphism of $\infty$-effect modules $h: \operatorname{Pred}(X) \rightarrow[0,1]$ we take:

$$
\mathcal{V}^{-1}(h):=\sum_{x \in X} h\left(\mathbf{1}_{x}\right)|x\rangle .
$$

In order to see that the probabilities $h\left(\mathbf{1}_{x}\right)$ add up to one, we write the set $X$ as $X=\left\{x_{1}, x_{2}, \ldots\right\}$. We express the sum over these $x_{i}$ via a join of an ascending chain:

$$
\begin{aligned}
\sum_{x \in X} h\left(\mathbf{1}_{x}\right)=\bigvee_{n} \sum_{k \leq n} h\left(\mathbf{1}_{x_{k}}\right) & =\bigvee_{n} h\left(\bigvee_{k \leq n} \mathbf{1}_{x_{k}}\right) \\
& =h\left(\bigvee_{n} \mathbf{1}_{\left\{x_{1}, \ldots, x_{n}\right\}}\right)=h(\mathbf{1})=1 .
\end{aligned}
$$

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 Theorem6.8.1

## Exercises

6.8.1 Consider the representation isomorphism in Theorem 6.8.1

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 $h \in \operatorname{EMod}(\operatorname{Pred}(X),[0,1])$ and $k \in \operatorname{EMod}(\operatorname{Pred}(Y),[0,1])$ one has:

$$
(h \otimes k)[1,0]=h .
$$

6.8.2 Recall from Theorem4.2.5 that $\operatorname{Pred}(X)$ 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:

$$
\mathbf{E A}(\mathcal{P}(X),[0,1]) \cong \mathbf{E M o d}(\operatorname{Pred}(X),[0,1])
$$

Describe this isomorphism in detail.
2 Conclude that an alternative version of the representation isomorphism (6.24) is:

$$
\mathcal{D}(X) \longrightarrow \mathbf{E A}(\mathcal{P}(X),[0,1])
$$

6.8.3 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 (2.3). Describe the corresponding homomorphism of $\infty$-effect $\operatorname{modules} \operatorname{Pred}(\mathbb{N}) \rightarrow[0,1]$ via (6.25).
6.8.4 Let $(A, \leq)$ be an $\infty$-cpo.

1 Suppose that $A$ has a least element $\mathbf{0} \in A$. For an $\infty$-continuous function $f: A \rightarrow A$ define:

$$
f i x(f):=\bigvee_{n \in \mathbb{N}} f^{n}(\mathbf{0}) \in A
$$

Check that this is well-defined, i.e. that the chain $\left(f^{n}(0)\right)$ of iterated applications of $f$ is ascending. Prove that $\operatorname{fix}(f) \in A$ is the least fixed point of $f$, i.e. that: $f(f i x(f))=f i x(f)$ and for any $x \in A$, if $f(x)=x$, then $\operatorname{fix}(f) \leq x$.
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 $\left(a_{n}\right)$,

$$
\bigvee_{n} a_{n} \in U \Longrightarrow a_{n} \in U, \text { for some } n \in \mathbb{N}
$$

Check that the Scott open subsets form a topology on $A$.
3 Let $f: A \rightarrow B$ be a monotone function, where $B$ is also an $\infty$-cpo. Show that:
$f$ is $\infty$-continuous $\Longleftrightarrow f$ is topologically continuous,
where $\infty$-continuity means that $f$ preserves $\infty$-joins (see Definition 6.8.4 (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 $\infty$-effect module, and thus implicitly as $\infty$-cpo. Check that the Scott open subsets are of the form $(r, 1] \subseteq[0,1]$ for $r \geq 0$.

## 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 $X \rightarrow Y$ into a channel $Y \leadsto X$ in the other direction. This reversal of channels is also called Bayesian inversion. 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 [24]). In the last section of this chapter we show that there is also neat categorical structure involved: channel inversion forms a 'dagger' functor $\mathbf{K r n} \rightarrow \mathbf{K r n}{ }^{\text {op }}$, on a category Krn of kernels (as in [27]).

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:

into:


Thus, disintegration changes an outgoing wire (here labeled $X$ ) into an incoming wire. Doing so in a probabilistic setting involves re-normalisation. 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, like learning with missing data, naive Bayesian classification, and in decision tree learning. Subsequently,
in Section7.6the 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.

We also use dagger channels 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. It is in general not well understood when one should use which rule. 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. In Section 7.7 we introduce the basics of Jeffrey's rule, with several illustrations. Further explanations and examples will be given in Chapter ?? on statistical learning.

As already announced, the final section of this chapter captures channel reversal in categorical terms and illustrates the analogies 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 $X \mapsto Y$ into a channel $Y \leadsto X$, 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 [27]. 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 [28]. This categorical approach is sketched in Section 7.9 below.

Suppose we have a channel $c: X \leadsto Y$ and a distribution $\omega \in \mathcal{D}(X)$. How can we obtain a channel $Y \leftrightarrow X$. 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 $c=\ll \mathbf{1}_{y}$ on $X$ via predicate transformation. The update $\left.\omega\right|_{c \approx<\mathbf{1}_{y}}$ now gives a new distribution on $X$, via backward inference. In this way we have constructed a function $Y \rightarrow \mathcal{D}(X)$.

There is one caveat: the update $\left.\omega\right|_{c=<\mathbf{1}_{y}}$ must exist, that is we need to have that the non-zero validity of:

$$
\omega \vDash c=<\mathbf{1}_{y} \stackrel{\sqrt{4.11}}{=} c \gg=\omega \vDash \mathbf{1}_{y}=(c \gg=\omega)(y) .
$$

Thus, in order to form the dagger, we need to require that the image distribution $c \gg=\omega$ has full support, so that $(c \gg=\omega)(y) \neq 0$ 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 $c: X \mapsto Y$ 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 $c \gg=\omega \in \mathcal{D}(Y)$ has full support.

The reversed 'dagger' channel $c_{\omega}^{\dagger}: Y \nrightarrow X$ is defined as on $y \in Y$ as:

$$
\begin{align*}
c_{\omega}^{\dagger}(y):=\left.\omega\right|_{c=\ll \mathbf{1}_{y}} & =\sum_{x \in X} \frac{\omega(x) \cdot c(x)(y)}{(c \gg \omega)(y)}|x\rangle  \tag{7.1}\\
& =\sum_{x \in X} \frac{\omega(x) \cdot c(x)(y)}{\sum_{z} \omega(z) \cdot c(z)(y)}|x\rangle .
\end{align*}
$$

We have already implicitly used the dagger of a channel in many situations. We list several of them below, together with some new examples.

Example 7.1.2. In Example 6.2.3 we used a channel $c:\{H, T\} \leadsto\{W, B\}$ from the two sides of a coin to two colors of balls in an urn, together with a fair coin 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:

$$
c_{u n i f}^{\dagger}(W)=\left.u n i f\right|_{c:\left\langle\mathbf{1}_{W}\right.}=\frac{22}{67}|H\rangle+\frac{45}{67}|T\rangle .
$$

The same redescription in terms of Bayesian inversion can be used for Examples 6.2.4 and 6.2.10, 6.2.11 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 make some special cases of the dagger of a channel explicit.
Lemma 7.1.3. Let $X, Y$ be sets, where $Y$ is finite.
1 When also $X$ is finite, $c: X \leadsto Y$ is a channel, and the uniform distribution unif $_{X}=\sum_{x \in X} \frac{1}{|X|}|x\rangle$ is used as prior, then:

$$
c_{u n i i_{X}}^{\dagger}(y)=\sum_{x \in X} \frac{c(x)(y)}{\sum_{z \in X} c(z)(y)}|x\rangle .
$$

2 When $f: X \rightarrow Y$ is a surjective function, used as deterministic channel $X \mapsto Y$, then for an $\omega \in \mathcal{D}(X)$ such that $\mathcal{D}(f)(\omega)$ has full support,

$$
f_{\omega}^{\dagger}(y)=\sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\mathcal{D}(f)(\omega)(y)}|x\rangle
$$

In this situation one has $f \odot f_{\omega}^{\dagger}=i d$, or, more explicitly, for $y \in Y$,

$$
\left(f \odot f_{\omega}^{\dagger}\right)(y)=f \gg f_{\omega}^{\dagger}(y)=\mathcal{D}(f)\left(f_{\omega}^{\dagger}(y)\right)=1|y\rangle .
$$

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 2.4.6.

$$
f_{u n i f_{X}}^{\dagger}=f^{\sim 1} .
$$

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 $\langle f\rangle=$ unit $\circ f: X \rightarrow \mathcal{D}(Y)$ is the deterministic channel corresponding to $f$, where the $\langle\cdot\rangle$ brackets are often omitted. Thus:

$$
\begin{aligned}
f_{\omega}^{\dagger}(y) & \stackrel{77.1}{=} \sum_{x \in X} \frac{\omega(x) \cdot\langle f\rangle(x)(y)}{\sum_{z \in X} \omega(z) \cdot\langle f\rangle(z)(y)}|x\rangle \\
& =\sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\sum_{z \in f^{-1}(y)} \omega(z)}|x\rangle \\
& =\sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\mathcal{D}(f)(\omega)(y)}|x\rangle \quad \text { see Lemma[2.1.3. }
\end{aligned}
$$

Then:

$$
\begin{aligned}
\mathcal{D}(f)\left(f_{\omega}^{\dagger}(y)\right) & =\sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\mathcal{D}(f)(\omega)(y)}|f(x)\rangle \\
& =\sum_{x \in f^{-1}(y)} \frac{\omega(x)}{\mathcal{D}(f)(\omega)(y)}|y\rangle \\
& =1|y\rangle .
\end{aligned}
$$

3 We now have, for a surjective function $f: X \rightarrow Y$,

$$
\begin{aligned}
f_{\text {uni }_{X}}^{\dagger}(y)=\sum_{x \in f^{-1}(y)} \frac{\text { unif }_{X}(x)}{\sum_{z \in f^{-1}(y)} \text { unif }_{X}(z)}|x\rangle & =\sum_{x \in f^{-1}(y)} \frac{1 /|X|}{\sum_{z \in f^{-1}(y)} 1 /|X|}|x\rangle \\
& =\sum_{x \in f^{-1}(y)} \frac{1}{\left|f^{-1}(y)\right|}|x\rangle \\
& \text { [2.29) } f^{\sim 1}(y) .
\end{aligned}
$$

Theorem6.3.2 describes updating when a joint distribution can be written in two different ways as graph. In the special case (6.7) 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 the earlier examples, now as daggers / reversals of channels.

Example 7.1.4. In Example 6.3.3 we have seen several instances of backward inference $\left.\omega\right|_{c \lll \mathbf{1}_{y}}$ that we now recognise as daggers of the channel $c$.

Originally, in Example 2.4.5 we have introduced the arrangement channel arr: $\mathcal{N}[K](X) \rightsquigarrow X^{K}$ as the probabilistic inverse of the accumulation function acc: $X^{K} \rightarrow \mathcal{N}[K](X)$, see Definition 2.4.6. Using Lemma7.1.3 3) we see that we can now write:

$$
\operatorname{arr}=\operatorname{acc}^{\sim 1}=\operatorname{acc}_{u n i f_{X} K}^{\dagger}
$$

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 7.1.3 11,

$$
\operatorname{arr}_{u n i f_{N[K](x)}}^{\dagger}(\vec{x})=\sum_{\varphi \in \mathcal{N}[K](X)} \frac{\operatorname{arr}(\varphi)(\vec{x})}{\sum_{\psi \in \mathcal{N}[K](X)} \operatorname{arr}(\psi)(\vec{x})}|\varphi\rangle=1|\operatorname{acc}(\vec{x})\rangle=\langle\operatorname{acc}\rangle(\vec{x}) .
$$

Calculating the dagger of a channel may require some work. We illustrate this for the hypergeometric channel, see Definition 2.6.1 and Section 3.4 for details. We describe its dagger for four different priors: uniform, multinomial, Pólya and finally multinomial itself.

Theorem 7.1.5. Consider the hypergeometric channel $h g[K]: \mathcal{N}[L](X) \rightarrow$ $\mathcal{N}[K](X)$, for numbers $L \geq K$

1 Let $X$ be a non-empty and finite set, say with $N:=|X|>0$ elements. We take as prior the uniform distribution on the (finite) set $\mathcal{N}[L](X)$ of multisets of size $L$. The resulting dagger acts on $\varphi \in \mathcal{N}[K](X)$ and produces a distribution on $\mathcal{N}[L](X)$, of the form:

$$
h g[K]_{u n i f_{\mathcal{N}[L](x)}}^{\dagger}(\varphi)=\sum_{\psi \in \mathcal{N}[L-K](X)} \frac{\binom{\varphi+\psi}{\varphi}}{\binom{L+N-1}{K+N-1}}|\varphi+\psi\rangle .
$$

2 Let $\omega \in \mathcal{D}(X)$ be a distribution; taking the multinomial distribution $m n[L](\omega)$
on $\mathcal{N}[L](X)$ as prior gives:

$$
h g[K]_{m n[L](\omega)}^{\dagger}(\varphi)=\sum_{\psi \in \mathcal{N}[L-K](X)} m n[L-K](\omega)(\psi)|\varphi+\psi\rangle .
$$

3 Let $X$ be non-empty and finite again and let $v \in \mathcal{N}_{f s}(X)$ be an urn with full support; then:

$$
h g[K]_{p l[L](v)}^{\dagger}(\varphi)=\sum_{\psi \in \mathcal{N}[L-K](X)} p l[L-K](v+\varphi)(\psi)|\varphi+\psi\rangle
$$

4 Let $v \in \mathcal{N}[M](X)$ be a multiset of size $M \geq L$, where by assumption $L \geq K$. We take the hypergeometric distribution $h g[L](v)$ as prior. In order to keep things well-defined we have to restrict the domain of the dagger to a subset of $\mathcal{N}[K](X)$, as in:

$$
\left\{\varphi \in \mathcal{N}[K](X) \mid \varphi \leq_{K} v\right\} \xrightarrow{\substack{h g[K]_{g,[L](v)}^{\dagger}}} \mathcal{N}[L](X)
$$

On such $a \varphi \leq_{K} v$ this dagger is given by:

$$
h g[K]_{h g[L](v)}^{\dagger}(\varphi)=\sum_{\psi \leq L-K^{v}-\varphi} \operatorname{hg}[L-K](v-\varphi)(\psi)|\varphi+\psi\rangle .
$$

Proof. 1 Since we have a uniform distribution as prior, Lemma 7.1.3 (1) applies:

$$
\begin{aligned}
& h g[K]_{u_{n i f}}^{\dagger}(\varphi)=\sum_{v \in \mathcal{N}[L](X), \varphi \leq_{K}} \frac{h g[K](v)(\varphi)}{\sum_{v \in \mathcal{N}[L](X), \varphi \leq_{K^{v}}} \operatorname{hg}[K](v)(\varphi)}|v\rangle
\end{aligned}
$$

$$
\begin{aligned}
& =\sum_{\psi \in \mathcal{N}[L-K](X)} \frac{\binom{\varphi+\psi}{\varphi}}{\binom{L+N-1}{K+N-1}}|\varphi+\psi\rangle \quad \text { by Exercise } 1.8 .8
\end{aligned}
$$

2 We use that $h g[K] \odot m n[L]=m n[K]$, see Corollary 3.4.2 (6).

$$
\begin{aligned}
h g[K]_{m n[L](\omega)}^{\dagger}(\varphi) & =\sum_{v \in \mathcal{N}[L](X)} \frac{h g[K](v)(\varphi) \cdot m n[L](\omega)(v)}{(h g[K] \gg m n[L](\omega))(\varphi)}|v\rangle \\
& =\sum_{v \in \mathcal{N}[L](X), \varphi \leq_{K} v} \frac{\binom{v}{\varphi}}{\binom{L}{K}} \cdot \frac{L!}{v \rrbracket} \cdot \frac{\varphi \rrbracket}{K!} \cdot \frac{\prod_{x \in X} \omega(x)^{v(x)}}{\prod_{x \in X} \omega(x)^{\varphi(x)}}|v\rangle \\
& =\sum_{v \in \mathcal{N}[L](X), \varphi \leq K_{v}} \frac{(L-K)!}{(v-\varphi) \rrbracket} \cdot \prod_{x \in X} \omega(x)^{v(x)-\varphi(x)}|v\rangle \\
& =\sum_{\psi \in \mathcal{N}[L-K](X)} \operatorname{mn}[L-K](\omega)(\psi)|\varphi+\psi\rangle .
\end{aligned}
$$

3 Let urn $v \in \mathcal{N}_{f s}(X)$ have size $M:=\|v\|>0$. We use that $h g[K] \odot p l[L]=$ $p l[K]$, see Proposition 3.5.2 (3). Then, for $\varphi \in \mathcal{N}[K](X)$,

$$
\begin{aligned}
& h g[K]_{p l[L](v)}^{\dagger}(\varphi)=\sum_{\chi \in \mathcal{N}[L](X)} \frac{h g[K](\chi)(\varphi) \cdot p l[L](v)(\chi)}{(h g[K] \gg p l[L](v))(\varphi)}|\chi\rangle \\
& =\sum_{\chi \in \mathcal{N}[L](X), \varphi \leq_{K} \chi} \frac{\binom{\chi}{\varphi}}{\binom{L}{K}} \cdot \frac{\left(\binom{v}{\chi}\right.}{\left(\binom{M}{L}\right)} \cdot \frac{\left(\binom{M}{K}\right)}{\left(\binom{v}{\varphi}\right)}|\chi\rangle \\
& =\sum_{\chi \in \mathcal{N}[L](X), \varphi \leq K \chi} \frac{\chi \rrbracket \cdot K!\cdot(L-K)!}{\varphi \rrbracket \cdot(\chi-\varphi) \rrbracket \cdot L!} \cdot \frac{(v+\chi-\mathbf{1}) \rrbracket \cdot L!\cdot(M-1)!}{\chi \square \cdot(v-\mathbf{1}) \rrbracket \cdot(M+L-1)!} \\
& \cdot \frac{(M+K-1)!\cdot \varphi \rrbracket \cdot(v-\mathbf{1}) \rrbracket}{K!\cdot(M-1)!\cdot(v+\varphi-\mathbf{1}) \rrbracket}|\chi\rangle \\
& =\sum_{\chi \in \mathcal{N}[L](X), \varphi \leq{ }_{K} \chi} \frac{(L-K)!\cdot(v+\chi-\mathbf{1}) \rrbracket \cdot(M+K-1)!}{(\chi-\varphi)!\cdot(M+L-1)!\cdot(v+\varphi-\mathbf{1}) \rrbracket}|\chi\rangle \\
& =\sum_{\psi \in \mathcal{N}[L-K](X)} \frac{(L-K)!\cdot(v+\varphi+\psi-\mathbf{1}) \rrbracket \cdot(M+K-1)!}{\psi \rrbracket \cdot(M+L-1)!\cdot(v+\varphi-\mathbf{1}) \rrbracket}|\varphi+\psi\rangle \\
& =\sum_{\psi \in \mathcal{N}[L-K](X)} \frac{\left(\binom{v+\varphi}{\psi}\right.}{\binom{M+K}{L-K}}|\varphi+\psi\rangle \\
& =\sum_{\psi \in \mathcal{N}[L-K](X)} p l[L-K](v+\varphi)(\psi)|\varphi+\psi\rangle .
\end{aligned}
$$

4 We assume that $\varphi \in \mathcal{N}[K](X)$ satisfies $\varphi \leq_{K} v$ and we write $M:=\|v\| \geq L \geq$
$K$. Then, by Corollary 3.4.2 (5),

$$
\begin{aligned}
& h g[K]_{h g[L](v)}^{\dagger}(\varphi) \\
& =\sum_{\chi \in \mathcal{N}[L](X)} \frac{h g[K](\chi)(\varphi) \cdot h g[L](v)(\chi)}{(h g[K] \gg h g[L](v))(\varphi)}|\chi\rangle \\
& =\sum_{\chi \in \mathcal{N}[L](X), \varphi \leq_{K \chi, \chi \leq L}} \frac{\binom{\chi}{\varphi}}{\binom{L}{K}} \cdot \frac{\binom{v}{\chi}}{\binom{M}{L}} \cdot \frac{\binom{M}{K}}{\binom{v}{\varphi}}|\chi\rangle \\
& =\sum_{\chi \in \mathcal{N}[L](X), \varphi \leq_{K \chi, \chi \leq L}} \frac{\chi \square \cdot K!\cdot(L-K)!}{\varphi \rrbracket \cdot(\chi-\varphi) \rrbracket \cdot L!} \cdot \frac{v \rrbracket \cdot L!\cdot(M-L)!}{\chi \square \cdot(v-\chi) \square \cdot M!} \\
& \cdot \frac{M!\cdot \varphi!\cdot(v-\varphi) \rrbracket}{K!\cdot(M-K)!\cdot v \rrbracket}|\chi\rangle \\
& =\sum_{\chi \in \mathcal{N}[L](X), \varphi \leq_{K} \chi, \chi \leq_{L^{\nu}}} \frac{(L-K)!\cdot(M-L)!\cdot(v-\varphi) \rrbracket}{(\chi-\varphi)!\cdot(M-K)!\cdot(v-\chi) \rrbracket}|\chi\rangle \\
& =\sum_{\psi \in \mathcal{N}[L-K](X), \varphi+\psi \leq L v} \frac{(L-K)!\cdot(M-L)!\cdot(v-\varphi) \rrbracket}{\psi \rrbracket \cdot(M-K)!\cdot(v-\varphi-\psi) \rrbracket}|\varphi+\psi\rangle \\
& =\sum_{\psi \leq L-K \cup-\varphi} \frac{\binom{v-\varphi}{\psi}}{\binom{M-K}{L-K}}|\varphi+\psi\rangle \\
& =\sum_{\psi \leq L-K^{v}-\varphi} h g[L-K](v-\varphi)(\psi)|\varphi+\psi\rangle \text {. }
\end{aligned}
$$

The above hypergeometric-dagger channels give for a draw $\varphi$ of size $K$, a distribution over urns of size $L$ from which $\varphi$ can be drawn, with the corresponding (normalised) likelihoods. For instance, using the uniform prior, for $\varphi=1|a\rangle+2|b\rangle$ of size $K=3$, when we consider urns of size $L=5$ with balls of colours $a, b$, from which $\varphi$ can be drawn, we get:

$$
\begin{aligned}
& h g[3]_{u n i f_{N(S \mid(a, b)]}}^{\dagger}(1|a\rangle+2|b\rangle) \\
& \left.\left.\left.\left.\left.\left.\quad=\frac{1}{5}|3| a\right\rangle+2|b\rangle\right\rangle+\frac{2}{5}|2| a\right\rangle+3|b\rangle\right\rangle+\frac{2}{5}|1| a\right\rangle+4|b\rangle\right\rangle .
\end{aligned}
$$

We can apply the hypergeometric channel to these urns, giving:

$$
\begin{aligned}
& h g[3] \gg=\left(h g[3]_{u n i f_{N(5)(a, b))}^{\dagger}}^{\dagger}(1|a\rangle+2|b\rangle)\right) \\
& \left.\left.\left.\left.\left.\left.\left.\left.\quad=\frac{1}{50}|3| a\right\rangle\right\rangle+\frac{6}{25}|2| a\right\rangle+1|b\rangle\right\rangle+\frac{27}{50}|1| a\right\rangle+2|b\rangle\right\rangle+\frac{1}{5}|3| b\right\rangle\right\rangle .
\end{aligned}
$$

Unsurprisingly, the original draw $\varphi=1|a\rangle+2|b\rangle$ gets the highest probability.
In general, in the situation of Definition 7.1.1, the pushforward $c \gg c_{\omega}^{\dagger}(y)$ of the inversion is called the posterior predictive distribution.

The following result gives the string-diagrammatic characterisation of dagger channels that is used in [24].

Theorem 7.1.6. Let $c: X \leadsto Y$ be a channel, with finite codomain $Y$, together with a distribution $\omega \in \mathcal{D}(X)$ such that $c \gg=\omega$ has full support. The resulting dagger channel $c_{\omega}^{\dagger}$ is the unique channel $Y \rightsquigarrow X$ satisfying:


In particular, by marginalising out the first wires we get:

$$
\begin{equation*}
\omega=c_{\omega}^{\dagger} \gg=(c \gg=\omega) . \tag{7.3}
\end{equation*}
$$

This is in essence the law of total probability, see Exercise 7.1.6 below.
Proof. We first check that (7.2) holds.

$$
\begin{aligned}
\left\langle\text { id }, c_{\omega}^{\dagger}\right\rangle \gg=(c \gg=\omega) & =\sum_{y \in Y} \sum_{x \in X}(c \gg=\omega)(y) \cdot c_{\omega}^{\dagger}(y)(x)|y, x\rangle \\
& \stackrel{7.1}{=} \sum_{y \in Y} \sum_{x \in X}(c \gg=\omega)(y) \cdot \frac{\omega(x) \cdot c(x)(y)}{(c \gg \omega)(y)}|y, x\rangle \\
& =\sum_{y \in Y} \sum_{x \in X} \omega(x) \cdot c(x)(y)|y, x\rangle=\langle c, \text { id }\rangle \gg=\omega .
\end{aligned}
$$

For uniqueness, let Equation (7.2) hold with dagger $d: Y \rightsquigarrow X$ instead of $c_{\omega}^{\dagger}$. Then we can apply Equation 6.7) from Theorem6.3.2, giving for each $y \in Y$,

$$
d(y)=\left.\omega\right|_{c=<\mathbf{1}_{y}} \stackrel{7.1}{-} c_{\omega}^{\dagger}(y)
$$

A consequence (from [79]) 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 $\tau:=c \gg=\omega$.
1 Given a factor $q$ on $Y$, we can express backward inference as forward inference with the dagger:

$$
\left.\omega\right|_{c \approx \ll q}=c_{\omega}^{\dagger} \gg=\left.\tau\right|_{q} .
$$

2 Given a factor pon $X$, we can express forward inference as backward inference with the dagger:

$$
\left.c \gg \omega\right|_{p}=\left.\tau\right|_{c_{\omega}^{\star}=\ll p} .
$$

Proof. We combine Theorems 7.1.6 and 6.3.2 Item (1) follows directly from Equation (6.6). Also item (2) follows from Equation (6.6), after left-right mirroring of the two string diagrams in 7.2 .

We conclude 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. In the following we assume that daggers are well-defined, i.e. that the relevant pushforwards have full support, see Definition 7.1.1.

1 For channels $c: X \rightsquigarrow U$ and $d: Y \rightsquigarrow V$, with distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$,

$$
(c \otimes d)_{\omega \otimes \rho}^{\dagger}=c_{\omega}^{\dagger} \otimes d_{\rho}^{\dagger}
$$

2 For composable channels $X \xrightarrow{c} Y \xrightarrow{d} Z$ and a distribution $\omega \in \mathcal{D}(X)$, there is the following 'chain rule'.

$$
(d \odot c)_{\omega}^{\dagger}=c_{\omega}^{\dagger} \odot d_{c \gg=\omega}^{\dagger} .
$$

3 Identity channels are stable under reversal:

$$
i d_{\omega}^{\dagger}=i d
$$

4 Double reversal returns the original channel:

$$
\left(c_{\omega}^{\dagger}\right)_{c \geqslant \omega}^{\dagger}=c .
$$

Proof. 1 For elements $u \in U$ and $v \in V$ we have:

$$
\begin{aligned}
(c \otimes d)_{\omega \otimes \rho}^{\dagger}(u, v) & =\sum_{(x, y) \in X \times Y} \frac{(\omega \otimes \rho)(x, y) \cdot(c \otimes d)(x, y)(u, v)}{((c \otimes d) \gg(\omega \otimes \rho))(u, v)}|x, y\rangle \\
& =\sum_{(x, y) \in X \times Y} \frac{\omega(x) \cdot \rho(y) \cdot c(x)(u) \cdot d(y)(v)}{((c \gg \omega) \otimes(d \gg \rho))(u, v)}|x, y\rangle \\
& =\sum_{(x, y) \in X \times Y} \frac{\omega(x) \cdot c(x)(u)}{(c \gg \omega)(u)} \cdot \frac{\rho(y) \cdot d(y)(v)}{(d \gg \rho)(v)}|x, y\rangle \\
& =\sum_{(x, y) \in X \times Y} c_{\omega}^{\dagger}(u)(x) \cdot d_{\rho}^{\dagger}(u)(y)|x, y\rangle \\
& =c_{\omega}^{\dagger}(u) \otimes d_{\rho}^{\dagger}(u) \\
& =\left(c_{\omega}^{\dagger} \otimes d_{\rho}^{\dagger}\right)(u, v) .
\end{aligned}
$$

2 For $z \in Z$ we have:

$$
\begin{aligned}
\left(c_{\omega}^{\dagger} \odot d_{c \gg \omega}^{\dagger}\right)(z) & =c_{c}^{\dagger} \gg d_{c \gg \omega}^{\dagger}(z) \\
& =\sum_{x \in X} \sum_{y \in Y} d_{c \gg \omega}^{\dagger}(z)(y) \cdot c_{\omega}^{\dagger}(y)(x)|x\rangle \\
& =\sum_{x \in X} \sum_{y \in Y} \frac{(c \gg \omega)(y) \cdot d(y)(z)}{(d \gg(c)>=\omega))(z)} \cdot \frac{\omega(x) \cdot c(x)(y)}{(c \gg=\omega)(y)}|x\rangle \\
& =\sum_{x \in X} \frac{\omega(x) \cdot \sum_{y \in Y} c(x)(y) \cdot d(y)(z)}{((d \odot c) \gg=\omega)(z)}|x\rangle \\
& =\sum_{x \in X} \frac{\omega(x) \cdot(d \odot c)(x)(z)}{((d \odot c) \gg \omega)(z)}|x\rangle \\
& =(d \odot c)_{\omega}^{\dagger}
\end{aligned}
$$

3 For an element $x \in X$,

$$
i d_{\omega}^{\dagger}(x)=\sum_{z \in X} \frac{\omega(z) \cdot \operatorname{id}(z)(x)}{(\operatorname{id} \gg=\omega)(x)}|z\rangle=\frac{\omega(x)}{\omega(x)}|x\rangle=1|x\rangle=\operatorname{id}(x) .
$$

4 Finally, for $x \in X$,

$$
\begin{aligned}
\left(c_{\omega}^{\dagger}\right)_{c \gg=\omega}^{\dagger}(x) & =\sum_{y \in Y} \frac{(c \gg=\omega)(y) \cdot c_{\omega}^{\dagger}(y)(x)}{\left(c_{\omega}^{\dagger} \gg(c \gg=\omega)\right)(x)}|y\rangle \\
& \stackrel{7.3]}{=} \sum_{y \in Y} \frac{(c \gg=\omega)(y)}{\omega(x)} \cdot \frac{\omega(x) \cdot c(x)(y)}{(c \gg=\omega)(y)}|y\rangle \\
& =\sum_{y \in Y} c(x)(y)|y\rangle \\
& =c(x) .
\end{aligned}
$$

One can prove these results via equational reasoning with string diagrams, using uniqueness of daggers. For instance preservation of composition $\odot$ by the dagger, in item (2), follows by uniqueness from:


The next result (from [94]) shows that the dagger (also) interacts nicely with
the extension $\mathcal{N}[K]$ : Chan $\rightarrow$ Chan of the multinomial functor to the category of channels, see Corollary 3.7.8

Theorem 7.1.9. Let $c: X \rightarrow Y$ be a channel to a finite set $Y$, with a distribution $\omega \in \mathcal{D}(X)$ such that $c \gg=\omega$ has full support. Fix a number $K \in \mathbb{N}$.

1 The $K$-fold tensor product of channels $c^{K}=c \otimes \cdots \otimes c: X^{K} \mapsto Y^{K}$ satisfies the following equation.


As a result, the dagger commutes with $K$-fold tensor product:

$$
\begin{equation*}
\left(c^{K}\right)_{i i d[K](\omega)}^{\dagger}=\left(c_{\omega}^{\dagger}\right)^{K} \tag{7.5}
\end{equation*}
$$

2 The multiset extension of channels $\mathcal{N}[K](c): \mathcal{N}[K](X) \leadsto \mathcal{N}[K](Y)$ satisfies:


We can now conclude that the dagger commutes with multiset-extension:

$$
\begin{equation*}
\mathcal{N}[K](c)_{\operatorname{mn}[K](\omega)}^{\dagger}=\mathcal{N}[K]\left(c_{\omega}^{\dagger}\right) . \tag{7.7}
\end{equation*}
$$

Proof. 1 We start on the right-hand-side of (7.4):

$$
\begin{aligned}
& \left\langle i d,\left(c_{\omega}^{\dagger}\right)^{K}\right\rangle \gg=\operatorname{iid}[K](c \gg=\omega) \\
& =\sum_{\vec{y} \in Y^{K}} \sum_{\vec{x} \in X^{K}}\left(c_{\omega}^{\dagger}\right)^{K}(\vec{y})(\vec{x}) \cdot \operatorname{iid}[K](c \gg=\omega)(\vec{y})|\vec{y}, \vec{x}\rangle \\
& =\sum_{\vec{y} \in Y^{K}} \sum_{\vec{x} \in X^{K}} \prod_{1 \leq i \leq K} c_{\omega}^{\dagger}\left(y_{i}\right)\left(x_{i}\right) \cdot(c \gg=\omega)\left(y_{i}\right)|\vec{y}, \vec{x}\rangle \\
& =\sum_{\vec{y} \in Y^{K}} \sum_{\vec{x} \in X^{K}} \prod_{1 \leq i \leq K} \frac{\omega\left(x_{i}\right) \cdot c\left(x_{i}\right)\left(y_{i}\right)}{(c \gg=\omega)\left(y_{i}\right)} \cdot(c \gg=\omega)\left(y_{i}\right)|\vec{y}, \vec{x}\rangle \\
& =\sum_{\vec{y} \in Y^{K}} \sum_{\vec{x} \in X^{K}} \prod_{1 \leq i \leq K} c\left(x_{i}\right)\left(y_{i}\right) \cdot \omega\left(x_{i}\right)|\vec{y}, \vec{x}\rangle \\
& =\sum_{\vec{y} \in Y^{K}} \sum_{\vec{x} \in X^{K}} c^{K}(\vec{x})(\vec{y}) \cdot \operatorname{iid}[K](\omega)(\vec{x})|\vec{y}, \vec{x}\rangle \\
& =\left\langle i d, c^{K}\right\rangle \gg=\operatorname{iid}[K](\omega) .
\end{aligned}
$$

2 We recall that $\mathcal{N}[K](c)=\operatorname{acc} \odot c^{K} \odot \operatorname{arr}: \mathcal{N}[K](X) \rightsquigarrow X^{K} \mapsto Y^{K} \leadsto \mathcal{N}[K](Y)$, see Corollary 3.7.8. We use the previous point, in combination with the accarr equation 3.14.


A direct proof, in the style of item (1), is much harder.
Notice that the naturality of the $\operatorname{iid}[K]$ and $m n[K]$ channels - with respect product and multiset functors extended to the category Chan, in Exercise 2.4.8 (4) and in Theorem 3.7.12 - is a consequence of Equations (7.4) and (7.6), via marginalising out the second wires. Concretely, in this way one obtains the first equation in (3.43). In fact, marginalising out the first wires also gives naturality equations, in combination with (7.3).

## Exercises

7.1.1 Consider the sets $X=\{1,2\}$ and $Y=\{a, b, c\}$ with channel $c: X \leadsto Y$ given by:

$$
c(1)=\frac{1}{2}|a\rangle+\frac{1}{3}|b\rangle+\frac{1}{6}|c\rangle \quad c(2)=\frac{1}{4}|a\rangle+\frac{3}{8}|b\rangle+\frac{1}{8}|c\rangle .
$$

Consider the following two distributions on $X$.

$$
\omega_{1}=\frac{1}{2}|1\rangle+\frac{1}{2}|2\rangle \quad \omega_{1}=\frac{1}{3}|1\rangle+\frac{2}{3}|2\rangle
$$

1 Check that $c \gg=\omega_{1}$ and $c \gg=\omega_{2}$ both have full support.
2 Compute the dagger channels $c_{\omega_{1}}^{\dagger}$ and $c_{\omega_{2}}^{\dagger}$.
7.1.2 In Exercises 6.2.1 and 6.2.2 we have seen a channel $c:\left\{d, d^{+}\right\} \rightarrow$ $\{p, n\}$ 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:

$$
\begin{aligned}
& P P V=c_{\omega}^{\dagger}(p)(d)=\left.\omega\right|_{c \ll 1_{p}}(d)=\frac{18}{117} \\
& N P V=c_{\omega}^{\dagger}(n)\left(d^{\perp}\right)=\left.\omega\right|_{c \ll 1_{n}}\left(d^{\perp}\right)=\frac{1881}{1883},
\end{aligned}
$$

where $\omega=\frac{1}{100}|d\rangle+\frac{99}{100}\left|d^{\perp}\right\rangle$ is the prior distribution.
7.1.3 Let $c: X \leadsto Y$ be a channel with distribution $\omega \in \mathcal{D}(X)$, where $c \gg=\omega$ has full support. Consider the following distribution of distributions:

$$
\Omega:=\sum_{y \in Y}(c \gg=\omega)(y)\left|c_{\omega}^{\dagger}(y)\right\rangle \in \mathcal{D}(\mathcal{D}(X)) .
$$

Recall the probabilistic 'flatten' operation from Section 2.4 and show:

$$
\operatorname{flat}(\Omega)=\omega
$$

This says that $\Omega$ is 'Bayes plausible' in the terminology of [110], in the context of Bayesian persuasion. The construction of $\Omega$ is one part of the bijective correspondence in Exercise 7.2.4
7.1.4 Consider the draw-delete channel $D D: \mathcal{N}[K+1](X) \mapsto \mathcal{N}[K](X)$ from Definition 3.2.1 (2). Let $X$ be a finite set, say with $n$ elements. Consider the the uniform state $v:=u_{i f} f_{\mathcal{N}[K+1](X)}$ on $\mathcal{N}[K+1](X)$, see Exercise 2.4.3 Show that DD's dagger, with respect to $v$, can be described on $\varphi \in \mathcal{N}[K](X)$ as:

$$
\left.\left.D D_{v}^{\dagger}(\varphi)=\sum_{x \in X} \frac{\varphi(x)+1}{K+n}|\varphi+1| x\right\rangle\right\rangle
$$

This dagger differs from the draw-add map $D A: \mathcal{N}[K](X) \multimap \mathcal{N}[K+$ $1](X)$.
7.1.5 Let $p$ be a predicate on $X$, considered as a channel $p: X \hookrightarrow \mathbf{2}$, and let $\omega \in \mathcal{D}(X)$ be a distribution. Compute the dagger $p_{\omega}^{\dagger}: \mathbf{2} \rightsquigarrow X$ and show that:

$$
p_{\omega}^{\dagger}(1)=\left.\omega\right|_{p} \quad \text { and } \quad p_{\omega}^{\dagger}(0)=\left.\omega\right|_{p^{\perp}}
$$

7.1.6 $\quad$ Let $\omega \in \mathcal{D}(X)$ be a state.

1 Consider a predicate $p: X \rightarrow[0,1]$ as a channel, as in the previous exercise. Show that the binary version of the law of total probability, see Exercise 6.1.6, can be expressed as:

$$
\omega=p_{\omega}^{\dagger} \gg=\left((\omega \models p)|1\rangle+\left(\omega \vDash p^{\perp}\right)|0\rangle\right) .
$$

2 Let $c: X \leadsto n$ be a channel and define $n$ predicates on $X$ by $p_{i}:=$ $c=\ll \mathbf{1}_{i}$. Show that these $p_{0}, \ldots, p_{n-1}$ form a test:

$$
\otimes_{i} p_{i}=\mathbf{1} \quad \text { with } \quad c_{\omega}^{\dagger}(i)=\left.\omega\right|_{p_{i}} .
$$

3 Conclude that the equation $c_{\omega}^{\dagger} \gg(c \gg=\omega)=\omega$ from Equation 7.3 can also be understood as the ( $n$-ary version of the) law of total probability.
7.1.7 Let $c: X \leadsto Y$ and $d: X \leadsto Z$ be channels with a distribution $\omega \in$ $\mathcal{D}(X)$ on their domain such that both pushforwards $c \gg=\omega$ and $d \gg=\omega$ have full support. Show that:

$$
\begin{aligned}
\langle c, d\rangle_{\omega}^{\dagger}(y, z) & =\left.\omega\right|_{\left(c=\ll \mathbf{1}_{y}\right) \&\left(d \approx<\mathbf{1}_{z}\right)}=d_{c_{\omega}^{\dagger}(y)}^{\dagger}(z) \\
& =\left.\omega\right|_{\left(d \lll \mathbf{1}_{z}\right) \&\left(c \approx\left\langle\mathbf{1}_{y}\right)\right.}=c_{d_{\omega}^{\dagger}(z)}^{\dagger}(y) .
\end{aligned}
$$

7.1.8 In Definition 6.5.4 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 $t: X \mapsto X$, emission channel $e: X \mapsto Y$ and initial state $\sigma \in \mathcal{D}(X)$

$$
\sigma_{1}=\sigma \quad \text { and } \quad \sigma_{i+1}=\left(t \odot e_{\sigma_{i}}^{\dagger}\right)\left(y_{i}\right) .
$$

### 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 $A \rightarrow \mathcal{P}(B)$, and similarly how a multiset $\psi \in \mathcal{M}(A \times B)$ corresponds to an $\mathcal{M}$-channel $A \rightarrow \mathcal{M}(B)$. 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 [57]). 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:

so that:


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 $X \stackrel{\pi_{1}}{\longleftarrow} X \times$ $Y \xrightarrow{\pi_{2}} Y$ as deterministic channels. Let $\omega \in \mathcal{D}(X \times Y)$ be a joint distribution, whose two marginals $\omega_{1}:=\omega[1,0]=\pi_{1} \gg \omega \in \mathcal{D}(X)$ and $\omega_{2}:=\omega[0,1]=$ $\pi_{2} \gg \omega \in \mathcal{D}(Y)$ both have full support.

Extract from the joint distribution $\omega$ the two channels:

$$
c:=\left(X \xrightarrow{\left(\pi_{1}\right)_{\omega}^{\dagger}} X \times Y \xrightarrow{\pi_{2}} Y\right) \quad d:=\left(Y \xrightarrow{\left(\pi_{2}\right)_{\infty}^{\dagger}} X \times Y \xrightarrow{\pi_{1}} X\right)
$$

Then:
1 These channels $c$ and $d$ can be described explicitly as:

$$
\begin{equation*}
c(x):=\sum_{y \in Y} \frac{\omega(x, y)}{\omega_{1}(x)}|y\rangle \quad d(y):=\sum_{x \in X} \frac{\omega(x, y)}{\omega_{2}(y)}|x\rangle . \tag{7.9}
\end{equation*}
$$

2 The joint distribution $\omega$ is the graph of both these channels: $\langle i d, c\rangle \gg=\omega_{1}=$ $\omega=\langle d, i d\rangle \gg=\omega_{2}$, that is:


3 The extracted channels are each other's daggers:

$$
c_{\omega_{1}}^{\dagger}=d \quad \text { and } \quad d_{\omega_{2}}^{\dagger}=c .
$$

We sometimes use the following notation for these extracted channels:

$$
\omega[1,0 \mid 0,1]=c: X \rightarrow Y \quad \text { and } \quad \omega[0,1 \mid 1,0]=d: Y \rightsquigarrow X .
$$

Proof. For all three items we only do the first equations, since the second ones follow by symmetry.

1 By assumption $\pi_{1} \gg=\omega=\omega_{1}$ has full support, so we can form the dagger channel $c:=\left(\pi_{1}\right)_{\omega}^{\dagger}$. It is:

$$
c(x) \stackrel{\sqrt[77.1]{-}}{=} \sum_{x^{\prime} \in X} \sum_{y \in Y} \frac{\omega\left(x^{\prime}, y\right) \cdot\left\langle\pi_{1}\right\rangle\left(x^{\prime}, y\right)(x)}{\left(\pi_{1} \gg \omega\right)(x)}|y\rangle=\sum_{y \in Y} \frac{\omega(x, y)}{\omega_{1}(x)}|y\rangle .
$$

2 Using this formulation, we get, for $x \in X$ and $y \in Y$,

$$
\left(\langle i d, c\rangle \gg=\omega_{1}\right)(x, y)=\omega_{1}(x) \cdot c(x)(y)=\omega(x, y) .
$$

3 Similarly,

$$
\begin{aligned}
& c_{\omega_{1}}^{\dagger}(y)(x) \stackrel{\sqrt[7.1]]{-}}{-\frac{\omega_{1}(x) \cdot c(x)(y)}{\left(c \gg \omega_{1}\right)(y)} \stackrel{\boxed{7.9}}{=} \frac{\omega(x, y)}{\sum_{x \in X} \omega_{1}(x) \cdot c(x)(y)}} \\
&=\frac{\omega(x, y)}{\sum_{x \in X} \omega(x, y)} \\
&=\frac{\omega(x, y)}{\omega_{2}(y)} \frac{7.9}{=} d(y)(x) .
\end{aligned}
$$

From a joint distribution $\omega \in \mathcal{D}(X \times Y)$ we can extract two channels, in both direction, written as $\omega[0,1 \mid 1,0]: X \leadsto Y$ and $\omega[1,0 \mid 0,1]: Y \leadsto X$. The direction of the channels is thus in a certain sense arbitrary, and does not reflect any form of causality, see Chapter ??
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 7.2.2. Let $c: X \leadsto Y$ be channel with a distribution $\sigma \in \mathcal{D}(X)$ such that $c \gg=\sigma$ has full support. Write $\omega:=\langle i d, c\rangle\rangle=\sigma \in \mathcal{D}(X \times Y)$. The channel $\omega[1,0 \mid 0,1]: Y \rightsquigarrow X$ extracted via disintegration is then the dagger $c_{\sigma}^{\dagger}$.

Proof. The distribution $\pi_{2} \gg \omega=\omega \gg=\sigma \in \mathcal{D}(Y)$ has full support. The extraction $\omega[1,0 \mid 0,1]: Y \rightarrow X$ 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 7.2.3. We shall look at two examples, involving spaces $A=\left\{a, a^{\perp}\right\}$ and $B=\left\{b, b^{+}\right\}$.

1 Consider the following state $\omega \in \mathcal{D}(A \times B)$,

$$
\omega=\frac{1}{4}|a, b\rangle+\frac{1}{2}\left|a, b^{\perp}\right\rangle+\frac{1}{4}\left|a^{\perp}, b^{\perp}\right\rangle .
$$

We have as first marginal $\omega_{1}:=\omega[1,0]=\frac{3}{4}|a\rangle+\frac{1}{4}\left|a^{\perp}\right\rangle$ with full support.

The extracted channel $c:=\omega[0,1 \mid 1,0]: A \nrightarrow B$ is given by:

$$
\begin{aligned}
& c(a) \stackrel{77.9}{=} \frac{\omega(a, b)}{\omega_{1}(a)}|b\rangle+\frac{\omega\left(a, b^{\perp}\right)}{\omega_{1}(a)}\left|b^{\perp}\right\rangle=\frac{1 / 4}{3 / 4}|b\rangle+\frac{1 / 2}{3 / 4}\left|b^{\perp}\right\rangle=\frac{1}{3}|b\rangle+\frac{2}{3}\left|b^{\perp}\right\rangle \\
& c\left(a^{\perp}\right) \stackrel{\boxed{77.9}}{=} \frac{\omega\left(a^{\perp}, b\right)}{\omega_{1}\left(a^{\perp}\right)}|b\rangle+\frac{\omega\left(a^{\perp}, b^{\perp}\right)}{\omega_{1}\left(a^{\perp}\right)}\left|b^{\perp}\right\rangle=\frac{0}{1 / 4}|b\rangle+\frac{1 / 4}{1 / 4}\left|b^{\perp}\right\rangle=1\left|b^{\perp}\right\rangle .
\end{aligned}
$$

Then indeed, $\langle i d, c\rangle \gg=\omega_{1}=\omega$.
2 Now let us start from a different joint distribution:

$$
\omega=\frac{1}{3}|a, b\rangle+\frac{2}{3}\left|a, b^{\perp}\right\rangle .
$$

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 $c: A \rightarrow B$ as:

$$
c(a)=\frac{1}{3}|b\rangle+\frac{2}{3}\left|b^{\perp}\right\rangle \quad \text { and } \quad c\left(a^{\perp}\right)=\tau
$$

We then still get an equation $\langle\mathrm{id}, c\rangle \gg=\omega_{1}=\omega$, no matter what $\tau$ 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.

Example 7.2.4. 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 [1, 18] - in particular in relation to Bell tables, see Exercise 2.4.15

Let $\omega \in \mathcal{D}(X \otimes Y)$ and $\rho \in \mathcal{D}(X \otimes Z)$ be two joint distributions with equal first marginal: $\omega[1,0]=\rho[1,0]$. A natural join, if it exists, is a distribution $\omega \bowtie \rho \in \mathcal{D}(X \otimes Y \otimes Z)$ which marginalises both to $\omega$ and two $\rho$, as in:

$$
(\omega \bowtie \rho)[1,1,0]=\omega \quad \text { and } \quad(\omega \bowtie \rho)[1,0,1]=\rho .
$$

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:

$$
c:=\omega[0,1 \mid 1,0]: X \rightarrow Y \quad d:=\rho[0,1 \mid 1,0]: X \leadsto Z .
$$

Now we define:


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 (7.11) 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 [132, 133].

Theorem 7.2.5. Let $X, Y$ be two sets, where $Y$ is finite.
1 There is a bijective correspondence between:

$$
\frac{\tau \in \mathcal{D}(X \times Y) \text { where } \tau[0,1] \text { has full support }}{\omega \omega \in \mathcal{D}(X) \text { and } c: X \hookrightarrow Y \text { such that } c \gg \omega \text { has full support }}
$$

2 For each natural number $N \geq 1$ there is a bijective correspondence between:

$$
\frac{\Omega \in \mathcal{D}(\mathcal{D}(X)) \text { with }|\operatorname{supp}(\Omega)|=N}{\rho \rho \mathcal{D}(X \times N) \text { where } \rho[0,1] \text { has full support }}
$$

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 $c:=\tau[0,1 \mid 1,0]: X \leadsto Y$ 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 $\omega, c$ to the joint state $\tau:=\langle i d, c\rangle\rangle=\omega \in$ $\mathcal{D}(X \times Y)$. By assumption, $\tau[0,1]=c \gg=\omega$ 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 $\operatorname{supp}(\Omega)=$ $\left\{\omega_{0}, \ldots, \omega_{N-1}\right\}$ with $\omega_{i} \in \mathcal{D}(X)$. We take:

$$
\begin{equation*}
\rho(x, i):=\Omega\left(\omega_{i}\right) \cdot \omega_{i}(x) . \tag{7.12}
\end{equation*}
$$

This yields a distribution $\rho \in \mathcal{D}(X \times N)$ since these probabilities add up to one:

$$
\sum_{x \in X, i \in N} \rho(x, i):=\sum_{i \in N} \Omega\left(\omega_{i}\right) \cdot \sum_{x \in X} \omega_{i}(x)=\sum_{i \in N} \Omega\left(\omega_{i}\right)=1 .
$$

Furthermore, $\rho$ 's second marginal has full support, since for each $i \in N$,

$$
\rho[0,1](i)=\sum_{x \in X} \rho(x, i)=\Omega\left(\omega_{i}\right) \cdot \sum_{x \in X} \omega_{i}(x)=\Omega\left(\omega_{i}\right)>0 .
$$

In the upward direction, starting from $\rho \in \mathcal{D}(X \times N)$ we from the channel $d:=\rho[1,0 \mid 0,1]: N \leadsto X$ by disintegration and use it to define $\Omega \in$ $\mathcal{D}(\mathcal{D}(X))$ as:

$$
\begin{equation*}
\Omega:=\sum_{i \in N} \rho[0,1](i)|d(i)\rangle . \tag{7.13}
\end{equation*}
$$

Since $\rho[0,1](i)>0$ for each $i \in N$, the support of $\Omega$ has $N$ elements.
Starting from $\Omega$ with support $\left\{\omega_{0}, \ldots, \omega_{N-1}\right\}$, we can get $\rho$ as in 7.12\} with extracted channel $d$ satisfying:

$$
d(i)(x)=\frac{\rho(x, i)}{\rho[0,1](i)}=\frac{\Omega\left(\omega_{i}\right) \cdot \omega_{i}(x)}{\Omega\left(\omega_{i}\right)}=\omega_{i}(x)
$$

Hence the definition (7.13) yields the original state $\Omega \in \mathcal{D}(\mathcal{D}(X))$ :

$$
\sum_{i \in N} \rho[0,1](i)|d(i)\rangle=\sum_{i \in N} \Omega\left(\omega_{i}\right)\left|\omega_{i}\right\rangle=\Omega
$$

In the other direction, starting from a joint state $\rho \in \mathcal{D}(X \times N)$ whose second marginal has full support, we can form $\Omega$ as in (7.13) and turn it into a joint state again via 7.12, whose probability at $(x, i)$ is:

$$
\begin{aligned}
\Omega(d(i)) \cdot d(i)(x) & =\rho[0,1](i) \cdot \rho[1,0 \mid 0,1](i)(x) \\
& =(\langle i d, \rho[1,0 \mid 0,1]\rangle \gg \rho[0,1])(x, i) \\
& =\rho(x, i)
\end{aligned}
$$

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 $\omega \in \mathcal{D}(X)$, $c: X \leadsto Y$, see Exercise 7.2.4 below. This corresponce is used in Bayesian persuasion, see [110], where the channel $c$ is called a signal.
With disintegration well-understood we can formulate a follow-up of crossover results in Theorem6.3.1 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 7.2.6. Let $c: X \mapsto Y$ be a channel with state $\omega \in \mathcal{D}(X)$ on its domain, and let $p \in \operatorname{Fact}(X)$ and $q \in \operatorname{Fact}(Y)$ be factors.

1 The extraction on an updated graph state yields:

$$
\left.(\langle i d, c\rangle \gg \omega)\right|_{p \otimes q}[0,1 \mid 1,0]=\left.c\right|_{q} \quad \text { where }\left.\quad c\right|_{q}(x):=\left.c(x)\right|_{q} .
$$

2 And as a result:

$$
\left.(\langle i d, c\rangle \gg \omega)\right|_{p \otimes q}=\left\langle i d,\left.c\right|_{q}\right\rangle \gg=\left.\omega\right|_{p \&(c \approx\langle q)} .
$$

Proof. 1 For $x \in X$ and $y \in Y$ we have:

$$
\begin{aligned}
\left.(\langle i d, c\rangle \gg \omega)\right|_{p \otimes q}[0,1 \mid 1,0](x)(y) & \stackrel{\text { 7.9. }}{-} \frac{\left.(\langle i d, c\rangle \gg=\omega)\right|_{p \otimes q}(x, y)}{\left.(\langle i d, c\rangle \gg \omega)\right|_{p \otimes q}[1,0](x)} \\
& =\frac{(\langle i d, c\rangle \gg=\omega)(x, y) \cdot(p \otimes q)(x, y)}{\sum_{v}(\langle i d, c\rangle \gg \omega)(x, v) \cdot(p \otimes q)(x, v)} \\
& =\frac{\omega(x) \cdot c(x)(y) \cdot p(x) \cdot q(y)}{\sum_{v} \omega(x) \cdot c(x)(v) \cdot p(x) \cdot q(v)} \\
& =\frac{c(x)(y) \cdot q(y)}{\sum_{v} c(x)(v) \cdot q(v)} \\
& =\frac{c(x)(y) \cdot q(y)}{c(x) \vDash q} \\
& =\left.c(x)\right|_{q}(y) \\
& =\left.c\right|_{q}(x)(y) .
\end{aligned}
$$

2 An (updated) joint state like $\left.(\langle i d, c\rangle \gg=\omega)\right|_{p \otimes q}$ can always be written as graph $\langle i d, d\rangle \gg \tau$. The channel $d$ is obtained by distintegration of the joint state, and equals $\left.c\right|_{q}$ by the previous point. The state $\tau$ is the first marginal of the joint state. Hence we are done by charactersing this first marginal, in:

$$
\begin{array}{ll}
\left.(\langle i d, c\rangle \gg=\omega)\right|_{p \otimes q}[1,0] & \\
=\left.(\langle i d, c\rangle \gg \omega)\right|_{(1 \otimes q) \&(p \otimes \mathbf{1}}[1,0] & \\
\text { by Exercise 4.3.8 } \\
=\left.\left.(\langle i d, c\rangle \gg \omega)\right|_{1 \otimes q}\right|_{p \otimes 1}[1,0] & \\
\text { by Lemmat6.1.6 (3) } \\
=\left.\left.(\langle i d, c\rangle \gg \omega)\right|_{1 \otimes q}[1,0]\right|_{p} & \\
=\left.\left.\omega\right|_{c=\langle q}\right|_{p} & \text { by Lemmat6.1.6 6) } \\
=\left.\omega\right|_{p \&(c=\alpha)} &
\end{array}
$$

Remark 7.2.7. In the preface to this book - to be precise, on page viii 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 $q: Y \rightarrow \mathbb{R}_{\geq 0}$ on the second component.

- The 'weaken-update-marginalise' approach first weakens $q$ to a factor $\mathbf{1} \otimes$ $q=\pi_{2} \approx \ll q$ 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 $\left.\omega\right|_{1_{\otimes q}} \in \mathcal{D}(X \times Y)$. Its first marginal $\left.\omega\right|_{1_{\otimes q}}[1,0] \in \mathcal{D}(X)$ then gives the effect in $X$ of updating in $Y$.
- The 'extract-infer' approach works differently: it first extracts a channel $c:=$ $\omega[0,1 \mid 1,0]: X \leadsto Y$ from $\omega$, via disintegration, such that $\omega=\langle i d, c\rangle \gg \sigma$, where $\sigma=\omega[1,0] \in \mathcal{D}(X)$ is $\omega$ 's first marginal. One then obtains the effect in $X$ of updating in $Y$ via backward inference, as $\left.\sigma\right|_{c \lll q} \in \mathcal{D}(X)$.

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 $X \mapsto Y$ extracted from $\omega$ by disintegration is the constant function $x \mapsto \tau$. Give a string diagrammatic account of this situation.
7.2.2 Disintegrate the distribution $\operatorname{Flrn}(\tau) \in \mathcal{D}(\{H, L\} \times\{1,2,3\})$ in Subsection 1.6.1 to a channel $\{H, L\} \rightarrow\{1,2,3\}$.
7.2.3 Consider sets $X=\left\{x_{1}, x_{2}\right\}, Y=\left\{y_{1}, y_{2}, y_{3}\right\}, Z=\left\{z_{1}, z_{2}, z_{3}\right\}$ with distribution $\omega \in \mathcal{D}(X \times Y)$ given by:

$$
\frac{1}{4}\left|x_{1}, y_{1}\right\rangle+\frac{1}{4}\left|x_{1}, y_{2}\right\rangle+\frac{1}{6}\left|x_{2}, y_{1}\right\rangle+\frac{1}{6}\left|x_{2}, y_{2}\right\rangle+\frac{1}{6}\left|x_{2}, y_{3}\right\rangle
$$

and $\rho \in \mathcal{D}(X \times Z)$ by:

$$
\frac{1}{12}\left|x_{1}, z_{1}\right\rangle+\frac{1}{3}\left|x_{1}, z_{2}\right\rangle+\frac{1}{12}\left|x_{1}, z_{3}\right\rangle+\frac{1}{8}\left|x_{2}, z_{1}\right\rangle+\frac{1}{8}\left|x_{2}, z_{2}\right\rangle+\frac{1}{4}\left|x_{2}, z_{3}\right\rangle .
$$

1 Check that $\omega$ and $\rho$ have equal $X$-marginals.
2 Explicitly describe the extracted channels $c:=\omega[0,1 \mid 1,0]: X \rightarrow$ $Y$ and $d:=\omega[0,1 \mid 1,0]: X \leadsto Z$.
3 Show that the natural join $\omega \bowtie \rho \in \mathcal{D}(X \times Y \times Z)$ according to 7.11) is:

$$
\begin{aligned}
& \frac{1}{24}\left|x_{1}, y_{1}, z_{1}\right\rangle+\frac{1}{6}\left|x_{1}, y_{1}, z_{2}\right\rangle+\frac{1}{24}\left|x_{1}, y_{1}, z_{3}\right\rangle \\
& \quad+\frac{1}{24}\left|x_{1}, y_{2}, z_{1}\right\rangle+\frac{1}{6}\left|x_{1}, y_{2}, z_{2}\right\rangle+\frac{1}{24}\left|x_{1}, y_{2}, z_{3}\right\rangle \\
& \quad+\frac{1}{24}\left|x_{2}, y_{1}, z_{1}\right\rangle+\frac{1}{24}\left|x_{2}, y_{1}, z_{2}\right\rangle+\frac{1}{12}\left|x_{2}, y_{1}, z_{3}\right\rangle \\
& \quad+\frac{1}{24}\left|x_{2}, y_{2}, z_{1}\right\rangle+\frac{1}{24}\left|x_{2}, y_{2}, z_{2}\right\rangle+\frac{1}{12}\left|x_{2}, y_{2}, z_{3}\right\rangle \\
& \quad+\frac{1}{24}\left|x_{2}, y_{3}, z_{1}\right\rangle+\frac{1}{24}\left|x_{2}, y_{3}, z_{2}\right\rangle+\frac{1}{12}\left|x_{2}, y_{3}, z_{3}\right\rangle .
\end{aligned}
$$

7.2.4 Combining the two items of Theorem 7.2.5 gives a bijective correspondence between:

$$
\frac{\Omega \in \mathcal{D}(\mathcal{D}(X)) \text { with }|\operatorname{supp}(\Omega)|=N}{\omega \in \mathcal{D}(X) \text { and } c: X \rightarrow N \text { such that } c \gg \omega \text { has full support }}
$$

Define this correspondence in detail and check that it is bijective.
7.2.5 Prove the following possibilitistic analogues of Theorem7.2.5

1 There is a bijective correspondence between:

$$
\frac{R \in \mathcal{P}(X \times Y)}{\overline{U \in \mathcal{P}(X) \text { with } f: U \rightarrow \mathcal{P}_{*}(Y)}}
$$

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:

$$
\frac{A \in \mathcal{P}(\mathcal{P}(X)) \text { with }|A|=N}{\overline{R \in \mathcal{P}(X \times N) \text { with } \forall i \neq j . \exists x . \neg(R(x, i) \Leftrightarrow R(x, j))}}
$$

7.2.6 Prove that the equation

$$
\left.(\langle i d, c\rangle \gg \sigma)\right|_{p \otimes 1}[0,1 \mid 1,0]=c .
$$

can be obtained both from Theorem6.3.4 and from Theorem7.2.6.
7.2.7 Prove that for a joint distribution $\omega \in \mathcal{D}(X \times Y)$ with full support one has:

$$
\omega[1,0 \mid 0,1]=(\omega[0,1 \mid 1,0])^{\dagger}: Y \leadsto X
$$

7.2.8 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 $p \in \operatorname{Pred}(X)$ and $q \in \operatorname{Pred}(Y)$ be arbitrary predicates.

Prove that there are predicates $q_{1}$ on $X$ and $p_{2}$ on $Y$ such that:

$$
\omega_{1} \vDash p \& q_{1}=\omega \vDash p \otimes q=\omega_{2} \vDash p_{1} \& q .
$$

This is a discrete version of [142, Prop. 6.7].
Hint: Disintegrate!

### 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 $Z \leadsto X \times Y$, and then more specifically to channels $X \leadsto Y_{1} \times \cdots \times Y_{n}$. 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 $P(Z \mid X, Y)$ such that

$$
P(X, Y \mid Z)=P(Y \mid X, Z) \cdot P(X \mid Z)
$$

This is captured diagrammatically in 7.14 below.
Definition 7.3.1. Consider a channel / box $f: Z \leadsto X \times Y$


We say that $f$ admits disintegration if there is a unique 'disintegration' channel $f^{\prime}: X \times Z \rightsquigarrow Y$ satisfying the equation:


Graphically, one can think of the disintegration $f^{\prime}$ as being obtained via bending the $X$-wire downward, as described below in (7.16).

Uniqueness of the channel $f^{\prime}$ in this definition means: for each channel $g: X \times Z \leadsto Y$ 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 $\bar{\mp}$ 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 $g=f^{\prime}$ in the conclusion expresses uniqueness of the disintegration box $f^{\prime}$.
We shall use the notation $f[0,1 \mid 1,0]$ for this disintegration box $f^{\prime}$. This notation will be explained in greater generality below.

The disintegration channel $f^{\prime}: X \times Z \leadsto Y$ is a pointwise version of the formulation for joint distributions on $X \times Y$ from (7.9):

$$
\begin{equation*}
f^{\prime}(x, z):=\sum_{y \in Y} \frac{f(z)(x, y)}{f[1,0](z)(x)}|y\rangle . \tag{7.15}
\end{equation*}
$$

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 7.2.3 (2) for an illustration. We first describe an illustration involving multinomial and hypergeometric channels from Chapter 3

## Example 7.3.2.

1 In Theorem 7.1.5 we have seen daggers of the hypergeometric channel, for four different prior distributions. Corrolary 7.2.2 tells that such daggers can be obtained via disintegration. We elaborate one case, with multinomials as prior.

We fix a finite set $X$ and numbers $L \geq K$. We define a channel $f: \mathcal{D}(X) \leadsto$ $\mathcal{N}[K] \times \mathcal{N}[L](X)$ of the form:

$$
\mathcal{D}(X) \xrightarrow{m n[L]} \mathcal{O}[L](X) \xrightarrow{\langle h g[K], i d\rangle} \mathcal{O} \mathcal{N}[K](X) \times \mathcal{N}[L](X)
$$

Notice that this channel's first marginal $f[1,0]$ is the multinomial channel $m n[K]$, by Corollary 3.4.2 (6). Where Corrolary 7.2.2 uses a joint state, we now use a channel, in line with the general approach to disintegration of this section.

We seek a disintegration $d: \mathcal{N}[K](X) \times \mathcal{D}(X) \rightsquigarrow \mathcal{N}[L](X)$ of the channel $f$. We define it via the dagger from Theorem 7.1.5)(2):

$$
d(\varphi, \omega):=h g[K]_{m n[L](\omega)}^{\dagger}(\varphi)=\sum_{\chi \in \mathcal{N}[L-K](X)} m n[L-K](\omega)(\chi)|\varphi+\chi\rangle .
$$

The disintegration equation 7.14 now holds, where the diagram on the right-hand-side is the above channel $f$.


We check this equation via the following calculation.

$$
\begin{aligned}
& \langle\text { id, } d(-, \omega)\rangle \gg m n[K](\omega) \\
& =\sum_{\varphi \in \mathcal{N}[K](X)} \sum_{\psi \in \mathcal{N}[L](X)} d(\varphi, \omega)(\psi) \cdot m n[K](\omega)(\varphi)|\varphi, \psi\rangle \\
& =\sum_{\psi \in \mathcal{N}[L](X)} \sum_{\varphi \leq K} m n[L-K](\omega)(\psi-\varphi) \cdot m n[K](\omega)(\varphi)|\varphi, \psi\rangle \\
& =\sum_{\psi \in \mathcal{N}[L](X)} \sum_{\varphi \leq \kappa \psi} \frac{(L-K)!}{(\psi-\varphi)]} \cdot \prod_{x} \omega(x)^{(\psi-\varphi)(x)} \cdot \frac{K!}{\varphi!} \cdot \prod_{x} \omega(x)^{\varphi(x)}|\varphi, \psi\rangle \\
& =\sum_{\psi \in \mathcal{N}[L](X)} \sum_{\varphi \leq K \psi} \frac{\binom{\psi}{\varphi}}{\binom{L}{K}} \cdot \frac{L!}{\psi!} \cdot \prod_{x} \omega(x)^{\psi(x)}|\varphi, \psi\rangle \quad \text { see Definition1.8.1 } \\
& =\sum_{\psi \in \mathcal{N}[L](X)} \sum_{\varphi \leq K \psi} h g[K](\psi)(\varphi) \cdot m n[L](\omega)(\psi)|\varphi, \psi\rangle \\
& =\langle h g[K], i d\rangle \gg m n[L](\omega) .
\end{aligned}
$$

2 In Diagram 3.21) in Theorem 3.3.8 one can recognise how the sum of numbers function sum: $\mathbb{N}^{m} \rightarrow \mathbb{N}$ 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 $m n[-](-) \odot(i d \otimes F l r n)$. It has a 'dependent type' $\prod_{K: \mathbb{N}} \mathcal{D}(\boldsymbol{m}) \rightsquigarrow \mathcal{N}[K](\boldsymbol{m})$.
3 Theorem 3.9.5, about Poisson multinomials, also expresses a disintegration situation, but we have to tweak it a little bit to make it fit the mold of Diagram (7.14). This works via additional copiers, as in 2.32, followed by
discarding. The first equation below is from Theorem 3.9.5


Next we elaborate on the notation that we will use for disintegration.
Remark 7.3.3. In traditional notation in probability theory one simply omits variables to express marginalisation. For instance, for a distribution $\omega \in \mathcal{D}\left(X_{1} \times\right.$ $\left.X_{2} \times X_{3} \times X_{4}\right)$, considered as function $\omega\left(x_{1}, x_{2}, x_{3}, x_{4}\right)$ in four variables $x_{i}$, one writes:

$$
\omega\left(x_{2}, x_{3}\right) \quad \text { for the marginal } \quad \sum_{x_{1}, x_{4}} \omega\left(x_{1}, x_{2}, x_{3}, x_{4}\right) .
$$

We have been using masks instead: lists containing as only elements 0 or 1 . We then write $\omega[0,1,1,0] \in \mathcal{D}\left(X_{2} \times X_{3}\right)$ to express the above marginalisation, with a 0 (resp. a 1 ) at position $i$ in the mask meaning that the $i$-th variable in the distribution is discarded (resp. kept).

We like to use a similar mask-style notation for disintegration, mimicking the traditional notation $\omega\left(x_{1}, x_{4} \mid x_{2}, x_{3}\right)$. This requires two masks, separated by the sign ' $\mid$ ' 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 $f[N \mid M]$ 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 $\vec{Y} \cap N$.
$6 f[N \mid M]$ is unique in satisfying an "obvious" adaptation of Equation (7.14), 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 $f[1,0,0,0,1 \mid 0,1,0,1,0]$, as on the right.


The disintegration box $f[1,0,0,0,1 \mid 0,1,0,1,0]$ is unique in satisfying:


In traditional notation one could express this equation as:

$$
f\left(Y_{1}, Y_{5} \mid Y_{2}, Y_{4}, X\right) \cdot f\left(Y_{2}, Y_{4} \mid X\right)=f\left(Y_{1}, Y_{2}, Y_{4}, Y_{5} \mid X\right) .
$$

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 $f[1,0,0,0,1 \mid 0,1,0,1,0]$ 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 7.15 . 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 box around the result in order to indicate that its interior is closed off and has become inaccessible. Thus:


This 'shaded box' notation can also be used for more complicated forms of disintegration, as described above. This notation is useful to express some basic properties of disintegration, see Exercise 7.3.5

Using such shading we can describe the dagger of a channel $c: X \leadsto Y$ as:


## Exercises

7.3.1 Consider an extracted channel $\omega[1,0,0 \mid 0,0,1]$ for some state $\omega$.

1 Write down the defining equation for this channel, as string diagram.
2 Check that $\omega[1,0,0 \mid 0,0,1]$ is the same as $\omega[1,0,1][1,0 \mid 0,1]$.
7.3.2 Check that a marginalisation $\omega M$ can also be described as disintegration $\omega[M \mid 0, \ldots, 0]$ 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 $f[0,0,0,1,1 \mid 0,1,1,0,0]$ in Remark 7.3.3 from the formulation in Definition 7.3.1 via rearranging and combining wires (via $\times$ ).
7.3.5 (From [57]) Prove the following 'sequential' and 'parallel' properties of disintegration. The best way is to give a diagrammatic proof, using uniqueness of disintegration.

7.3.6 Let $f: X \times Y \leadsto X \times Z$ be a channel, where $X, Y, Z$ are finite sets. Define the "probabilistic trace" $\operatorname{prtr}(f): Y \leadsto Z$ as:

where $\perp$ is used for the uniform distribution.

1 Check that:

$$
\operatorname{prtr}(f)(y)=\sum_{z \in Z}\left(\sum_{x \in X} \frac{1}{|X|} \cdot \frac{f(x, y)(x, z)}{\sum_{y^{\prime} \in Y} f(x, y)\left(x, y^{\prime}\right)}\right)|z\rangle
$$

A categorically oriented reader might now ask if the above definition $\operatorname{prtr}(f)$ yields a proper 'trace' construction in the symmetric monoidal Kleisli category 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 $f: X \mapsto 2 \times 2$ and $g: 2 \mapsto X$ defined by:

$$
\begin{array}{ll}
f(a)=\frac{1}{4}|0,0\rangle+\frac{3}{4}|1,0\rangle & g(0)=\frac{1}{3}|a\rangle+\frac{2}{3}|c\rangle \\
f(b)=\frac{2}{5}|0,0\rangle+\frac{3}{5}|1,1\rangle & g(1)=\frac{1}{2}|a\rangle+\frac{1}{2}|b\rangle . \\
f(c)=\frac{1}{2}|0,1\rangle+\frac{1}{2}|1,0\rangle &
\end{array}
$$

The aim is to show an inequality of states:


Both sides are an instance of 7.18 with $B=1$.
3 Check that $(g \otimes i d) \odot f: X \leadsto X \times 2$ is:

$$
\begin{aligned}
a & \mapsto \frac{11}{24}|a, 0\rangle+\frac{3}{8}|b, 0\rangle+\frac{1}{6}|c, 0\rangle \\
b & \mapsto \frac{2}{15}|a, 0\rangle+\frac{3}{10}|a, 1\rangle+\frac{3}{10}|b, 1\rangle+\frac{4}{15}|c, 0\rangle \\
c & \mapsto \frac{1}{4}|a, 0\rangle+\frac{1}{6}|a, 1\rangle+\frac{1}{4}|b, 0\rangle+\frac{1}{3}|c, 1\rangle .
\end{aligned}
$$

And that $f \odot g: 2 \hookrightarrow 2 \times 2$ is:

$$
\begin{aligned}
0 & \mapsto \frac{1}{12}|0,0\rangle+\frac{1}{3}|0,1\rangle+\frac{7}{12}|1,0\rangle \\
1 & \mapsto \frac{13}{40}|0,0\rangle+\frac{3}{8}|1,0\rangle+\frac{3}{10}|1,1\rangle .
\end{aligned}
$$

4 Now show that the disintegration $((g \otimes i d) \odot f)[0,1 \mid 1,0]: X \times X \leadsto 2$ is:
$(a, a) \mapsto 1|0\rangle$
$(b, a) \mapsto \frac{4}{13}|0\rangle+\frac{9}{13}|1\rangle$
$(c, a) \mapsto \frac{3}{5}|0\rangle+\frac{2}{5}|1\rangle$
$(a, b) \mapsto 1|0\rangle$
$(b, b) \mapsto 1|1\rangle$
$(c, b) \mapsto 1|0\rangle$
$(a, c) \mapsto 1|0\rangle$
$(b, c) \mapsto 1|0\rangle$
$(c, c) \mapsto 1|1\rangle$.

And that $(f \odot g)[0,1 \mid 1,0]: 2 \times 2 \leadsto 2$ is:

$$
\begin{aligned}
(0,0) & \mapsto \frac{1}{5}|0\rangle+\frac{4}{5}|1\rangle & (1,0) & \mapsto 1|0\rangle \\
(0,1) & \mapsto 1|0\rangle & (1,1) & \mapsto \frac{5}{9}|0\rangle+\frac{4}{9}|1\rangle
\end{aligned}
$$

5 Conclude that:

$$
\operatorname{prtr}((g \otimes i d) \odot f)=\frac{1}{3}|0\rangle+\frac{2}{3}|1\rangle,
$$

whereas:

$$
\operatorname{prtr}(f \odot g)=\frac{17}{45}|0\rangle+\frac{28}{45}|1\rangle .
$$

### 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 [102, §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:


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 [102] 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 |
| :---: | :---: | :---: | :---: |
| $\mathbf{1}$ | $?$ | $b$ | $u$ |
| $\mathbf{2}$ | $p$ | $b^{\perp}$ | $u$ |
| $\mathbf{3}$ | $p$ | $b$ | $?$ |
| $\mathbf{4}$ | $p$ | $b$ | $u^{\perp}$ |
| $\mathbf{5}$ | $?$ | $b^{\perp}$ | $?$ |

We wish to learn a distribution on $P$ and two channels $P \leadsto B$ and $P \leadsto U$ as interpretation of the Bayesian network (7.19). 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 (1.28), 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 7.20 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:

$$
q_{2}:=\mathbf{1}_{\left(p, b^{\perp}, u\right)}=\mathbf{1}_{p} \otimes \mathbf{1}_{b^{\perp}} \otimes \mathbf{1}_{u} \quad q_{4}:=\mathbf{1}_{\left(p, b, u^{\perp}\right)}=\mathbf{1}_{p} \otimes \mathbf{1}_{b} \otimes \mathbf{1}_{u^{\perp}}
$$

For cases $1,3,5$ we cannot use a point predicate where there is a missing item ?. There we use the truth predicate 1, in the parallel conjunction formulation $\otimes$. This indicates that all options are equally likely. Thus:

$$
q_{1}:=\mathbf{1} \otimes \mathbf{1}_{b} \otimes \mathbf{1}_{u} \quad q_{3}:=\mathbf{1}_{p} \otimes \mathbf{1}_{b} \otimes \mathbf{1} \quad q_{5}:=\mathbf{1} \otimes \mathbf{1}_{b^{\perp}} \otimes \mathbf{1} .
$$

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 updates, corresponding to five rows in Table 7.20

$$
\begin{aligned}
\tau:= & \frac{1}{5} \cdot \text { unif }_{q_{1}}+\frac{1}{5} \cdot \text { unif }_{q_{2}}+\frac{1}{5} \cdot \text { unif }_{q_{3}}+\frac{1}{5} \cdot \text { unif }_{q_{4}}+\frac{1}{5} \cdot \text { unif }_{q_{5}} \\
= & \frac{1}{10}|p, b, u\rangle+\frac{1}{10}\left|p^{\perp}, b, u\right\rangle+\frac{1}{5}\left|p, b^{\perp}, u\right\rangle+\frac{1}{10}|p, b, u\rangle+\frac{1}{10}\left|p, b, u^{\perp}\right\rangle \\
& \quad+\frac{1}{5}\left|p, b, u^{1}\right\rangle+\frac{1}{20}\left|p, b^{\perp}, u\right\rangle+\frac{1}{20}\left|p^{\perp}, b^{\perp}, u\right\rangle+\frac{1}{20}\left|p, b^{\perp}, u^{\perp}\right\rangle+\frac{1}{20}\left|p^{\perp}, b^{\perp}, u^{\perp}\right\rangle \\
= & \frac{1}{5}|p, b, u\rangle+\frac{3}{10}\left|p, b, u^{\perp}\right\rangle+\frac{1}{4}\left|p, b^{\perp}, u\right\rangle+\frac{1}{20}\left|p, b^{\perp}, u^{\perp}\right\rangle \\
& +\frac{1}{10}\left|p^{\perp}, b, u\right\rangle+\frac{1}{20}\left|p^{\perp}, b^{\perp}, u\right\rangle+\frac{1}{20}\left|p^{\perp}, b^{\perp}, u^{\perp}\right\rangle .
\end{aligned}
$$

The first marginal $\tau[1,0,0]=\frac{4}{5}|p\rangle+\frac{1}{5}\left|p^{\perp}\right\rangle \in \mathcal{D}(P)$ is the learned pregnancy
distribution. The two 'Blood test' and 'Urine Test' channels $P \leadsto B$ and $P \mapsto U$ in 7.19) can be extracted via disintegration.

- We get $c:=\tau[0,1,0 \mid 1,0,0]=\tau[1,1,0][0,1 \mid 1,0]: P \leadsto B$ with:

$$
\begin{aligned}
c(p)= & \frac{\tau(p, b, u)+\tau\left(p, b, u^{\perp}\right)}{\tau(p, b, u)+\tau\left(p, b, u^{\perp}\right)+\tau\left(p, b^{\perp}, u\right)+\tau\left(p, b^{\perp}, u^{\perp}\right)}|b\rangle \\
& \quad+\frac{\tau\left(p, b^{\perp}, u\right)+\tau\left(p, b^{\perp}, u^{\perp}\right)}{\tau(p, b, u)+\tau\left(p, b, u^{\perp}\right)+\tau\left(p, b^{\perp}, u\right)+\tau\left(p, b^{\perp}, u^{\perp}\right)}\left|b^{\perp}\right\rangle \\
= & \frac{1 / 5+3 / 10}{1 / 5+3 / 10+1 / 4+1 / 20}|b\rangle+\frac{1 / 4+1 / 20}{1 / 5+3 / 10+1 / 4+1 / 20}\left|b^{\perp}\right\rangle \\
= & \frac{5}{8}|b\rangle+\frac{3}{8}\left|b^{\perp}\right\rangle . \\
c\left(p^{\perp}\right)= & \cdots=\frac{1}{2}|b\rangle+\frac{1}{2}\left|b^{\perp}\right\rangle .
\end{aligned}
$$

We see that a pregnant cow has a slightly higher probability than a nonpregnant cow of getting a positive blood test.

- In the same way we get $d:=\tau[0,0,1 \mid 1,0,0]=\tau[1,0,1][0,1 \mid 1,0]: P \mapsto U$ with:

$$
d(p)=\frac{9}{16}|u\rangle+\frac{7}{16}\left|u^{\perp}\right\rangle \quad \text { and } \quad d\left(p^{\perp}\right)=\frac{3}{4}|u\rangle+\frac{1}{4}\left|u^{\perp}\right\rangle .
$$

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 [102], 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 [37, §17.3.1], where we do have prior information, namely in the form of a Bayesian network, involving four sets
$A=\left\{a, a^{\perp}\right\}, B=\left\{b, b^{\perp}\right\}, C=\left\{c, c^{\perp}\right\}$ and $D=\left\{d, d^{\perp}\right\}$, in the string diagram:


$$
\left\{\begin{align*}
\omega & =\frac{1}{5}|a\rangle+\frac{4}{5}\left|a^{\perp}\right\rangle  \tag{7.21}\\
f(a) & =\frac{3}{4}|b\rangle+\frac{1}{4}\left|b^{\perp}\right\rangle \\
f\left(a^{\perp}\right) & =\frac{1}{10}|b\rangle+\frac{9}{10}\left|b^{\perp}\right\rangle \\
h(b) & =\frac{1}{5}|d\rangle+\frac{4}{5}\left|d^{\perp}\right\rangle \\
h\left(b^{\perp}\right) & =\frac{7}{10}|d\rangle+\frac{3}{10}\left|d^{\perp}\right\rangle \\
g(a) & =\frac{1}{2}|c\rangle+\frac{1}{2}\left|c^{\perp}\right\rangle \\
g\left(a^{\perp}\right) & =\frac{1}{4}|c\rangle+\frac{3}{4}\left|c^{\perp}\right\rangle
\end{align*}\right.
$$

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 7.21).

$$
\begin{aligned}
& \tau:=((i d \otimes i d \otimes s w a p) \odot(i d \otimes i d \otimes h \otimes i d) \\
& \left.\quad \circ\left(i d \otimes \Delta_{2} \otimes i d\right) \odot(i d \otimes f \otimes g) \odot \Delta_{3}\right) \gg=\omega
\end{aligned}
$$

That is:


$$
\begin{aligned}
& \frac{3}{200}|a, b, c, d\rangle+\frac{3}{50}\left|a, b, c, d^{\perp}\right\rangle+ \\
& \frac{3}{200}\left|a, b, c^{\perp}, d\right\rangle+\frac{3}{50}\left|a, b, c^{\perp}, d^{\perp}\right\rangle+ \\
& \frac{7}{400}\left|a, b^{\perp}, c, d\right\rangle+\frac{3}{400}\left|a, b^{\perp}, c, d^{\perp}\right\rangle+ \\
& =\quad \frac{7}{400}\left|a, b^{\perp}, c^{\perp}, d\right\rangle+\frac{3}{400}\left|a, b^{\perp}, c^{\perp}, d^{\perp}\right\rangle+ \\
& \frac{1}{250}\left|a^{\perp}, b, c, d\right\rangle+\frac{2}{125}\left|a^{\perp}, b, c, d^{\perp}\right\rangle+ \\
& \frac{3}{250}\left|a^{\perp}, b, c^{\perp}, d\right\rangle+\frac{6}{125}\left|a^{\perp}, b, c^{\perp}, d^{\perp}\right\rangle+ \\
& \frac{63}{500}\left|a^{\perp}, b^{\perp}, c, d\right\rangle+\frac{27}{500}\left|a^{\perp}, b^{\perp}, c, d^{\perp}\right\rangle+ \\
& \frac{189}{500}\left|a^{\perp}, b^{\perp}, c^{\perp}, d\right\rangle+\frac{81}{500}\left|a^{\perp}, b^{\perp}, c^{\perp}, d^{\perp}\right\rangle .
\end{aligned}
$$

This distribution $\tau$ will be used as prior.
The table with data provided in [37] is:

| case | $A$ | $B$ | $C$ | $D$ |
| :---: | :---: | :---: | :---: | :---: |
| $\mathbf{1}$ | $?$ | $b$ | $c^{\perp}$ | $?$ |
| $\mathbf{2}$ | $?$ | $b$ | $?$ | $d^{\perp}$ |
| $\mathbf{3}$ | $?$ | $b^{\perp}$ | $c$ | $d$ |
| $\mathbf{4}$ | $?$ | $b^{\perp}$ | $c$ | $d$ |
| $\mathbf{5}$ | $?$ | $b$ | $?$ | $d^{\perp}$ |

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 Theorem6.3.1 we know that this inference can also be done via the joint state $\tau$, namely as:

$$
\tau_{1}:=\left.\tau\right|_{\mathbf{1} \otimes \mathbf{1}_{b} \otimes \mathbf{1}_{c} \perp \otimes \mathbf{1}} .
$$

The missing probabilities for $A, D$ are then obtained via the marginal, as:

$$
\begin{aligned}
& \tau_{1}[1,0,0,1] \\
& \quad=\frac{1}{9}|a, d\rangle+\frac{4}{9}\left|a, d^{\perp}\right\rangle+\frac{4}{45}\left|a^{\perp}, d\right\rangle+\frac{16}{45}\left|a^{\perp}, d^{\perp}\right\rangle \\
& =0.1111|a, d\rangle+0.4444\left|a, d^{\perp}\right\rangle+0.08889\left|a^{\perp}, d\right\rangle+0.3556\left|a^{\perp}, d^{\perp}\right\rangle
\end{aligned}
$$

This distribution appears in [37, Fig. 17.4 (a)].
In a similar way we define:

$$
\tau_{2}:=\left.\tau\right|_{\mathbf{1} \otimes \mathbf{1}_{b} \otimes \mathbf{1}_{\otimes 1} \mathbf{1}_{d^{\perp}}} \quad \tau_{3}:=\left.\tau\right|_{\mathbf{1} \otimes \mathbf{1}_{b^{\perp}} \otimes \mathbf{1}_{c} \otimes \mathbf{1}_{d}} \quad \tau_{4}:=\tau_{3} \quad \tau_{5}:=\tau_{2}
$$

These five states $\tau_{1}, \ldots, \tau_{5}$ are combined in a convex combination, like in the previous subsection, giving a new joint state:

$$
\tau^{\prime}:=\frac{1}{5} \cdot \tau_{1}+\frac{1}{5} \cdot \tau_{2}+\frac{1}{5} \cdot \tau_{3}+\frac{1}{5} \cdot \tau_{4}+\frac{1}{5} \cdot \tau_{5}=\frac{1}{5} \cdot \tau_{1}+\frac{2}{5} \cdot \tau_{2}+\frac{2}{5} \cdot \tau_{3}
$$

From $\tau^{\prime}$ we can obtain a newly learned interpretation of the Bayesian network (7.21), via marginalisation and disintegration. In detail, the new distribution $\omega^{\prime}$ on $A$ is:

$$
\begin{align*}
\omega^{\prime}:=\tau^{\prime}[1,0,0,0] & =\frac{3571}{8487}|a\rangle+\frac{4916}{8487}\left|a^{\perp}\right\rangle  \tag{7.23}\\
& \approx 0.4208|a\rangle+0.5792\left|a^{\perp}\right\rangle .
\end{align*}
$$

We get as newly learned channels:

$$
\begin{aligned}
& f^{\prime}:=\tau^{\prime}[0,1,0,0 \mid 1,0,0,0] \quad\left\{\begin{aligned}
f^{\prime}(a) & =\frac{3157}{3571}|b\rangle+\frac{414}{3571}\left|b^{\perp}\right\rangle \\
& \approx 0.8841|b\rangle+0.1159\left|b^{\perp}\right\rangle \\
f^{\prime}\left(a^{\perp}\right) & =\frac{2419}{6145}|b\rangle+\frac{3726}{6145}\left|b^{\perp}\right\rangle \\
& \approx 0.3937|b\rangle+0.6063\left|b^{\perp}\right\rangle
\end{aligned}\right. \\
& g^{\prime}:=\tau^{\prime}[0,0,1,0 \mid 1,0,0,0] \quad\left\{\begin{aligned}
g^{\prime}(a) & =\frac{1521}{3511}|c\rangle+\frac{2050}{3571}\left|c^{\perp}\right\rangle \\
& \approx 0.4259|c\rangle+0.5741\left|c^{\perp}\right\rangle \\
g^{\prime}\left(a^{\perp}\right) & =\frac{819}{1229}|c\rangle+\frac{410}{1229}\left|c^{\perp}\right\rangle
\end{aligned}\right. \\
& \begin{aligned}
g^{\prime}:=\tau[0,0,1,0 \mid 1,0,0,0]
\end{aligned} \quad \begin{aligned}
g^{\prime}\left(a^{\perp}\right) & =\frac{819}{1229}|c\rangle+\frac{410}{129}\left|c^{\perp}\right\rangle \\
& \approx 0.6664|c\rangle+0.3336\left|c^{\perp}\right\rangle \\
h^{\prime}:=\tau^{\prime}[0,0,0,1 \mid 0,1,0,0] \quad & =\frac{1}{15}|d\rangle+\frac{14}{15}\left|d^{\perp}\right\rangle \\
& \approx 0.06667|d\rangle+0.9333\left|d^{\perp}\right\rangle \\
h^{\prime}\left(b^{\perp}\right) & =1|d\rangle .
\end{aligned}
\end{aligned}
$$

These outcomes are as in [37, Fig. 17.5] - up to some small differences, probably due to rounding.

Interestingly, Table (7.22) gives no information about $A$, but still we learn a new distribution $\omega^{\prime}$ on $A$, basically from the prior information $\tau$.

## Exercises

7.4.1 Check that the pregnancy distribution in Subsection 7.4.1, obtained from Table (7.20), is $\frac{4}{5}|p\rangle+\frac{1}{5}\left|p^{\perp}\right\rangle$.
7.4.2 Calculate in detail that, in Subsection 7.4.2.

$$
\tau_{1}[1,0,0,1]=\frac{1}{9}|a, d\rangle+\frac{4}{9}\left|a, d^{\perp}\right\rangle+\frac{4}{45}\left|a^{\perp}, d\right\rangle+\frac{16}{45}\left|a^{\perp}, d^{\perp}\right\rangle
$$

7.4.3 Check that the new state $\omega^{\prime}$ in (7.23) can also be obtained as:

$$
\begin{aligned}
& \omega^{\prime}=\left.\frac{1}{5} \cdot \omega\right|_{\left(f=\ll \mathbf{1}_{b}\right) \&\left(g=\ll \mathbf{1}_{c^{\perp}}\right)} \\
& +\left.\frac{2}{5} \cdot \omega\right|_{\left(f=\ll \mathbf{1}_{b}\right) \&\left(h \text { <<< } \mathbf{1}_{d^{\perp}}\right)} \\
& \left.\left.+\left.\frac{2}{5} \cdot \omega\right|_{\left(f=\ll \mathbf{1}_{b} \perp\right.} \&\left(h=<\mathbf{1}_{c}\right)\right)\right) \&\left(g=<\mathbf{1}_{c}\right) .
\end{aligned}
$$

Argue why this works.

### 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 [181]; we follow the analysis of [24]. 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 [181] 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

| 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 $[181$ Table 1.2].
the table in Figure 7.1 We choose obvious abbreviations for the entries in the table:

$$
O=\{s, o, r\} \quad T=\{h, m, c\} \quad H=\{h, n\} . \quad W=\{t, f\} \quad P=\{y, n\}
$$

These sets are joined into a single product space:

$$
S:=O \times T \times H \times W \times P
$$

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:

$$
\tau=\frac{1}{14}|s, h, h, f, n\rangle+\frac{1}{14}|s, h, h, t, n\rangle+\cdots+\frac{1}{14}|r, m, h, t, n\rangle
$$

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:


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 $\tau$, that is, we 'factorise' $\tau$ according to this string diagram (7.24). Obviously, the play state $\pi \in \mathcal{D}(P)$ is obtained as the last marginal:

$$
\pi:=\tau[0,0,0,0,1]=\frac{9}{14}|y\rangle+\frac{5}{14}|n\rangle .
$$

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.

$$
\begin{align*}
c_{O} & :=\tau[1,0,0,0,0 \mid 0,0,0,0,1]: P \longrightarrow \\
c_{T} & :=\tau[0,1,0,0,0 \mid 0,0,0,0,1]: P \longrightarrow \\
c_{H} & :=\tau[0,0,1,0,0 \mid 0,0,0,0,1]: P \longrightarrow T  \tag{7.25}\\
c_{W} & :=\tau[0,0,0,1,0 \mid 0,0,0,0,1]: P \longrightarrow H
\end{align*}
$$

For instance the 'outlook' channel $c_{O}: P \leadsto O$ looks as follows.

$$
\begin{equation*}
c_{O}(y)=\frac{2}{9}|s\rangle+\frac{4}{9}|o\rangle+\frac{3}{9}|r\rangle \quad c_{O}(n)=\frac{3}{5}|s\rangle+\frac{2}{5}|r\rangle . \tag{7.26}
\end{equation*}
$$

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:

$$
P \xrightarrow{c:=\left\langle c_{0}, c_{T}, c_{H}, c_{W}\right\rangle} O \times T \times H \times W
$$

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:

$$
\begin{aligned}
\left.c_{\pi}^{\dagger}(s, c, h, t) \stackrel{\boxed{7.11}}{-} \pi\right|_{c=<1_{(s, c, h, t)}} & =\frac{125}{611}|y\rangle+\frac{486}{611}|n\rangle \\
& \approx 0.2046|y\rangle+0.7954|n\rangle .
\end{aligned}
$$

This corresponds to the probability $20.5 \%$ calculated in [181] - 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:

$$
O \times T \times H \times W \xrightarrow{c_{\pi}^{*}} P
$$

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:

$$
\left\langle c_{O}, c_{T}, c_{H}, c_{W}, i d\right\rangle \gg=\pi .
$$

This state differs considerably from the original table / state $\tau$. It shows that the shape (7.24) 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 $\omega \in \mathcal{D}\left(X_{1} \times \cdots \times X_{n} \times Y\right)$ 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 $c_{i}: Y \leadsto X_{i}$ via disintegration:

$$
\begin{aligned}
c_{1} & :=\omega[1,0, \ldots, 0 \mid 0, \ldots, 0,1]: Y \rightsquigarrow X_{1} \\
c_{2} & :=\omega[0,1,0 \ldots, 0 \mid 0, \ldots, 0,1]: Y \rightsquigarrow X_{2} \\
& \vdots \\
c_{n} & :=\omega[0, \ldots, 0,1,0 \mid 0, \ldots, 0,1]: Y \nrightarrow X_{n} .
\end{aligned}
$$

3 Form the tuple channel $c:=\left\langle c_{1}, \ldots, c_{n}\right\rangle: Y \leadsto X_{1} \times \cdots \times X_{n}$.
4 Take the Bayesian inversion (dagger channel) $c_{\pi}^{\dagger}: X_{1} \times \cdots \times X_{n} \rightsquigarrow Y$, as classifier. It gives for each $n$-tuple of input features $x_{1}, \ldots, x_{n}$ a distribution $c_{\pi}^{\dagger}\left(x_{1}, \ldots, x_{n}\right) \in \mathcal{D}(Y)$ 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:

$$
\pi=\frac{\overline{\bar{\chi}} \ldots \overline{\text { 구 }}}{}{ }^{Y}
$$



The classification channel $X_{1} \times \cdots \times X_{n} \leadsto Y$ is then obtained via the dagger

| 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 [154].
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 wrt. this set.

### 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 [154, 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..

$$
A=\{k, u\} \quad T=\{n, f\} \quad L=\{l, s\} \quad W=\{h, w\} \quad U=\{s, r\} .
$$

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:

$$
\begin{equation*}
\omega=\frac{1}{18}|k, n, l, h, s\rangle+\frac{1}{18}|u, n, s, w, r\rangle+\cdots+\frac{1}{18}|u, n, s, w, r\rangle . \tag{7.27}
\end{equation*}
$$

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 $\left(x_{1}, x_{2}, x_{3}, x_{4}\right) \in A \times T \times L \times W$ 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 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 $\sigma$.

1 Check if we are done, that is, if the $U$-marginal (for User action) of the current joint state $\sigma$ 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 $\sigma$, and write $X$ as a circle in the tree.
3 For each element of $x \in X$, update (and marginalise) $\sigma$ 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 $\omega$ in (7.27), corresponding to the table in Figure 7.2 .


Figure 7.3 Decision tree for reading (r) or skipping (s) derived from the data 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 $\omega$. This is where disintegration comes in. We first extract four channels and states:

$$
\begin{aligned}
c_{A} & :=\omega[0,0,0,0,1 \mid 1,0,0,0,0]: A \nrightarrow U \\
\omega_{A} & :=\omega[1,0,0,0,0] \in \mathcal{D}(A) \\
c_{T} & :=\omega[0,0,0,0,1 \mid 0,1,0,0,0]: T \rightsquigarrow U \\
\omega_{T} & :=\omega[0,1,0,0,0] \in \mathcal{D}(T) \\
c_{L} & :=\omega[0,0,0,0,1 \mid 0,0,1,0,0]: L \nrightarrow U \\
\omega_{L} & :=\omega[0,0,1,0,0] \in \mathcal{D}(L) \\
c_{W} & :=\omega[0,0,0,0,1 \mid 0,0,0,1,0]: W \rightsquigarrow U \\
\omega_{W} & :=\omega[0,0,0,1,0] \in \mathcal{D}(W) .
\end{aligned}
$$

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:

$$
\begin{array}{rlrl}
c_{A}(k) & =\frac{1}{2}|s\rangle+\frac{1}{2}|r\rangle & c_{A}(u) & =\frac{1}{2}|s\rangle+\frac{1}{2}|r\rangle \\
c_{T}(n) & =\frac{3}{10}|s\rangle+\frac{7}{10}|r\rangle & c_{T}(f) & =\frac{3}{4}|s\rangle+\frac{1}{4}|r\rangle \\
c_{L}(l) & =\frac{1}{3}|s\rangle & \omega_{T} & =\frac{5}{9}|n\rangle+\frac{1}{3}|u\rangle \\
c_{W}(h) & =\frac{1}{2}|s\rangle+\frac{1}{2}|r\rangle \\
c_{L}(s) & =\frac{2}{11}|s\rangle+\frac{9}{11}|r\rangle & \omega_{L} & =\frac{7}{18}|l\rangle+\frac{11}{18}|s\rangle \\
c_{W}(w) & =\frac{1}{2}|s\rangle+\frac{1}{2}|r\rangle & \omega_{W} & =\frac{4}{9}|h\rangle+\frac{5}{9}|w\rangle .
\end{array}
$$

In order to determine which of the input features $A, T, L, W$ is most dominant we compute the expected entropy - sometimes called intrinsice value - for each of these features. Recall the Shannon entropy function $H: \mathcal{D}(X) \rightarrow \mathbb{R}_{\geq 0}$ from Exercise 4.1.11 Post-composing it with the above channels gives a factor, like $H \circ c_{A}: A \rightarrow \mathbb{R}_{\geq 0}$. It computes the entropy $H\left(c_{A}(x)\right)$ for each $x \in A$. Hence we can compute its validity in the marginal state $\omega_{A} \in \mathcal{D}(A)$, giving the expected entropy. Explicitly:

$$
\begin{aligned}
\omega_{A} \models H \circ c_{A}= & \sum_{x \in A} \omega_{A}(x) \cdot H\left(c_{A}(x)\right) \\
= & \omega_{A}(k) \cdot H\left(c_{A}(k)\right)+\omega_{A}(u) \cdot H\left(c_{A}(u)\right) \\
= & \frac{2}{3} \cdot\left(\frac{1}{2} \cdot-\log \left(\frac{1}{2}\right)+\frac{1}{2} \cdot-\log \left(\frac{1}{2}\right)\right) \\
& +\frac{1}{3} \cdot\left(\frac{1}{2} \cdot-\log \left(\frac{1}{2}\right)+\frac{1}{2} \cdot-\log \left(\frac{1}{2}\right)\right) \\
= & \frac{2}{3}+\frac{1}{3} \\
= & 1 .
\end{aligned}
$$

In the same way one computes the other expected entropies as:

$$
\omega_{T} \vDash H \circ c_{T}=0.85 \quad \omega_{L} \vDash H \circ c_{L}=0.42 \quad \omega_{W} \vDash H \circ c_{W}=1
$$

One then picks the lowest entropy value, which is 0.42 , 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 $\omega$ 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 notation.

$$
\begin{aligned}
& \omega / l:=\left.\omega\right|_{1 \otimes 1 \otimes \mathbf{1}_{l} \otimes \otimes 1}[1,1,0,1,1] \in \mathcal{D}(A \times T \times W \times U) \\
& \omega / s:=\left.\omega\right|_{1 \otimes 1 \otimes \mathbf{1}_{s} \otimes 1 \otimes \mathbf{1}}[1,1,0,1,1] \in \mathcal{D}(A \times T \times W \times U) .
\end{aligned}
$$

We now go into a recursive loop and repeat the previous steps for both these states $\omega / l$ and $\omega / s$. Notice that they are 'shorter' than $\omega$, 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 $\omega / s$ is not a point state.
2.2.2 We will now have to determine the dominant feature in $\omega / s$. We compute the three expected entropies for $A, T, W$ as:

$$
\begin{aligned}
& \omega / s[1,0,0,0] \vDash H \circ \omega / s[0,0,0,1 \mid 1,0,0,0]=0.44 \\
& \omega / s[0,1,0,0] \vDash H \circ \omega / s[0,0,0,1 \mid 0,1,0,0]=0.36 \\
& \omega / s[0,0,1,0] \vDash H \circ \omega / s[0,0,0,1 \mid 0,0,1,0]=0.68 .
\end{aligned}
$$

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:

$$
\begin{aligned}
& \omega / s / n:=\omega /\left.s\right|_{1 \otimes 1_{n} \otimes 1 \otimes 1}[1,0,1,1] \in \mathcal{D}(A \times W \times U) \\
& \omega / s / f:=\omega /\left.s\right|_{1 \otimes 1_{f} \otimes 1 \otimes 1}[1,0,1,1] \in \mathcal{D}(A \times W \times U) .
\end{aligned}
$$

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:

$$
\omega / \operatorname{s} / n[0,0,1]=1|r\rangle .
$$

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 expected entropies again, in order to find out which of the remaining input features $A, T$ is dominant.

$$
\begin{aligned}
& \omega / s / f[1,0,0] \vDash H \circ \omega / s / f[0,0,1 \mid 1,0,0]=0 \\
& \omega / s / f[0,1,0] \vDash H \circ \omega / s / f[0,0,1 \mid 0,1,0]=1 .
\end{aligned}
$$

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 $\omega / s / f$.

$$
\begin{aligned}
& \omega / s / f / k:=\omega / s /\left.f\right|_{\mathbf{1}_{k} \otimes 1 \otimes 1}[0,1,1] \in \mathcal{D}(W \times U) \\
& \omega / s / f / u:=\omega / s /\left.f\right|_{\mathbf{1}_{u} \otimes 1 \otimes 1}[0,1,1] \in \mathcal{D}(W \times U)
\end{aligned}
$$

The next cycle continues with these two states.
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 [102], 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 [156], which uses expected entropies (intrinsic values). Sometimes it is described in terms of 'gains', see [135], where in the above step 1.2 we can define for feature $X \in\{A, T, L, W\}$,

$$
\operatorname{gain}(X):=H\left(\omega_{U}\right)-\left(\omega_{X} \vDash H \circ c_{X}\right) .
$$

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 [181, 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 $\sigma[1,0,0,0,1] \in \mathcal{M}(O \times P)$.


Figure 7.4 Decision tree for playing (y) or not (n) derived from the data in Figure 7.1

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 $P \leadsto O$ from this table, and compare it to the description 7.26) of the channel $c_{O}$ given in Subsection 7.5.1see also Lemma 2.2.4
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 $\sigma$, notice that instead of taking the dagger of the tuple of channels $\left\langle c_{O}, c_{T}, c_{H}, c_{W}\right\rangle: P \rightarrow O \times$ $T \times H \times W$ in Example 7.5.1 we could have used instead the 'direct' disintegration:

$$
O \times T \times H \times W \xrightarrow{\sigma[0,0,0,0,1 \mid 1,1,1,1,0]} P
$$

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:
review, send, us, your, password, account : S $\quad \mathbf{2}$.
Write $c: S \rightarrow \mathbf{2}^{6}=\mathbf{2} \times \mathbf{2} \times \mathbf{2} \times \mathbf{2} \times \mathbf{2} \times \mathbf{2}$ for the tuple channel:
$c:=\langle$ review, send, us, your, password, account〉.
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:

$$
\begin{aligned}
c=\ll \mathbf{1}_{(1,0,1,0,0,0)}= & \left(\text { review }=\ll \mathbf{1}_{1}\right) \&\left(\text { send }=\ll \mathbf{1}_{0}\right) \\
& \&\left(\text { us }=\ll \mathbf{1}_{1}\right) \&\left(\text { your }=\ll \mathbf{1}_{0}\right) \\
& \&\left(\text { password }=\ll \mathbf{1}_{0}\right) \&\left(\text { account }=\ll \mathbf{1}_{0}\right) \\
= & \frac{9}{2048} \cdot \mathbf{1}_{s}+\frac{1}{16} \cdot \mathbf{1}_{h}
\end{aligned}
$$

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:

$$
\begin{aligned}
c_{\omega}^{\dagger}(1,0,1,0,0,0)=\left.\omega\right|_{c \lll \mathbf{1}_{(1,0,1,0,0,0)}} & =\frac{9}{73}|s\rangle+\frac{64}{73}|h\rangle \\
& \approx 0.1233|s\rangle+0.8767|h\rangle .
\end{aligned}
$$

7.5.4 Here is another example in naive Bayesian classification that is commonly used as illustration 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 $\left\{l, l^{\perp}\right\},\left\{s, s^{\perp}\right\}$ and $\left\{y, y^{\perp}\right\}$. The training data is summarised in the following table.

[^10]|  | long | not <br> long | sweet | not <br> sweet | yellow | not <br> 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:

$$
\{b, o, n\} \rightarrow 0 \rightarrow\left\{l, l^{\perp}\right\} \quad\{b, o, n\} \rightarrow\left\{s, s^{\perp}\right\} \quad\{b, o, n\} \rightarrow\left\{y, y^{\perp}\right\},
$$

and then combine them into a single channel

$$
\{b, o, n\} \longrightarrow \longrightarrow\left\{l, l^{\perp}\right\} \times\left\{s, s^{\perp}\right\} \times\left\{y, y^{\perp}\right\} .
$$

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 ( $\left.s, c, h, t\right)$. 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 [154]. There the following two questions are asked: what can be said about reading / skipping for the two queries:

$$
\begin{aligned}
& Q_{1}=(\text { unknown }, \text { new, long, work }) \\
& Q_{2}=(\text { unknown }, \text { follow-up, short, home }) .
\end{aligned}
$$

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.
(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 Sufficient statistics

It sometimes happens in an update $\left.c(x)\right|_{q}$, for a channel $c$, that the input $x$ drops out, in the sense that the updated distribution $\left.c(x)\right|_{q}$ 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 [53]. 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 [57]. Various examples are given below, based partly on [88, 83]. 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 $[K]: \mathcal{D}(X) \rightsquigarrow X^{K}$, 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):

$$
\begin{array}{rlr}
\left.\operatorname{iid}[K](\omega)\right|_{\operatorname{acc} \approx<1_{\varphi}} & =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{\omega^{K}(\vec{x})}{\mathcal{D}(\operatorname{acc})\left(\omega^{K}\right)(\varphi)}|\vec{x}\rangle & \text { by Lemma[7.1.3] 2] } \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{\omega^{K}(\vec{x})}{m n[K](\omega)(\varphi)}|\vec{x}\rangle & \text { by Theorem 2.6.7 } \\
& =\sum_{\vec{x} \in \operatorname{acc}^{-1}(\varphi)} \frac{\omega^{K}(\vec{x})}{(\varphi) \cdot \prod_{x \in X} \omega(x)^{\varphi(x)}}|\vec{x}\rangle & \\
& =\sum_{\overrightarrow{\vec{x} \in \operatorname{acc}^{-1}(\varphi)}} \frac{1}{(\varphi)}|\vec{x}\rangle & \text { since } \operatorname{acc}(\vec{x})=\varphi \\
\frac{\operatorname{arr}(\varphi) .}{} \quad
\end{array}
$$

In the latter expression the distribution $\omega$ has disappeared. This depends on the fact that sequences $\vec{x}, \vec{y} \in X^{K}$ with $\operatorname{acc}(\vec{x})=\varphi=\operatorname{acc}(\vec{y})$ 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 $\varphi$ : if we condition $\operatorname{iid}[K](\omega)=\omega^{K}$ to those sequences that accumulate to $\varphi$, the outcome no longer depends on $\omega$.

The axiomatisation of this phenomenon uses (surjective) functions $s: X \rightarrow$ $Y$, 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 $s: X \rightarrow Y$ to a channel $Y \leadsto X$. 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 [57] Defn. 14.3]. Subsequently we explain the relation to disintegration.

Definition 7.6.1. Let $c: A \rightarrow X$ be a channel, where we think of $A$ as the space of parameters.

1 A statistic for the channel $c$ is a function $s: X \rightarrow Y$.
2 Such a statistic $s$ is sufficient if there is a channel $d: Y \leadsto X$ such that:


This means that $d=s_{c(a)}^{\dagger}$, for each $a \in A$.
We shall soon look at examples, and non-examples too, but the shape on the right-hand-side of 7.28 may look familiar already since it occurred at several earlier stages in this book.

## Remark 7.6.2.

1 The channel $d$ on the right-hand-side in (7.28 results in a special way from disintegration, namely disintegration of the composite channel $\langle s, i d\rangle \odot c$ on the left-hand-side in 7.28 ). If we write $D: Y \times A \mapsto X$ for its disintegration channel, as in 7.14, 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.


The situation on the left can be expressed by the equation on the right.
2 The fact that the dashed wire is missing in (7.29) 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 7.29 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 4.2.3). Explicitly, via Lemma 7.1.3 (2), since the statistic $s$ is a function, not a channel:

$$
\begin{equation*}
D(a, y)=s_{c(a)}^{\dagger}(y)=\left.c(a)\right|_{s \ll \mathbf{1}_{y}}=\sum_{x \in s^{-1}(y)} \frac{c(a)(x)}{\mathcal{D}(s)(c(a))(y)}|x\rangle . \tag{7.30}
\end{equation*}
$$

The absence of the dashed arrow in 7.29 corresponds to the non-dependence of the latter expressions on $a$. This is the essence of the Fisher-Neyman factorisation theorem, see [57, Thm. 14.5] and [14, Prop 4.10] or [172, §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 7.30 and check if the parameter $a$ drops out.

## Examples 7.6.3.

1 We have already seen several sufficient statistics situations:

- The bitsum function sum: $\mathbf{2}^{K} \rightarrow\{0, \ldots, K\}$ is a sufficient statistic for the $K$-ary tuple channel $\langle$ flip,$\ldots$, flip $\rangle:[0,1] \rightsquigarrow \mathbf{2}^{K}$, see Theorem 2.5.1 A related situation occurs in Exercise 7.6.1 below.
- The function size: $\mathcal{N}(X) \rightarrow \mathbb{N}$ is a sufficient statistic for the multiset Poisson channel mpois from Example 2.1.7 (2), see Exercise 4.3.3.
- The accumulation function acc: $X^{K} \rightarrow \mathcal{N}[K](X)$ is a sufficient statistic for the $K$-fold tensor channel $\operatorname{iid}[K]: \mathcal{D}(X) \rightsquigarrow X^{K}$, see Theorem 3.3.1. Accumulation is also a sufficient statistic for the Poisson-iid channel, see Theorem 3.9.6.
- The sum of two multisets sum: $\mathcal{N}[K](X) \times \mathcal{N}[L](X) \rightarrow \mathcal{N}[K+L](X)$ is a sufficient statistic for the tuple channel $\langle m n[K], m n[L]\rangle: \mathcal{D}(X) \rightarrow$ $\mathcal{N}[K](X) \times \mathcal{N}[L](X)$ of two multinomials, see Theorem 3.4.4
- Decoupling dcpl $=\left\langle\mathcal{N}\left(\pi_{1}\right), \mathcal{N}\left(\pi_{2}\right)\right\rangle: \mathcal{N}[K](X \times Y) \rightarrow \mathcal{N}[K](X) \times \mathcal{N}[K](Y)$ is a sufficient statistic for the channel $m n[K] \circ \otimes: \mathcal{D}(X) \times \mathcal{D}(Y) \leadsto$ $\mathcal{N}[K](X \times Y)$, see Theorem 3.3.4
2 As listed above, accumulation acc: $X^{K} \rightarrow \mathcal{N}[K](X)$ is a sufficient statistic for iid $[K]$. What about matching mat: $X^{K} \rightarrow S P(K)$, see Definition 1.5.8. It turns out not to be sufficient. We compute the relevant dagger 7.30, for a distribution $\omega \in \mathcal{D}(X)$ and $P \in S P(K)$,

$$
\begin{aligned}
\operatorname{mat}_{\text {jid }[K](\omega)}^{\dagger}(P) & =\sum_{\vec{x} \in \operatorname{mat}^{-1}(P)} \frac{\omega^{K}(\vec{x})}{\mathcal{D}(\operatorname{mat})\left(\omega^{K}\right)(P)}|\vec{x}\rangle \\
& =\sum_{\vec{x} \in \operatorname{mat}^{-1}(P)} \frac{\omega^{K}(\vec{x})}{\sum_{\vec{y} \in \operatorname{mat}(P)} \omega^{K}(\vec{y})}|\vec{x}\rangle
\end{aligned}
$$

What we would now need is a result saying: $\operatorname{mat}(\vec{x})=\operatorname{mat}(\vec{y})$ 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 $(a, b)=$ $\{\{1\},\{2\}\}=\operatorname{mat}(a, c)$, but $\omega^{2}(a, b)=\frac{1}{2} \cdot \frac{1}{3} \neq \frac{1}{2} \cdot \frac{1}{6}=\omega^{2}(a, c)$.

We include a formulation of the Fisher-Neyman factorisation theorem (see e.g. [14, Prop 4.10] or [172, §3.3]) in diagrammatic form in [57, Thm. 14.5] and in [88, Lem. 3.4]. It uses that $\mathcal{K} \ell(\mathcal{D})$ is a positive Markov category, justifying the last equation below, see [57, Rem.11.23].
 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 $f_{\omega}^{\dagger} \odot f$, for a function $f$, see Lemma 7.1.3 (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 2.5.3.

The following result can be seen as a special case of Proposition 7.6.4, involving the probabilistic inverse $f^{\sim 1}$ from Definition 2.4.6. given by $f^{\sim 1}(y)=$
$\sum_{x \in f^{-1}(y)} \frac{1}{\left|f^{-1}(y)\right|}|x\rangle$. We recall that we write $\operatorname{ker}(f)=\left\{\left(x, x^{\prime}\right) \mid f(x)=f\left(x^{\prime}\right)\right\}$ for the kernel relation, for a function $f$.

Proposition 7.6.5. Let $c: A \rightarrow X$ be a channel and $f: X \rightarrow Y$ a surjective function with finite inverse images, satisfying $\operatorname{ker}(f) \subseteq \operatorname{ker}(c(a))$, for each $a \in A$, then $f$ is a sufficient statistic, with its probabilistic inverse $f^{\sim 1}: Y \leadsto X$ as associated channel, in a situation:


Proof. For each $a \in A$ we have:

$$
\begin{aligned}
& \langle f, \text { id }\rangle \gg=c(a) \\
& =\sum_{x \in X} \sum_{y \in Y}\langle f\rangle(x)(y) \cdot c(a)(x)|y, x\rangle \\
& =\sum_{y \in Y} \sum_{x \in f^{-1}(y)} c(a)(x)|y, x\rangle \\
& =\sum_{y \in Y} \sum_{x \in f^{-1}(y)} \sum_{z \in f^{-1}(y)} \frac{c(a)(z)}{\left|f^{-1}(y)\right|}|y, x\rangle \quad \text { since } \operatorname{ker}(f) \subseteq \operatorname{ker}(c(a)) \\
& =\sum_{x \in X} \sum_{y \in Y} f^{\sim 1}(y)(x) \cdot \mathcal{D}(f)(c(a))(y)|y, x\rangle \\
& =\left\langle i d, f^{\sim 1}\right\rangle \gg(f \odot c)(a) .
\end{aligned}
$$

Theorem 3.3.1 is an instance of this result. We present one more, involving the size count function sc: $S P(K) \rightarrow M P(K)$ from Definition 1.9.2 (2). It records the size of blocks in a partition $P$ as a multiset partition, namely as $s c(P)=\sum_{B \in P} 1| | B| \rangle$.

## Proposition 7.6.6.

1 For a distribution $\omega \in \mathcal{D}(X)$,

$$
\operatorname{ker}(s c) \subseteq \operatorname{ker}(c d[K](\omega))
$$

This means, for all set partitions $P, Q \in S P(K)$,

$$
\operatorname{sc}(P)=\operatorname{sc}(Q) \Longrightarrow c d[K](\omega)(P)=c d[K](\omega)(Q)
$$

where $c d[K](\omega)=\mathcal{D}(\operatorname{mat})\left(\omega^{K}\right)$ is the coincidence distribution from Definition 2.3.7(1).

2 This size count function sc: $S P(K) \rightarrow M P(K)$ is a sufficient statistic for the coincidence channel $c d[K]: \mathcal{D}(X) \mapsto S P(K)$, in a situation:


By Lemma 1.9.7 the probabilistic inverse channel sc ${ }^{\sim 1}: M P(K) \rightsquigarrow S P(K)$ on the right-hand-side can be described on $\alpha \in M P(K)$ as:

$$
\operatorname{sc}^{\sim 1}(\alpha)=\sum_{P \in \mathrm{sc}^{-1}(\alpha)} \frac{\alpha \rrbracket}{(\alpha)_{p}}|P\rangle .
$$

Proof. By Proposition 7.6 .5 it suffices to do the first item. We give an exemplaric proof, for $K=4$. Consider $P, Q \in S P(4)$ of the form:

$$
P=\{\{1\},\{2,4\},\{3\}\} \quad \text { and } \quad Q=\{\{1,3\},\{2\},\{4\}\} .
$$

Clearly they have the same size count: two blocks of size 1 , and one block of size 2. Each sequence of elements $s=\left\langle x_{1}, x_{2}, x_{3}, x_{4}\right\rangle$ that matches $P$ must have $x_{2}=x_{4}$. Then it can be easily transposed into a sequence $s^{\prime}=\left\langle x_{2}, x_{4}, x_{1}, x_{3}\right\rangle$ that matches $Q$, so that $\omega^{4}(s)=\omega^{4}\left(s^{\prime}\right)$. Thus, in general, when $\operatorname{sc}(P)=\operatorname{sc}(Q)$,

$$
c d[K](\omega)(P)=\sum_{\vec{x} \in \operatorname{mat}^{-1}(P)} \omega^{K}(\vec{x})=\sum_{\vec{x} \in \operatorname{mat}^{-1}(Q)} \omega^{K}(\vec{x})=c d[K](\omega)(Q) .
$$

## Exercises

7.6.1 Prove that the $K$-fold sum sum: $\mathbb{N}^{K} \rightarrow \mathbb{N}$ is a sufficient statistic for the $K$-tuple of Poisson channels $\langle$ pois, $\ldots$, pois $\rangle: \mathbb{R}_{>0} \leadsto \mathbb{N}^{K}$ in a situation:


Define yourself what this channel $d: \mathbb{N} \leadsto \mathbb{N}^{K}$ does.
Hint: Use Exercise 1.7.7
7.6.2 Check that Proposition 7.6.6 (1) does not hold for multiplicity count mc , see Definition 1.9.2 (1), and the multinomial distribution, that is, in general:

$$
m c(\varphi)=m c(\psi) \nRightarrow m n[K](\omega)(\varphi)=m n[K](\omega)(\psi) .
$$

### 7.7 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 $c: X \leadsto Y$, 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, where the evidence is not given by a predicate on the codomain of the channel, but by distribution. This alternative was introduced by Jeffrey [101], see also [21, 37, 43, 176], or [19] for a recent application in physics. Jeffrey's rule can be formulated conveniently (see [79]) 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 7.7 .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 [157, 56, 70, 26].

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 7.7.1. Let $c: X \leadsto Y$ be channel with a prior state $\sigma \in \mathcal{D}(X)$.

1 Pearl's rule is backward inference: it involves updating of the prior $\sigma$ with evidence in the form of a factor $q \in \operatorname{Fact}(Y)$ to the posterior:

$$
\left.\sigma\right|_{c \approx<q} \in \mathcal{D}(X) .
$$

2 Jeffrey's rule involves update of the prior $\sigma$ with evidence in form of a distribution $\tau \in \mathcal{D}(Y)$ to the posterior:

$$
c_{\sigma}^{\dagger} \gg \tau \in \mathcal{D}(X) .
$$

We illustrate the difference between Pearl's and Jeffreys's rules in a standard medical test situation.

Example 7.7.2. Consider the situation from Exercise 6.2.1 with set $D=\left\{d, d^{\perp}\right\}$ for disease (or not) and $T=\{p, n\}$ for a positive or negative test outcome, with a test channel $c: D \mapsto T$ given by:

$$
c(d)=\frac{9}{10}|p\rangle+\frac{1}{10}|n\rangle \quad c\left(d^{\perp}\right)=\frac{1}{20}|p\rangle+\frac{19}{20}|n\rangle .
$$

The test thus has a sensitivity of $90 \%$ and a specificity of $95 \%$. We assume a prevalence of $10 \%$ via a prior state $\omega=\frac{1}{10}|d\rangle+\frac{9}{10}\left|d^{\perp}\right\rangle$.

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:

$$
\left.\omega\right|_{c \approx<q}=\frac{74}{281}|d\rangle+\frac{207}{281}\left|d^{\perp}\right\rangle \approx 0.26|d\rangle+0.73\left|d^{\perp}\right\rangle .
$$

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:

$$
\begin{aligned}
c_{\sigma}^{\dagger} \gg \tau & =\frac{4}{5} \cdot c_{\sigma}^{\dagger}(p)+\frac{1}{5} \cdot c_{\sigma}^{\dagger}(n) \\
& \left.\stackrel{7.1 \mid}{=} \frac{4}{5} \cdot \omega\right|_{c=<\mathbf{1}_{p}}+\left.\frac{1}{5} \cdot \omega\right|_{c=<\mathbf{1}_{n}} \\
& =\frac{4}{5} \cdot\left(\frac{2}{3}|d\rangle+\frac{1}{3}\left|d^{\perp}\right\rangle\right)+\frac{4}{5} \cdot\left(\frac{2}{173}|d\rangle+\frac{171}{173}\left|d^{\perp}\right\rangle\right) \\
& =\frac{278}{519}|d\rangle+\frac{241}{519}\left|d^{\perp}\right\rangle \\
& \approx 0.54|d\rangle+0.46\left|d^{\perp}\right\rangle .
\end{aligned}
$$

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 likelihoods, in order to produce an advice about medical treatment?

The following theorem (from [81]) captures the essences of the update rules of Pearl and Jeffrey: validity increase or divergence decrease. For convenience we (again) repeat the Pearl case.

Theorem 7.7.3. Consider the situation in Definition 7.7.1 with a prior state $\sigma \in \mathcal{D}(X)$ and a channel $c: X \leadsto Y$.

1 For a factor $q$ on Y Pearl's update rule gives an increase of validity:

$$
c \gg \sigma_{P} \vDash q \geq c \gg \sigma \vDash q \quad \text { for the updated state } \quad \sigma_{P}:=\left.\sigma\right|_{c \approx<q} .
$$

2 For an evidence distribution $\tau \in \mathcal{D}(Y)$, Jeffrey's update rule gives a decrease of Kullback-Leibler divergence:

$$
D_{K L}\left(\tau, c \gg=\sigma_{J}\right) \leq D_{K L}(\tau, c \gg=\sigma) \quad \text { for } \quad \sigma_{J}:=c_{\sigma}^{\dagger} \gg=\tau .
$$

We assume that the predicted state $c \gg \sigma$ has fulll support, so that the dagger $c_{\sigma}^{\dagger}$ is well-defined.
In this sitation $c \gg=\sigma$ is the predicted distribution, which is used to evaluate the evidence.

- In Pearl's case we look at the validity $c \gg \sigma \vDash q$ of the evidence $q$ in this predicted state. The above theorem tells that by switching to the Pearlupdate $\sigma_{P}$ that we get a higher validity $c \gg \sigma_{P} \vDash q$ of the evidence, in the newly predicted state $c \gg \sigma_{P}$.
- In Jeffrey's case we look at the divergence / mismatch $D_{K L}(\tau, c \gg=\sigma)$ between the evidence distribution $\tau$ and the prediction $c \gg \sigma$. By changing to the Jeffrey-update $\sigma_{J}$ we get a lower divergence $D_{K L}\left(\tau, c \gg \sigma_{J}\right)$. Thus, via Jeffrey's rule one reduces 'prediction errors', in the terminology of predictive coding theory.

Proof. Item (1) is a repetition of Theorem 6.2.2. A proof of item (2) is given in [81] using rather heavy machinery (the Gelfand spectral radius theorem). We give a more elementary proof in the next chapter, in ??.

We review an earlier illustration, now with Jeffrey's approach emphasising the decrease of divergence between the predication and the evidence distribution.

Example 7.7.4. 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 $\tau$ on the set $G$ of grades:

$$
\tau=\frac{1}{10}|1\rangle+\frac{3}{10}|2\rangle+\frac{3}{10}|3\rangle+\frac{2}{10}|4\rangle+\frac{1}{10}|5\rangle .
$$

There is an a priori divergence $D_{K L}(\tau, c »>\sigma) \approx 1.336$. With some effort one can prove that the Jeffrey-update of $\sigma$ is:

$$
\begin{aligned}
\sigma^{\prime}=c_{\sigma}^{\dagger} \gg=\tau & =\frac{972795}{3913520}|p\rangle+\frac{1966737}{3913520}|n\rangle+\frac{973988}{3913520}|o\rangle \\
& \approx 0.2486|p\rangle+0.5025|n\rangle+0.2489|o\rangle .
\end{aligned}
$$

The divergence has now dropped to: $D_{K L}\left(\tau, c \gg \sigma^{\prime}\right) \approx 1.087$.
In the end it is interesting to compare the orinal (prior) mood with its Pearland 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.


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 7.7.3 (2), supports the view that Jeffrey's update rule should be used in situations where one is confronted with 'surprises' [45] or with 'unanticipated knowledge' [44] that one has to adapt to, as a form of correction. Here we include an example from [45] (also used in [79]) that involves such error correction after a 'surprise'.

Example 7.7.5. 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 experience with many earlier candidates, is a joint distribution on the product space $C \times E$, for $C=\left\{c, c^{\perp}\right\}$ and $E=\left\{e, e^{\perp}\right\}$, given as:

$$
\omega=\frac{4}{10}|c, e\rangle+\frac{1}{10}\left|c, e^{\perp}\right\rangle+\frac{1}{10}\left|c^{\perp}, e\right\rangle+\frac{4}{10}\left|c^{\perp}, e^{\perp}\right\rangle .
$$

The first marginal of $\omega$ is the uniform distribution $\frac{1}{2}|c\rangle+\frac{1}{2}\left|c^{\perp}\right\rangle$. It is the neutral base rate for Bob's competence.

We use the two projection functions $C \stackrel{\pi_{1}}{\longleftrightarrow} C \times E \xrightarrow{\pi_{2}} E$ 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 $\left.\omega\right|_{\pi_{2}=<\mathbf{1}_{e}}=\left.\omega\right|_{1 \otimes 1_{e}}$, whose first marginal is $\frac{4}{5}|c\rangle+\frac{1}{5}\left|c^{\perp}\right\rangle$. 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 [45]:

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 $\rho=\frac{1}{8}|c\rangle+\frac{7}{8}\left|c^{\perp}\right\rangle$. Ann wants to adapt to this new surprising situation, so she uses Jeffrey's rule, giving a new joint state:

$$
\omega^{\prime}=\left(\pi_{1}\right)_{\omega}^{\dagger} \gg=\rho=\frac{1}{10}|c, e\rangle+\frac{1}{40}\left|c, e^{\perp}\right\rangle+\frac{7}{40}\left|c^{\perp}, e\right\rangle+\frac{7}{10}\left|c^{\perp}, e^{\perp}\right\rangle
$$

If the letter now tells that Bob has work experience, Ann will factor this in, in this new situation $\omega^{\prime}$, giving, like above, via Pearl's rule followed by marginalisation,

$$
\left.\omega^{\prime}\right|_{\pi_{2}=\left\langle\mathbf{1}_{e}\right.}[1,0]=\frac{4}{11}|c\rangle+\frac{7}{11}\left|c^{\perp}\right\rangle
$$

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 [45] in channelbased form, with the associated formulations of Pearl's and Jeffrey's rules from Definition 7.7.1 and produces exactly the same outcomes as in loc. cit.

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 7.7.6. Let $c: X \leadsto Y$ 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:

$$
\left.\sigma\right|_{c: \ll \mathbf{1}_{y}}=c_{\sigma}^{\dagger}(y)=c_{\sigma}^{\dagger} \gg=1|y\rangle .
$$

2 Pearl's updating with a constant predicate (no information) does not change the prior state $\omega$ :

$$
\left.\sigma\right|_{c:<r \cdot \mathbf{1}}=\sigma, \quad \text { for } r>0
$$

Jeffrey's update does not change anything when we update with what we already predict:

$$
c_{\sigma}^{\dagger} \gg=(c \gg=\sigma)=\sigma .
$$

3 For a factor $q \in \operatorname{Fact}(Y)$ we can both update the prior distribution $\sigma$, according to Pearl, and also the channel c, so that the updated predicted state is predicted:

$$
\left.c\right|_{q} \gg=\left(\left.\sigma\right|_{c \lll q}\right)=\left.(c \gg \sigma)\right|_{q} .
$$

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 $\tau$ is predicted:

$$
\left(c_{\sigma}^{\dagger}\right)_{\tau}^{\dagger} \gg\left(c_{\sigma}^{\dagger} \gg=\tau\right)=\tau .
$$

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 7.7.1.
2 The first equation follows from Lemma 6.1.6(1) and (4); the second one is Equation 7.3
3 The first claim is Theorem 6.3.4 and the second one is Lemma 7.1.8, (4).
4 By Lemma 6.1.6(3) we have:

$$
\left.\left.\sigma\right|_{c=\ll q_{1}}\right|_{c=\ll q_{2}}=\left.\sigma\right|_{\left(c=\ll q_{1}\right) \&\left(c=\ll q_{2}\right)}=\left.\left.\sigma\right|_{c=\ll q_{2}}\right|_{c=\ll q_{1}} .
$$

However, in general, for evidence states $\tau_{1}, \tau_{2} \in \mathcal{D}(Y)$,

$$
c_{c_{\sigma}^{\dagger} \gg \tau_{1}}^{\dagger} \gg \tau_{2} \neq c_{c_{\sigma}^{\dagger} \gg \tau_{2}}^{\dagger} \gg=\tau_{1} .
$$

Exercise 7.7.5 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. [65, 70]. 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.

There are translations back-and-forth between Pearl's and Jeffrey's rules, due to [21]; they are translated here to the current setting.

Proposition 7.7.7. Let $c: X \leadsto Y$ be a channel with a prior state $\sigma \in \mathcal{D}(X)$ on its domain, such that the predicted state $c \gg=\sigma$ has full support.

1 Pearl's updating can be expressed as Jeffrey's update, by turning a factor $q: Y \rightarrow \mathbb{R}_{\geq 0}$ into a state $\left.(c \gg \sigma)\right|_{q} \in \mathcal{D}(Y)$, so that it can be used as evidence in:

$$
c_{\sigma}^{\dagger} \gg=\left(\left.(c \gg=\sigma)\right|_{q}\right)=\left.\sigma\right|_{c=\ll q} .
$$

2 Jeffrey's updating can also be expressed as Pearl's updating: for an evidence state $\tau \in \mathcal{D}(Y)$ write $\tau /(c \gg \sigma)$ for the factor $y \mapsto \frac{\tau(y)}{(c \gg \sigma)(y)}$; then:

$$
c_{\sigma}^{\dagger} \gg=\tau=\left.\sigma\right|_{c=\langle\tau /(\omega ; \sigma \sigma} .
$$

Proof. The first item is exactly Theorem 7.1.7 11. For the second item we first note that the fraction $\tau /(c \gg \sigma)$ is a well-defined factor, since $c \gg \sigma$ has full support. Further, its validity $(c \gg \sigma) \vDash \tau /(c \gg \sigma)$ is one, since $\tau$ is a distribution:

$$
\begin{equation*}
(c \gg \sigma) \models \tau /(c \gg \sigma)=\sum_{y \in Y}(c \gg=\sigma)(y) \cdot \frac{\tau(y)}{(c \gg=\sigma)(y)}=\sum_{y \in Y} \tau(y)=1 . \tag{7.31}
\end{equation*}
$$

But then, for $x \in X$,

$$
\begin{aligned}
& \left(c_{\sigma}^{\dagger} \gg \tau\right)(x)=\sum_{y \in Y} \tau(y) \cdot c_{\sigma}^{\dagger}(y)(x) \\
& \text { [7.1) } \sum_{y \in Y} \tau(y) \cdot \frac{\sigma(x) \cdot c(x)(y)}{(c \gg \sigma)(y)} \\
& =\sigma(x) \cdot \sum_{y \in Y} c(x)(y) \cdot \frac{\tau(y)}{(c \gg \sigma)(y)} \\
& =\frac{\sigma(x) \cdot(c=\langle\tau /(c \gg \sigma))(x)}{c \gg \sigma \models \tau /(c \gg \sigma)} \quad \text { as just shown } \\
& =\frac{\sigma(x) \cdot(c=\langle\tau /(c \gg \sigma))(x)}{\sigma \vDash c=\langle/(c \gg \sigma)} \\
& =\left.\sigma\right|_{c \approx \ll /(c=\sigma)}(x) \text {. }
\end{aligned}
$$

Combination of Theorem 7.7.3 and Proposition 7.7.7 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 7.7.8. Let $c: X \leadsto Y$ be a channel with a prior state $\sigma \in \mathcal{D}(X)$ and a factor $q$ on $Y$. Then:

$$
D_{K L}\left(\left.(c \gg=\sigma)\right|_{q}, c \gg=\left(\left.\sigma\right|_{c=<q}\right)\right) \leq D_{K L}\left(\left.(c \gg=\sigma)\right|_{q}, c \gg=\sigma\right) .
$$

This result can be interpreted as follows. With my prior state $\sigma$ I can predict $c \gg \sigma$. I can update this prediction with evidence $q$ to $\left.(c \gg=\sigma)\right|_{q}$. The divergence between this update and my prediction is bigger than the divergence between the update and the prediction from the Pearl / Bayes posterior $\left.\sigma\right|_{c=\ll}$. Thus, the posterior gives a correction.

Proof. Take $\tau=\left.(c \gg=\sigma)\right|_{q}$. Theorem 7.7.3(2) says that:

$$
D_{K L}\left(\tau, c \gg=\left(c_{\sigma}^{\dagger} \gg \tau\right)\right) \leq D_{K L}(\tau, c \gg=\sigma) .
$$

But Proposition 7.7.7 (1) tells that $c_{\sigma}^{\dagger} \gg=\tau=\left.\sigma\right|_{c=<q}$.
Lemma 7.7.9. Consider a channel $c: X \rightarrow Y$ 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.

$$
\begin{aligned}
& D_{K L}\left(c_{\omega}^{\dagger} \gg=\rho, \omega\right) \leq D_{K L}(\rho, c \gg=\omega) \\
& D_{K L}\left(\omega, c_{\omega}^{\dagger} \gg \rho\right) \leq D_{K L}(c \gg=\omega, \rho) .
\end{aligned}
$$

Proof. Both inequalities are obtained via Equation 7.3 and Proposition 2.8.4 (2):

$$
\begin{aligned}
D_{K L}\left(c_{\omega}^{\dagger} \gg=\rho, \omega\right) & =D_{K L}\left(c_{\omega}^{\dagger} \gg=\rho, c_{\omega}^{\dagger} \gg(c \gg=\omega)\right) \\
& \leq D_{K L}(\rho, c \gg=\omega) \\
D_{K L}\left(\omega, c_{\omega}^{\dagger} \gg \rho\right) & =D_{K L}\left(c_{\omega}^{\dagger} \gg=(c \gg=\omega), c_{\omega}^{\dagger} \gg=\rho\right) \\
& \leq D_{K L}(c \gg=\omega, \rho) .
\end{aligned}
$$

## Remarks 7.7.10.

1 In Corrolary 7.7.8 we use that Pearl's rule can be expressed as Jeffrey's, see point (1) in Proposition 7.7.7. 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 $c: X \mapsto Y$ be a channel with states $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Y)$. Assuming that $c \gg \sigma$ has full support we can form the factor $q=\tau /(c \gg \sigma)$. We then get an inequality:

$$
c \gg=\left(c_{\sigma}^{\dagger} \gg=\tau\right) \vDash q=c \gg=\left(\left.\sigma\right|_{c \approx<q}\right) \vDash q \geq c \gg=\sigma \vDash q \stackrel{7.31}{=} 1 .
$$

This is not very useful.
2 We have emphasised the message "Pearl increases validity" and "Jeffrey decreases divergence". But Corollary 7.7 .8 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 $c: X \leadsto Y$ given by:

$$
c(0)=\frac{1}{9}|a\rangle+\frac{2}{3}|b\rangle+\frac{2}{9}|c\rangle \quad \text { and } \quad c(1)=\frac{7}{25}|a\rangle+\frac{7}{25}|b\rangle+\frac{11}{25}|c\rangle .
$$

The predicted state is then $c \gg \sigma=\frac{44}{225}|a\rangle+\frac{71}{150}|b\rangle+\frac{149}{450}|c\rangle$. We use as 'equal' evidence predicate and state:

$$
q=\frac{1}{2} \cdot \mathbf{1}_{a}+\frac{1}{3} \cdot \mathbf{1}_{b}+\frac{1}{6} \cdot \mathbf{1}_{c} \quad \text { and } \quad \tau=\frac{1}{2}|a\rangle+\frac{1}{3}|b\rangle+\frac{1}{6}|c\rangle .
$$

We then get the following updates, according to Pearl and Jeffrey, respectively:

$$
\begin{aligned}
\sigma_{P} & =\left.\sigma\right|_{c \approx<q} & \sigma_{J} & =c_{\sigma}^{\dagger} \gg \tau \\
& =\frac{425}{839}|0\rangle+\frac{414}{839}|1\rangle & & \text { and }
\end{aligned} \quad=\frac{805675}{1861904}|0\rangle+\frac{1056229}{1861904}|1\rangle .
$$

The validities are summarised in the following table.

| description | formula | value |
| :---: | :---: | :---: |
| prior validity | $c »>\sigma \vDash q$ | 0.31074 |
| after Pearl | $c \gg=\sigma_{P} \vDash q$ | 0.31079 |
| after Jeffrey | $c »>\sigma_{J} \vDash q$ | 0.31019 |

The differences are small, but relevant. Pearl's updating increases validity, as Theorem 7.7.3 1 dictates, but Jeffrey's updating does not, in this example.

The divergences in this example are as follows.

| description | formula | value |
| :---: | :---: | :---: |
| prior divergence | $D_{K L}(\tau, c \gg \sigma)$ | 0.238 |
| after Pearl | $D_{K L}\left(\tau, c \gg \sigma_{P}\right)$ | 0.240 |
| after Jeffrey | $D_{K L}\left(\tau, c \gg \sigma_{J}\right)$ | 0.221 |

Jeffrey's rule decreases divergence, in line with Theorem 7.7.3(2), but Pearl's updating does not. Is there a contradiction with Corollary 7.7.8, which does involve a divergence decrease? No, since there the state $\tau$ has a very particular shape, namely $\left.(c \gg=\sigma)\right|_{q}$. We conclude that for that particular state Pearl's rule gives a divergence decrease, but there is no such decrease in general.

Thus, we conclude that, in general, validity increase works exclusively for Pearl's rule, and divergence decrease works exclusively for Jeffrey's rule.

## Exercises

7.7.1 Check for yourself the claimed outcomes in Example 7.7.5

```
\(\left.\omega\right|_{\pi_{2}=<\mathbf{1}_{e}}[1,0]=\frac{4}{5}|c\rangle+\frac{1}{5}\left|c^{\perp}\right\rangle ;\)
\(\left.\omega^{\prime}:=\left(\pi_{1}\right)_{\omega}^{\dagger}\right\rangle=\rho=\frac{1}{10}|c, e\rangle+\frac{1}{40}\left|c, e^{\perp}\right\rangle+\frac{7}{40}\left|c^{\perp}, e\right\rangle+\frac{7}{10}\left|c^{\perp}, e^{\perp}\right\rangle ;\)
\(\left.\omega^{\prime}\right|_{\pi_{2}=《 \mathbf{1}_{e}}[1,0]=\frac{4}{11}|c\rangle+\frac{7}{11}\left|c^{\perp}\right\rangle\).
```

7.7.2 This example is taken from [43], 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 $f: X \rightarrow \mathbf{2}$.
2 Use the uniform distribution unif $=\frac{1}{2}|1\rangle+\frac{1}{2}|0\rangle$ on 2 and show that Jeffrey's rule gives as adapted distribution:

$$
f_{\omega}^{\dagger} \gg=\text { unif }=\left.\frac{1}{2} \cdot \omega\right|_{\left.\mathbf{1}_{A A}\right]}+\left.\frac{1}{2} \cdot \omega\right|_{\mathbf{1}_{[B, C)}}=\frac{1}{2}|A\rangle+\frac{2}{9}|B\rangle+\frac{5}{18}|C\rangle .
$$

7.7.3 The next illustration is attributed to Jeffrey, and is reproduced for instance in [21, 37]. 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 $S=\left\{s, s^{\perp}\right\}$. There is a prior joint distribution $\tau$ on $C \times S$, namely:
$\omega=\frac{3}{25}|g, s\rangle+\frac{9}{50}\left|g, s^{\perp}\right\rangle+\frac{3}{25}|b, s\rangle+\frac{9}{50}\left|b, s^{\perp}\right\rangle+\frac{8}{25}|v, s\rangle+\frac{2}{25}\left|v, s^{\perp}\right\rangle$.
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 $\omega$ in two ways as:

$$
\langle c, i d\rangle \gg=\omega_{2}=\omega=\langle i d, d\rangle \gg=\omega_{1}
$$

for channels $c:=\omega[1,0 \mid 0,1]: S \leadsto C$ and $d:=\omega[0,1 \mid 1,0]: C \rightarrow$ $S$, given by:

$$
\left\{\begin{array} { r } 
{ c ( s ) = \frac { 3 } { 1 1 } | b \rangle + \frac { 3 } { 1 } | b \rangle + \frac { 8 } { 1 4 } | v \rangle } \\
{ c ( s ^ { \perp } ) = \frac { 9 } { 2 2 } | b \rangle + \frac { 9 } { 2 2 } | b \rangle + \frac { 2 } { 1 1 } | v \rangle }
\end{array} \quad \left\{\begin{array}{l}
d(g)=\frac{2}{5}|s\rangle+\frac{3}{5}\left|s^{\perp}\right\rangle \\
d(b)=\frac{2}{5}|s\rangle+\frac{3}{5}\left|s^{\perp}\right\rangle \\
d(v)=\frac{4}{5}|s\rangle+\frac{1}{5}\left|s^{\perp}\right\rangle .
\end{array}\right.\right.
$$

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:

$$
\left.\omega_{2}\right|_{c \approx<q}=\left(\left.\omega\right|_{q \otimes 1}\right)[0,1]=d \gg=\left(\left.\omega_{1}\right|_{q}\right)=\frac{26}{61}|s\rangle+\frac{35}{61}\left|s^{\perp}\right\rangle .
$$

3 Describe the evidence now as a state $\tau$ on $C$ and show that Jeffrey's rule gives:

$$
d \gg=\tau=\left(\left(\pi_{1}\right)_{\omega}^{\dagger} \gg=\tau\right)[0,1]=\frac{21}{50}|s\rangle+\frac{29}{50}\left|s^{\perp}\right\rangle .
$$

4 Check also that:

$$
\begin{aligned}
\langle i d, c\rangle \gg=\tau=\left(\pi_{1}\right)_{\omega}^{\dagger} \gg \tau= & \left.\frac{7}{25}|g, s\rangle+\frac{21}{50}\left|g, s^{\perp}\right\rangle\right)+\frac{1}{10}|b, s\rangle \\
& \left.+\frac{3}{20}\left|b, s^{\perp}\right\rangle\right)+\frac{1}{25}|v, s\rangle+\frac{1}{100}\left|v, s^{\perp}\right\rangle .
\end{aligned}
$$

The latter outcome is given in [37].
7.7.4 The following alarm example is in essence due to Pearl [145], see Example 6.1.2 (2) and also [37, §3.6]; we have adapted the numbers in order to make the calculations a little bit easier. There is an 'alarm' set $A=\left\{a, a^{\perp}\right\}$ and a 'burglary' set $B=\left\{b, b^{\perp}\right\}$, with the following a priori joint distribution on $A \times B$.

$$
\omega=\frac{1}{200}|a, b\rangle+\frac{7}{500}\left|a, b^{\perp}\right\rangle+\frac{1}{1000}\left|a^{\perp}, b\right\rangle+\frac{98}{100}\left|a^{\perp}, b^{\perp}\right\rangle .
$$

Someone reports that the alarm went off, but with only $80 \%$ certainty because of deafness.

1 Translate the alarm information into a predicate $p: A \rightarrow[0,1]$ and show that crossover updating leads to a burglary distribution:

$$
\begin{aligned}
\left.\omega\right|_{p \otimes 1}[0,1] & =\frac{3}{151}|b\rangle+\frac{148}{151}\left|b^{\perp}\right\rangle \\
& \approx 0.02|b\rangle+0.98\left|b^{\perp}\right\rangle .
\end{aligned}
$$

2 Compute the extracted channel $c:=\omega[1,0 \mid 0,1]: B \leftrightarrows A$ as in Theorem7.2.1, 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 $d:=c_{\omega[0,1]}^{\dagger}: A \nrightarrow B$ of this channel $c$ to calculate the outcome of Jeffrey's update rule as:

$$
\begin{aligned}
d \gg\left(\frac{4}{5}|a\rangle+\frac{1}{5}\left|a^{\perp}\right\rangle\right) & =\frac{19639}{93195}|b\rangle+\frac{73556}{93195}\left|b^{\perp}\right\rangle \\
& \approx 0.21|b\rangle+0.79\left|b^{\perp}\right\rangle .
\end{aligned}
$$

(Notice again the considerable difference in outcomes between Pearl and Jeffrey.)
7.7.5 Recall from Example 7.7.2 the Jeffrey update:

$$
\sigma_{1}:=c_{\sigma}^{\dagger} \gg=\rho=\frac{278}{519}|d\rangle+\frac{278}{519}\left|d^{\perp}\right\rangle .
$$

Suppose we have another evidence state $\rho=\frac{3}{5}|p\rangle+\frac{2}{5}|n\rangle$. Compute:

$$
\begin{aligned}
& c_{\sigma_{1}}^{\dagger} \gg \rho ; \\
& \sigma_{2}:=c_{\sigma}^{\dagger} \gg \rho ; \\
& c_{\sigma_{2}}^{\dagger} \gg \tau .
\end{aligned}
$$

The first and third items produce different results, which proves Proposition 7.7.6 (4).
7.7.6 Consider a channel $c: X \rightarrow Y$ with prior $\sigma \in \mathcal{D}(X)$ and evidence distribution $\tau \in \mathcal{D}(Y)$. Assume that $\tau$ is a point state $1|z\rangle$, for some $z \in Y$. Write $\omega:=\langle i d, c\rangle \gg=\sigma \in \mathcal{D}(X \times Y)$.
1 Show that the Jeffrey-updated joint state $\omega^{\prime}:=\left(\pi_{2}\right)_{\omega}^{\dagger} \gg=\tau \in \mathcal{D}(X \times$ $Y$ ) is of the form:

$$
\omega^{\prime}=\left.\sigma\right|_{c \approx\left\langle\mathbf{1}_{z}\right.} \otimes 1|z\rangle .
$$

2 Verify that the double-dagger channel $c^{\prime}:=\left(c_{\sigma}^{\dagger}\right)_{\tau}^{\dagger}: X \leadsto Y$ is extremely trivial, namely a 'constant-point' channel, that is, of the form $c^{\prime}(x)=1|z\rangle$ for all $x \in X$.
7.7.7 Jeffrey's rule is frequently formulated (notably in [66], to which we refer for details) and used (like in Exercise 7.7.2), in situations where the channel involved is deterministic, see Lemma 7.1.3 Consider an arbitrary function $f: X \rightarrow I$, 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:

$$
f_{\omega}^{\dagger} \gg=\rho=\left.\sum_{i \in I} \rho(i) \cdot \omega\right|_{\mathbf{1}_{U_{i}}} \quad \text { satisfying } \quad f \gg=\left(f_{\omega}^{\dagger} \gg=\rho\right)=\rho .
$$

2 Prove the following minimal-distance result:

$$
d\left(f_{\omega}^{\dagger} \gg \rho, \omega\right)=\bigwedge\left\{d\left(\omega, \omega^{\prime}\right) \mid \omega^{\prime} \in \mathcal{D}(X) \text { with } f \gg=\omega^{\prime}=\rho\right\},
$$

where $d$ is the total variation distance from Section 4.5
7.7.8 Prove that for a general, not-deterministic channel $c: X \leadsto Y$ with prior state $\omega \in \mathcal{D}(X)$ and state $\rho \in \mathcal{D}(Y)$, there is an inequality:

$$
d\left(c_{\omega}^{\dagger} \gg \rho, \omega\right) \leq \bigwedge_{\omega^{\prime} \in \mathcal{D}(X)} d\left(\omega, \omega^{\prime}\right)+d\left(c \gg=\omega^{\prime}, \rho\right)
$$

### 7.8 Factorisation of joint states

Earlier, in Subsection 2.3.2, we have called a binary joint distribution nonentwined 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. For instance, the state:

$$
\omega=\frac{1}{4}|a, b\rangle+\frac{1}{2}\left|a, b^{\perp}\right\rangle+\frac{1}{12}\left|a^{\perp}, b\right\rangle+\frac{1}{6}\left|a^{\perp}, b^{\perp}\right\rangle
$$

is non-entwined: it is the product of its marginals $\omega[1,0]=\frac{3}{4}|a\rangle+\frac{1}{4}\left|a^{\perp}\right\rangle$ and $\omega[0,1]=\frac{1}{3}|b\rangle+\frac{2}{3}\left|b^{\perp}\right\rangle$. We will formulate this as:

$$
\omega \text { has shape }
$$

$\square$
or as:
$\omega$ factorises as $\qquad$ and write this as: $\quad \omega \approx \curvearrowleft \downarrow$.

We shall give a formal definition of $\approx$ below, but at this stage it suffices to read $\sigma \approx S$, for a distribution $\sigma$ and a string diagram $S$, as: there is an interpretation of the boxes in $S$ that produces $\sigma$.

In the above case of $\omega \approx \square \square$ we obtain $\omega=\omega_{1} \otimes \omega_{2}$, for some state $\omega_{1}$ that interpretes the box on the left, and some $\omega_{2}$ interpreting the box on the right. But then:

$$
\omega[1,0]=\left(\omega_{1} \otimes \omega_{2}\right)[1,0]=\omega_{1}
$$

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 $\omega$.
We conclude that non-entwinedness of an arbitrary binary joint state $\omega$ can be expressed as: $\omega \approx \square \square$. 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 more expressive.

In general there may be several interpretations of a string diagram (as shape). Consider for instance the statement:

$$
\frac{14}{25}|H\rangle+\frac{11}{25}|T\rangle \approx \square
$$

The distribution on the left-hand-side matches the shape on the right-hand-side in multiple ways, for instance as:

$$
c_{1} \gg=\sigma_{1}=\frac{14}{25}|H\rangle+\frac{11}{25}|T\rangle=c_{2} \gg=\sigma_{2}
$$

for:

$$
\begin{aligned}
\sigma_{1} & =\frac{1}{5}|1\rangle+\frac{4}{5}|0\rangle & \sigma_{2} & =\frac{2}{5}|1\rangle+\frac{3}{5}|0\rangle \\
c_{1}(1) & =\frac{4}{5}|H\rangle+\frac{1}{5}|T\rangle & \text { and } & c_{2}(1)
\end{aligned}=\frac{1}{2}|H\rangle+\frac{1}{2}|T\rangle,
$$

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 7.8.1. Consider two-element sets $A=\left\{a, a^{\perp}\right\}, B=\left\{b, b^{\perp}\right\}, C=\left\{c, c^{\perp}\right\}$ and $D=\left\{d, d^{\perp}\right\}$ and an (accessible) string diagram $S$ of the form:


Now suppose we have a joint distribution $\omega \in \mathcal{D}(A \times C \times D \times B)$ given by:

$$
\begin{aligned}
\omega=\frac{1}{25} & |a, c, d, b\rangle+\frac{9}{50}\left|a, c, d, b^{\perp}\right\rangle+\frac{3}{50}\left|a, c, d^{\perp}, b\right\rangle+\frac{1}{50}\left|a, c, d^{\perp}, b^{\perp}\right\rangle \\
& +\frac{1}{25}\left|a, c^{\perp}, d, b\right\rangle+\frac{9}{50}\left|a, c^{\perp}, d, b^{\perp}\right\rangle+\frac{3}{50}\left|a, c^{\perp}, d^{\perp}, b\right\rangle \\
& +\frac{1}{50}\left|a, c^{\perp}, d^{\perp}, b^{\perp}\right\rangle+\frac{3}{122}\left|a^{\perp}, c, d, b\right\rangle+\frac{9}{500}\left|a^{\perp}, c, d, b^{\perp}\right\rangle \\
& +\frac{9}{250}\left|a^{\perp}, c, d^{\perp}, b\right\rangle+\frac{1}{500}\left|a^{\perp}, c, d^{\perp}, b^{\perp}\right\rangle+\frac{12}{125}\left|a^{\perp}, c^{\perp}, d, b\right\rangle \\
& +\frac{9}{125}\left|a^{\perp}, c^{\perp}, d, b^{\perp}\right\rangle+\frac{18}{125}\left|a^{\perp}, c^{\perp}, d^{\perp}, b\right\rangle+\frac{1}{125}\left|a^{\perp}, c^{\perp}, d^{\perp}, b^{\perp}\right\rangle
\end{aligned}
$$

We ask ourselves: does $\omega \approx S$ hold? More specifically, can we somehow obtain interpretations of the boxes $\sigma, f$ and $g$ in $S$ so that $\omega$ 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 $\sigma$ as:

$$
\begin{aligned}
\sigma= & \omega[1,0,0,1] \\
= & \left(\sum_{u \in C, v \in D} \omega(a, u, v, b)\right)|a, b\rangle+\left(\sum_{u \in C, v \in D} \omega\left(a, u, v, b^{\perp}\right)\right)\left|a, b^{\perp}\right\rangle \\
& +\left(\sum_{u \in C, v \in D} \omega\left(a^{\perp}, u, v, b\right)\right)\left|a^{\perp}, b\right\rangle+\left(\sum_{u \in C, v \in D} \omega\left(a^{\perp}, u, v, b^{\perp}\right)\right)\left|a^{\perp}, b^{\perp}\right\rangle \\
= & \left(\frac{1}{25}+\frac{3}{50}+\frac{1}{25}+\frac{3}{50}\right)|a, b\rangle+\left(\frac{9}{50}+\frac{1}{50}+\frac{9}{50}+\frac{1}{50}\right)\left|a, b^{\perp}\right\rangle \\
& \quad+\left(\frac{3}{125}+\frac{9}{250}+\frac{12}{125}+\frac{18}{125}\right)\left|a^{\perp}, b\right\rangle+\left(\frac{9}{500}+\frac{1}{500}+\frac{9}{125}+\frac{1}{125}\right)\left|a^{\perp}, b^{\perp}\right\rangle \\
= & \frac{1}{5}|a, b\rangle+\frac{2}{5}\left|a, b^{\perp}\right\rangle+\frac{3}{10}\left|a^{\perp}, b\right\rangle+\frac{1}{10}\left|a^{\perp}, b^{\perp}\right\rangle .
\end{aligned}
$$

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 7.8 . In the same way one obtains $g$ from the marginal
$\omega[0,0,1,1]$. Thus:

$$
\begin{aligned}
f & =\omega[0,1,0,0 \mid 1,0,0,0] \\
& =\left\{\begin{aligned}
a & \mapsto \frac{\sum_{v, y} \omega(a, c, v, y)}{\sum_{u, v, y} \omega(a, u, v, y)}|c\rangle+\frac{\sum_{v, y} \omega\left(a, c^{\perp}, v, y\right)}{\sum_{u, v, v} \omega(a, u, v, y)}\left|c^{\perp}\right\rangle \\
a^{\perp} & \mapsto \frac{\sum_{v, y} \omega\left(a^{\perp}, c, v, y\right)}{\sum_{u, v, y} \omega\left(a^{\perp}, u, v, y\right)}|c\rangle+\frac{\sum_{v, y} \omega\left(a^{\perp}, c^{\perp}, v, y\right)}{\sum_{u, v, y} \omega\left(a^{\perp}, u, v, y\right)}\left|c^{\perp}\right\rangle
\end{aligned}\right. \\
& =\left\{\begin{aligned}
a & \mapsto \frac{1}{2}|c\rangle+\frac{1}{2}\left|c^{\perp}\right\rangle \\
a^{\perp} & \mapsto \frac{1}{5}|c\rangle+\frac{4}{5}\left|c^{\perp}\right\rangle
\end{aligned}\right. \\
g & =\omega[0,0,1,0 \mid 0,0,0,1] \\
& =\left\{\begin{aligned}
b & \mapsto \frac{\sum_{x, u} \omega(x, u, d, b)}{\sum_{x, u, v} \omega(x, u, v, b)}|d\rangle+\frac{\sum_{x, u} \omega\left(x, u, d^{\perp}, b\right)}{\sum_{x, u, v} \omega(x, u, v, b)}\left|d^{\perp}\right\rangle \\
b^{\perp} & \mapsto \frac{\sum_{x, u} \omega\left(x, u, d, b^{\perp}\right)}{\sum_{x, u, v} \omega\left(x, u, v, b^{\perp}\right)}|d\rangle+\frac{\sum_{x, v} \omega\left(x, u, d^{\perp}, b^{\perp}\right)}{\sum_{x, u, v} \omega\left(x, u, v, b^{\perp}\right)}\left|d^{\perp}\right\rangle
\end{aligned}\right. \\
& =\left\{\begin{aligned}
b & \mapsto \frac{2}{5}|d\rangle+\frac{3}{5}\left|d^{\perp}\right\rangle \\
b^{\perp} & \mapsto \frac{9}{10}|d\rangle+\frac{1}{10}\left|d^{\perp}\right\rangle .
\end{aligned}\right.
\end{aligned}
$$

At this stage one can check that the joint distribution $\omega$ can be reconstructed from these extracted state and channels, namely as:

$$
\begin{aligned}
\omega & =(i d \otimes f \otimes g \otimes i d) \gg=((\Delta \otimes \Delta) \gg=\sigma) \\
& =(\langle i d, f\rangle \otimes\langle g, i d\rangle) \gg=\sigma .
\end{aligned}
$$

This proves $\omega \approx S$.
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 7.8.2. Let a state $\omega \in \mathcal{D}(A \times B \times C)$ have full support.
1 There are equivalences:


2 These equivalences can be extended as:

$$
\omega \approx \underset{\square}{\square \square} \Longleftrightarrow\left(\left.\omega\right|_{1 \otimes \mathbf{1}_{b} \otimes \mathbf{1}}\right)[1,0,1] \approx \square \square \quad \text { for all } b \in B .
$$

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 $\omega$ 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 $\omega=\langle c, i d, d\rangle\rangle=\tau$, consisting of a state $\tau=\omega[0,1,0] \in \mathcal{D}(B)$ and channels $c: B \leadsto A$ and $d: B \leadsto C$. We write $\sigma=c \gg \tau=\omega[1,0,0]$ and take the Bayesian inversion $c_{\tau}^{\dagger}: A \leadsto B$. We now have an interpretation of the string diagram in the middle, which is equal to $\omega$, since by (7.2):


Similarly one obtains an interpretation of the string diagram on the right via $\rho=d \gg=\tau=\omega[0,0,1]$ and the inversion $d_{\tau}^{\dagger}: C \rightsquigarrow B$.

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, Proposition 7.8.3 1) below gives a slightly stronger result.

We concentrate on $(\Leftarrow)$. By assumption, for $b \in B$ we can write:

$$
\left(\left.\omega\right|_{1 \otimes 1_{b} \otimes \mathbf{1}}\right)[1,0,1]=\sigma_{b} \otimes \tau_{b} \quad \text { for } \sigma_{b} \in \mathcal{D}(A), \tau_{b} \in \mathcal{D}(C)
$$

In a next step we define channels $f: B \rightsquigarrow A$ and $g: B \multimap C$ and a state $\rho \in \mathcal{D}(B)$ as:

$$
f(b):=\sigma_{b} \quad g(b):=\tau_{b} \quad \rho:=\omega[0,1,0] .
$$

Then, for $x \in A$ and $z \in C$,

$$
\begin{aligned}
f(b)(x) \cdot g(b)(z)=(f(b) \otimes g(b))(x, z) & =\left(\sigma_{b} \otimes \tau_{b}\right)(x, z) \\
& =\left(\left.\omega\right|_{\left.1 \otimes \mathbf{1}_{b} \otimes 1\right)}\right)[1,0,1](x, z) \\
& =\sum_{y} \frac{\omega(x, y, z) \cdot \mathbf{1}_{b}(y)}{\omega \models \mathbf{1} \otimes \mathbf{1}_{b} \otimes \mathbf{1}} \\
& =\frac{\omega(x, b, z)}{\omega[0,1,0] \models \mathbf{1}_{b}} \\
& =\frac{\omega(x, b, z)}{\rho(b)} .
\end{aligned}
$$

Now we see that $\omega$ has a fork shape:

$$
\omega(y, b, z)=f(b)(x) \cdot g(b)(z) \cdot \rho(b)=(\langle f, \text { id, } g\rangle \gg=\rho)(y, b, z) .
$$

What the first item of this result shows is that (sub)shapes of the form:

can be changed into
and vice-versa. We knew this already from Theorem 7.2.1 By applying these transformations directly to the shapes in Theorem 7.8 .2 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 [97], for three basic shapes, called fork, chain and collider,

Proposition 7.8.3. Let $\omega \in \mathcal{D}(X \times Y \times Z)$ be an arbitrary distribution and let $q \in \operatorname{Fact}(Y)$ 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 $\omega$ have fork shape:


In the special case of conditioning with a point predicate we get:

$$
\left.\omega\right|_{\otimes 1_{a} \otimes 1} \approx \square \square \square .
$$

where the middle box is $1|a\rangle$.
2 If $\omega$ has a chain shape:


3 Let $\omega$ have collider shape:


For this shape it does not matter if $q$ is a point predicate or not.
Proof. 1 Let's assume we have an interpretation $\omega=\langle c, i d, d\rangle \geqslant=\sigma$, for

$$
\begin{aligned}
& c: Y \nrightarrow X, d: Y \nrightarrow Z \text { and } \sigma \in \mathcal{D}(Y) \text {. Then: } \\
& \left.\omega\right|_{1 \otimes q \otimes 1}=\left.(\langle c, i d, d\rangle \gg \sigma)\right|_{1 \otimes q \otimes \mathbf{1}} \\
& \left.=\left.\langle c, i d, d\rangle\right|_{{ }_{\otimes q \otimes 1}} \gg=\left.\sigma\right|_{\langle c, i d, d\rangle=\langle(\mathbf{1} \otimes q \otimes \mathbf{1})}\right) \quad \text { by Corrolary 6.3.5 (2) } \\
& =\langle c, i d, d\rangle \gg=\left.\sigma\right|_{q} \quad \text { by Exercises 6.1.12 } \\
& \text { and Lemma 4.3.2 7. }
\end{aligned}
$$

Hence we see the same shape that $\omega$ has.
In the special case when $q=\mathbf{1}_{a}$ we can extend the above calculation and obtain a parallel product of states:

$$
\begin{array}{rlrl}
\left.\omega\right|_{1 \otimes \mathbf{1}_{a} \otimes \mathbf{1}} & =\langle c, \text { id }, d\rangle \gg=\left(\left.\sigma\right|_{\mathbf{1}_{a}}\right) & & \text { as just shown } \\
& =\left((c \otimes i d \otimes d) \odot \Delta_{3}\right) \gg 1|a\rangle & & \text { by Lemma6.1.6] [2] } \\
& =(c \otimes i d \otimes d) \gg\left(\Delta_{3} \gg 1|a\rangle\right) & & \\
& =(c \otimes i d \otimes d) \gg(1|a\rangle \otimes 1|a\rangle \otimes 1|a\rangle) & & \\
& =(c \gg 1|a\rangle) \otimes 1|a\rangle \otimes(d \gg=1|a\rangle) & & \\
& =c(a) \otimes 1|a\rangle \otimes d(a) . &
\end{array}
$$

2 By the previous point and Theorem 7.8.2
3 Let's assume as interpretation of the collider shape:

$$
\omega=(\mathrm{id} \otimes c \otimes \mathrm{id}) \gg=((\Delta \otimes \Delta) \gg=(\sigma \otimes \tau)),
$$

for states $\sigma \in \mathcal{D}(X)$ and $\tau \in \mathcal{D}(Z)$ and channel $c: X \times Z \leadsto Y$. Then:

$$
\begin{aligned}
& \left.\omega\right|_{1 \otimes q \otimes 1} \\
& =\left.((\mathrm{id} \otimes c \otimes \mathrm{id}) \gg((\Delta \otimes \Delta) \gg=(\sigma \otimes \tau)))\right|_{1_{\otimes q \otimes 1}} \\
& =\left(i d \otimes c l_{q} \otimes i d\right) \gg=\left(\left.((\Delta \otimes \Delta) \gg=(\sigma \otimes \tau))\right|_{1 \otimes(c \ll q) \otimes 1}\right) \\
& \text { by Corrolary 6.3.5 (3) } \\
& =\left(\left.i d \otimes c\right|_{q} \otimes i d\right) »=\left(\left.(\Delta \otimes \Delta)\right|_{1 \otimes(c=<q) \otimes 1} \gg=\left(\left.(\sigma \otimes \tau)\right|_{(\Delta \otimes \Delta)=\ll(\mathbf{1} \otimes(c=<q) \otimes \mathbf{1})}\right)\right. \\
& \text { by Theorem6.3.4 } \\
& =\left(\left.i d \otimes c\right|_{q} \otimes \mathrm{id}\right) \gg\left((\Delta \otimes \Delta) \gg=\left(\left.(\sigma \otimes \tau)\right|_{(\Delta \otimes \Delta)=<(1 \otimes(c *<q) \otimes 1)}\right)\right. \\
& \text { by Exercise 6.1.11. }
\end{aligned}
$$

The updated state $\left.(\sigma \otimes \tau)\right|_{(\Delta \otimes \Delta)=\langle(\mathbf{1} \otimes(c=\langle q) \otimes \mathbf{1})}$ is typically entwined, even if $q$ is a point predicate, see also Exercise 6.1.9

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 [147, 148], see the causal surgery procedure in Section ??. 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.

## Exercises

7.8.1 Check the aim of Exercise 2.3.2 really is to prove the shape statement

$$
\omega \approx \square \square
$$

for the (ternary) state $\omega$ defined there.
7.8.2 Prove that a collider shape leads to non-entwinedness:

7.8.3 Prove the following items, which are known as the 'semigraphoid' properties, see [179, 60]; they are seen as the basic rules of conditional independence.



3 Weak union: $\omega \approx \square$ implies $\omega \approx \square \square$
Hint: Apply disintegration to the upper-left box.
4 Contraction: if $\omega \approx \square$ and also $\omega[1,1,1,0] \approx \square$ then $\omega \approx \underset{\square}{\square}$.
Hint: Form a suitable combination of the two upper-right boxes in the assumptions.

### 7.9 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 7.9.1 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 section 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 7.9.1. Recall the category $\operatorname{Chan}(\mathcal{P})$ of non-deterministic computations. Its objects are sets $X$ and its morphisms $f: X \leadsto Y$ are functions $f: X \rightarrow \mathcal{P}(Y)$. The identity morphism unit $X: X \mapsto X$ in $\operatorname{Chan}(\mathcal{P})$ is the singleton function unit $(x)=\{x\}$. Composition of $f: X \leadsto Y$ and $g: Y \leadsto Z$ is the function $g \odot f: X \leadsto Z$ given by:

$$
(g \odot f)(x)=\{z \in Z \mid \exists y \in Y . y \in f(x) \text { and } z \in g(y)\} .
$$

It is not hard to see that $\odot$ is associative and has unit as identity element. In fact, this has already been proven more generally, in Lemma 1.10 .3

There are two aspects of the category $\operatorname{Chan}(\mathcal{P})$ 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 Chan $(\mathcal{P})$ can be reversed. There is a bijective correspondence, indicated by the double lines, between morphisms in $\operatorname{Chan}(\mathcal{P})$ :

$$
\xlongequal[Y \longrightarrow \longrightarrow]{X \longrightarrow \longrightarrow} \quad \text { that is, between functions: } \quad \frac{X \xrightarrow{\circ \longrightarrow} \mathcal{P}(Y)}{Y \longrightarrow \mathcal{P}(X)}
$$

This correspondence sends $f: X \rightarrow \mathcal{P}(Y)$ to the function $f^{\dagger}: Y \rightarrow \mathcal{P}(X)$ with $f^{\dagger}(y):=\{x \mid y \in f(x)\}$. Hence $y \in f(x)$ iff $x \in f^{\dagger}(y)$. Similarly one sends $g: Y \rightarrow \mathcal{P}(X)$ to $g^{\dagger}: X \rightarrow \mathcal{P}(Y)$ 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 unit ${ }^{\dagger}=$ unit and also $(g \odot f)^{\dagger}=f^{\dagger} \odot g^{\dagger}$. This means that the dagger is functorial. It can be described as a functor $(-)^{\dagger}: \operatorname{Chan}(\mathcal{P}) \rightarrow$ $\operatorname{Chan}(\mathcal{P})^{\mathrm{op}}$, 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 $X \rightarrow Y$. The identity $X \rightarrow X$ is given by the equality relation $E q_{X} \subseteq X \times X$, with $E q_{X}=\{(x, x) \mid x \in X\}$. 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 $G: \operatorname{Chan}(\mathcal{P}) \rightarrow \operatorname{Rel}$, which is the identity on objects: $G(X)=X$. On a morphism $f: X \leadsto Y$, that is, on a function $f: X \rightarrow \mathcal{P}(Y)$, we define $G(f) \subseteq X \times Y$ to be $G(f)=\{(x, y) \mid x \in X, y \in f(x)\}$. Then: $G\left(\right.$ unit $\left._{X}\right)=E q_{X}$ and $G(g \odot f)=G(f) \bullet G(f)$.

In the other direction there is also a functor $F: \operatorname{Rel} \rightarrow \operatorname{Chan}(\mathcal{P})$, which is again the identity on objects: $F(X)=X$. On a morphism $R: X \rightarrow Y$ in Rel, that is, on a relation $R \subseteq X \times Y$, we define $F(R): X \rightarrow \mathcal{P}(Y)$ 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:

$$
F \circ G=i d: \operatorname{Chan}(\mathcal{P}) \rightarrow \mathbf{C h a n}(\mathcal{P}) \quad \text { and } \quad G \circ F=i d: \operatorname{Rel} \rightarrow \mathbf{R e l} .
$$

This establishes an isomorphism $\operatorname{Chan}(\mathcal{P}) \cong \operatorname{Rel}$ 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 $(-)^{\dagger}: \mathbf{R e l} \rightarrow \mathbf{R e l}^{\mathrm{op}}$, obviously with $(-)^{\dagger \dagger}=$ id .

Moreover, the above functors $G$ and $F$ commute with the daggers of $\operatorname{Chan}(\mathcal{P})$ and Rel, in the sense that:

$$
G\left(f^{\dagger}\right)=G(f)^{\dagger} \quad \text { and } \quad F\left(R^{\dagger}\right)=F(R)^{\dagger} .
$$

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 $f: X \rightarrow \mathcal{P}(Y)$ and elements $x \in X, y \in Y$,

$$
G\left(f^{\dagger}\right)(y, x) \Leftrightarrow x \in f^{\dagger}(y) \Leftrightarrow y \in f(x) \Leftrightarrow G(f)(x, y) \Leftrightarrow G(f)^{\dagger}(y, x) .
$$

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 $\operatorname{Chan}(\mathcal{D})$ of probabilistic channels. But the situation is a bit more subtle. Recall from the previous section that the dagger (Bayesian inversion) $c_{\omega}^{\dagger}: Y \rightarrow X$ of a probabilistic channel $c: X \rightarrow Y$ requires a 'prior' distribution $\omega \in \mathcal{D}(X)$ on the domain - with side-condition that $c \gg=\omega$ has full support. In order to conveniently deal with this situation we incorporate these
distributions $\omega$ into the objects of our category. We follow [27] and denote this category as Krn; its morphisms are 'kernels'.

Definition 7.9.2. The category Krn of kernels has:

- objects: pairs $(X, \sigma)$ where $X$ is a finite set and $\sigma \in \mathcal{D}(X)$ is a distribution on $X$ with full support;
- morphisms: $f:(X, \sigma) \rightarrow(Y, \tau)$ are probabilistic channels $f: X \rightarrow Y$ with $f \gg \sigma=\tau$.

Identity maps $(X, \sigma) \rightarrow(X, \sigma)$ in Krn are identity channels unit: $X \leadsto X$, given by unit $(x)=1|x\rangle$, which we write simply as id. Composition in Krn is ordinary composition $\odot$ of channels.

Theorem 7.9.3. Bayesian inversion forms a dagger functor $(-)^{\dagger}: \mathbf{K r n} \rightarrow$ Krn ${ }^{\text {op }}$ which is the identity on objects and which sends:

$$
((X, \sigma) \xrightarrow{f}(Y, \tau)) \longmapsto\left((Y, \tau) \xrightarrow{f_{\sigma}^{\ddagger}}(X, \sigma)\right)
$$

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.

$$
f_{\sigma}^{\dagger} \gg \tau=f_{\sigma}^{\dagger} \gg(f \gg=\sigma) \stackrel{\sqrt{7.3}}{=} \sigma .
$$

Aside: this does not mean that $f^{\dagger} \odot f=i d$.
Identities and composition in Krn are preserved by Lemma 7.1.8) 3) and (2). Finally we have $\left(f_{\sigma}^{\dagger}\right)_{\tau}^{\dagger}=f$ 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. [12].

Definition 7.9.4. We introduce a category $\mathbf{C p l}$ of couplings with the same objects as Krn. A morphism $(X, \sigma) \rightarrow(Y, \tau)$ in $\mathbf{C p l}$ 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 $\sigma$ and $\tau$ is called a coupling between $\sigma$ and $\tau$, see Definition 3.1.5

Composition of $\varphi:(X, \sigma) \rightarrow(Y, \tau)$ and $\psi:(Y, \tau) \rightarrow(Z, \rho)$ is the distribution $\psi \bullet \varphi \in \mathcal{D}(X \times Z)$ defined as:

$$
\begin{align*}
\psi \bullet \varphi & :=\langle\varphi[1,0 \mid 0,1], \psi[0,1 \mid 1,0]\rangle \gg=\tau \\
& =\sum_{x \in X, z \in Z}\left(\sum_{y \in Y} \frac{\varphi(x, y) \cdot \psi(y, z)}{\tau(y)}\right)|x, z\rangle . \tag{7.33}
\end{align*}
$$

The identity coupling $E q_{(X, \sigma)}:(X, \sigma) \rightarrow(X, \sigma)$ is the distribution $\Delta »=\sigma$.

The essence of the following result is due to [27], but there it occurs in slightly different form, namely in a setting of continuous probability. Here it is transferred to the discrete situation.

Theorem 7.9.5. Couplings as defined above indeed form a category $\mathbf{C p l}$.
1 This category carries a dagger functor $(-)^{\dagger}: \mathbf{C p l} \rightarrow \mathbf{C p l}^{\mathrm{op}}$ which is the identity on objects; on morphisms it is defined via swapping:

$$
((X, \sigma) \xrightarrow{\varphi}(Y, \tau))^{\dagger}:=\left((Y, \tau) \xrightarrow{\left\langle\pi_{2}, \pi_{1}\right\rangle \ggg}(X, \sigma)\right) .
$$

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 $\mathbf{K r n} \cong \mathbf{C p l}$, 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:

$$
\begin{aligned}
\sum_{x \in X, z \in Z}(\psi \bullet \varphi)(x, z) & =\sum_{x \in X, y \in Y, z \in Z} \frac{\varphi(x, y) \cdot \psi(y, z)}{\tau(y)} \\
& =\sum_{y \in Y, z \in Z} \frac{\left(\sum_{x \in X} \varphi(x, y)\right) \cdot \psi(y, z)}{\tau(y)} \\
& =\sum_{y \in Y, z \in Z} \frac{\varphi[0,1](y) \cdot \psi(y, z)}{\tau(y)} \\
& =\sum_{y \in Y, z \in Z} \frac{\tau(y) \cdot \psi(y, z)}{\tau(y)}=\sum_{y \in Y, z \in Z} \psi(y, z)=1
\end{aligned}
$$

We leave it to the reader to check that $E q_{(X, \sigma)}=\Delta \gg \sigma$ is neutral element for $\bullet$ We do verify that $\bullet$ is associative - and thus that $\mathbf{C p l}$ is indeed a category. Let $\varphi:(X, \sigma) \rightarrow(Y, \tau), \psi:(Y, \tau) \rightarrow(Z, \rho), \chi:(Z, \rho) \rightarrow(W, \kappa)$ be morphisms in Cpl. Then:

$$
\begin{aligned}
(\chi \bullet(\psi \bullet \varphi))(x, w) & =\sum_{z \in Z} \frac{(\psi \bullet \varphi)(x, z) \cdot \chi(z, w)}{\rho(z)} \\
& =\sum_{y \in Y, z \in Z} \frac{\varphi(x, y) \cdot \psi(y, z) \cdot \chi(z, w)}{\tau(y) \cdot \rho(z)} \\
& =\sum_{y \in Y} \frac{\varphi(x, y) \cdot(\chi \bullet \psi)(y, w)}{\tau(y)}=((\chi \bullet \psi) \bullet \varphi)(x, w) .
\end{aligned}
$$

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:

$$
\begin{aligned}
(\psi \bullet \varphi)^{\dagger}(z, x)=(\psi \bullet \varphi)(x, z) & =\sum_{y \in Y} \frac{\varphi(x, y) \cdot \psi(y, z)}{\tau(y)} \\
& =\sum_{y \in Y} \frac{\psi^{\dagger}(z, y) \cdot \varphi^{\dagger}(y, x)}{\tau(y)}=\left(\varphi^{\dagger} \bullet \psi^{\dagger}\right)(z, x) .
\end{aligned}
$$

The graph operation on channels gives rise to an identity-on-objects 'graph' functor $G: \mathbf{K r n} \rightarrow \mathbf{C p l}$ via:

$$
G((X, \sigma) \xrightarrow{f}(Y, \tau)):=((Y, \tau) \xrightarrow{\langle i d, f\rangle \gg \sigma}(X, \sigma)) .
$$

This yields a functor since:

$$
\begin{aligned}
G\left(i d_{(X, \sigma)}\right) & =\langle\text { id, id }\rangle \gg=\sigma \\
& =E q_{(X, \sigma)} \\
G(g \odot f)(x, z) & =(\langle i d, g \odot f\rangle \gg \sigma)(x, z) \\
& =\sigma(x) \cdot(g \odot f)(x)(z) \\
& =\sum_{y \in Y} \sigma(x) \cdot f(x)(y) \cdot g(y)(z) \\
& =\sum_{y \in Y} \frac{\sigma(x) \cdot f(x)(y) \cdot \tau(y) \cdot g(y)(z)}{\tau(y)} \\
& =\sum_{y \in Y} \frac{(\langle i d, f\rangle \gg=\sigma)(x, y) \cdot(\langle i d, g\rangle \gg \tau)(y, z)}{\tau(y)} \\
& =((\langle i d, g\rangle \gg \tau) \bullet(\langle i d, f\rangle \gg \sigma))(x, z) \\
& =(G(g) \bullet G(f))(x, z) .
\end{aligned}
$$

In the other direction we define a functor $F: \mathbf{C p l} \rightarrow$ Krn which is the identity on objects and uses disintegration on morphisms: for $\varphi:(X, \sigma) \rightarrow$ $(Y, \tau)$ in Cpl we get a channel $F(\varphi):=\varphi[0,1 \mid 1,0]: X \leadsto Y$ which satisfies, by construction 7.8):

$$
\varphi=\langle\mathrm{id}, \varphi[0,1 \mid 1,0]\rangle \gg=\varphi[1,0]=\langle\mathrm{id}, F(\varphi)\rangle \gg=\sigma=G F(\varphi) .
$$

Moreover, $F(\varphi)$ is a morphism $(X, \sigma) \rightarrow(Y, \tau)$ in Krn since:

$$
\begin{aligned}
F(\varphi) \gg \sigma & =\varphi[0,1 \mid 1,0] \gg=\varphi[1,0] \\
& =(\langle\text { id }, \varphi[0,1 \mid 1,0]\rangle \gg \varphi[1,0])[0,1] \stackrel{\boxed{77.8}}{=} \varphi[0,1]=\tau .
\end{aligned}
$$

We still need to prove that $F$ preserves identities and composition. This follows by uniqueness of disintegration. We get $F\left(E q_{(X, \sigma)}\right)=i d_{F(X, \sigma)}$ from:

$$
\left\langle i d, F\left(E q_{(X, \sigma)}\right)\right\rangle \gg=\sigma=E q_{(X, \sigma)}=\Delta \gg=\sigma=\langle i d, i d\rangle \gg=\sigma \text {. }
$$

Next, the equation $F(\psi \bullet \varphi)=F(\psi) \odot F(\varphi)$ follows from:

```
\(\langle i d, F(\psi \bullet \varphi)\rangle \gg=\sigma\)
    \(=\psi \bullet \varphi\)
        by definition
    [7.33) \(\langle\varphi[1,0 \mid 0,1], \psi[0,1 \mid 1,0]\rangle \gg \tau\)
    \(=(i d \otimes \psi[0,1 \mid 1,0])\rangle=(\langle\varphi[1,0 \mid 0,1], i d\rangle \gg=(\varphi[1,0 \mid 0,1]\rangle=\sigma))\)
    \(=(i d \otimes F(\psi)) \gg=\left(\left\langle F(\varphi)^{\dagger}, i d\right\rangle \gg(F(\varphi) \gg \sigma)\right) \quad\) by Exercise 7.2.7
    \(\stackrel{7.8}{=}(\) id \(\otimes F(\psi)) \geqslant=(\langle i d, F(\varphi)\rangle \gg=\sigma)\)
    \(=\langle\mathrm{id}, F(\psi) \odot F(\varphi)\rangle \gg \sigma\).
```

We have seen that, by construction, $G \circ F$ is the identity functor on the category Cpl. In the other direction we also have $F \circ G=i d:$ Krn $\rightarrow$ 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 $f:(X, \sigma) \rightarrow$ $(Y, \tau)$ we have:

$$
\begin{aligned}
G\left(f^{\dagger}\right)(y, x)=\left(\left\langle\mathrm{id}, f^{\dagger}\right\rangle \gg \tau\right)(y, x) & =\tau(y) \cdot f_{\sigma}^{\dagger}(y)(x) \\
\stackrel{\text { 有 }}{=} & (f \gg=\sigma)(y) \cdot \frac{\sigma(x) \cdot f(x)(y)}{(f \gg \sigma)(y)} \\
& =\sigma(x) \cdot f(x)(y) \\
& =(\langle i d, f\rangle \gg \sigma)(x, y) \\
& =G(f)(x, y)=G(f)^{\dagger}(y, x) .
\end{aligned}
$$

Next, for $\varphi:(X, \sigma) \rightarrow(Y, \tau)$ in $\mathbf{C p l}$,

$$
\begin{aligned}
F\left(\varphi^{\dagger}\right)(y)(x) & =\varphi^{\dagger}[0,1 \mid 1,0](y)(x) \\
& \stackrel{7.9}{=} \frac{\varphi^{\dagger}(y, x)}{\varphi^{\dagger}[1,0](y)} \\
& =\frac{\varphi(x, y)}{\varphi[0,1](y)} \\
& =\varphi[1,0 \mid 0,1](y)(x) \\
& =(\varphi[0,1 \mid 1,0])^{\dagger}(y)(x) \\
& =F(\varphi)^{\dagger}(y)(x) .
\end{aligned}
$$

We have done a lot of work in order to be able to say that Krn and $\mathbf{C p l}$ are isomorphic dagger categories, or, more informally, that there is a one-one correspondence between probabilistic computations (channels) and probabilistic relations.

## Exercises

7.9.1 Prove in the context of the powerset channels of Example 7.9.1 that:

1 unit $^{\dagger}=$ unit and $(g \odot f)^{\dagger}=f^{\dagger} \odot g^{\dagger}$.
$2 G\left(u n i t_{X}\right)=E q_{X}$ and $G(g \odot f)=G(f) \bullet G(f)$.
$3 \quad F\left(E q_{X}\right)=u n i t_{X}$ and $F(S \bullet R)=F(S) \odot F(R)$.
$4(S \bullet R)^{\dagger}=R^{\dagger} \bullet S^{\dagger}$.
$5 \quad F\left(R^{\dagger}\right)=F(R)^{\dagger}$.
7.9.2 1 Define tensors $\otimes$ on the category of kernels, as a functor $\mathbf{K r n} \times$ $\mathbf{K r n} \rightarrow \mathbf{K r n}$.
2 Prove that the dagger functor on Krn preseves these tensors: $(f \otimes$ $g)^{\dagger}=f^{\dagger} \otimes g^{\dagger}$.
7.9.3 Prove the equation $=$ in the definition (7.33) of composition $\bullet$ in the category $\mathbf{C p l}$. Give also a string diagrammatic description of $\bullet$.

## References

[1] S. Abramsky. Contextual semantics: From quantum mechanics to logic, databases, constraints, and complexity. EATCS Bulletin, 113, 2014.
[2] S. Abramsky and A. Jung. Domain theory. In S. Abramsky, Dov M. Gabbay, and T. Maibaum, editors, Handbook of Logic in Computer Science, volume 3, pages 1-168. Oxford Univ. Press, 1994.
[3] D. Aldous. Exchangeability and related topics. In P. Hennequin, editor, École d'Été de Probabilités de Saint-Flour XIII - 1983, number 1117 in Lect. Notes Math., pages 1-198. Springer, Berlin, 1985.
[4] E. Alfsen and F. Shultz. State spaces of operator algebras: basic theory, orientations and $C^{*}$-products. Mathematics: Theory \& Applications. Birkhauser Boston, 2001.
[5] C. Anderson. The end of theory. The data deluge makes the scientific method obsolete. Wired, 16:07, 2008. Available at https://www.wired.com/2008/ $06 / \mathrm{pb}$-theory/
[6] G. Andrews. The Theory of Partitions. Cambridge Univ. Press, 1998.
[7] S. Awodey. Category Theory. Oxford Logic Guides. Oxford Univ. Press, 2006.
[8] G. Bacci, R. Furber, D. Kozen, R. Mardare, P. Panangaden, and D. Scott. Boolean-valued semantics for the stochastic $\lambda$-calculus. In A. Dawar and E. Grädel, editors, Logic in Computer Science, pages 669-678. IEEE, Computer Science Press, 2018.
[9] D. Barber. Bayesian Reasoning and Machine Learning. Cambridge Univ. Press, 2012. Publicly available via http://web4.cs.ucl.ac.uk/staff/D. Barber/pmwiki/pmwiki.php?n=Brml.HomePage
[10] M. Barr and Ch. Wells. Toposes, Triples and Theories. Springer, Berlin, 1985. Revised and corrected version available from URL: www.tac.mta.ca/tac/ reprints/articles/12/tr12.pdf
[11] M. Barr and Ch. Wells. Category Theory for Computing Science. Prentice Hall, Englewood Cliffs, NJ, 1990.
[12] G. Barthe and J. Hsu. Probabilistic couplings from program logics. In G. Barthe, J.-P. Katoen, and A. Silva, editors, Foundations of Probabilistic Programming, pages 145-184. Cambridge Univ. Press, 2021.
[13] J. Beck. Distributive laws. In B. Eckman, editor, Seminar on Triples and Categorical Homology Theory, number 80 in Lect. Notes Math., pages 119-140. Springer, Berlin, 1969.
[14] J. Bernardo and A. Smith. Bayesian Theory. John Wiley \& Sons, 2000. Available via https://onlinelibrary.wiley.com/doi/book/10.1002/ 9780470316870
[15] C. Bishop. Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, 2006.
[16] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journ. Machince Learning Research, 3:993-1022, 2003.
[17] F. van Breugel, C. Hermida, M. Makkai, and J. Worrell. Recursively defined metric spaces without contraction. Theor. Comp. Sci., 380:143-163, 2007.
[18] P. Bruza and S. Abramsky. Probabilistic programs: Contextuality and relational database theory. In Quantum Interaction, pages 163-174, 2016.
[19] F. Buscemi and V. Scarani. Fluctuation theorems from bayesian retrodiction. Phys. Rev. E, 103:052111, 2021.
[20] A. Carboni and R. Walters. Cartesian bicategories I. Journ. of Pure $\mathcal{E}$ Appl. Algebra, 49(1-2):11-32, 1987.
[21] H. Chan and A. Darwiche. On the revision of probabilistic beliefs using uncertain evidence. Artif. Intelligence, 163:67-90, 2005.
[22] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. The Amer. Statistician, 49(4):327-335, 1995.
[23] K. Cho and B. Jacobs. The EfProb library for probabilistic calculations. In F. Bonchi and B. König, editors, Conference on Algebra and Coalgebra in Computer Science (CALCO 2017), volume 72 of LIPIcs. Schloss Dagstuhl, 2017.
[24] K. Cho and B. Jacobs. Disintegration and Bayesian inversion via string diagrams. Math. Struct. in Comp. Sci., 29(7):938-971, 2019.
[25] K. Cho, B. Jacobs, A. Westerbaan, and B. Westerbaan. An introduction to effectus theory. seehttps://arxiv.org/abs/1512.05813, 2015.
[26] A. Clark. Surfing Uncertainty. Prediction, Action, and the Embodied Mind. Oxford Univ. Press, 2016.
[27] F. Clerc, F. Dahlqvist, V. Danos, and I. Garnier. Pointless learning. In J. Esparza and A. Murawski, editors, Foundations of Software Science and Computation Structures, number 10203 in Lect. Notes Comp. Sci., pages 355-369. Springer, Berlin, 2017.
[28] B. Coecke and A. Kissinger. Picturing Quantum Processes. A First Course in Quantum Theory and Diagrammatic Reasoning. Cambridge Univ. Press, 2016.
[29] B. Coecke and R. Spekkens. Picturing classical and quantum Bayesian inference. Synthese, 186(3):651-696, 2012.
[30] J.B. Conway. A Course in Functional Analysis. Graduate Texts in Mathematics 96. Springer, $2^{\text {nd }}$ edition, 1990.
[31] T. Cover and J. Thomas. Elements of Information Theory. Wiley-Interscience, $2^{\text {nd }}$ edition, 2006.
[32] J. Culbertson and K. Sturtz. A categorical foundation for Bayesian probability. Appl. Categorical Struct., 22(4):647-662, 2014.
[33] F. Dahlqvist, V. Danos, and I. Garnier. Robustly parameterised higher-order probabilistic models. In J. Desharnais and R. Jagadeesan, editors, Int. Conf. on Concurrency Theory, volume 59 of LIPIcs, pages 23:1-23:15. Schloss Dagstuhl, 2016.
[34] F. Dahlqvist and D. Kozen. Semantics of higher-order probabilistic programs with conditioning. In Princ. of Programming Languages, pages 57:1-57:29. ACM Press, 2020.
[35] F. Dahlqvist, L. Parlant, and A. Silva. Layer by layer - combining monads. In B. Fischer and T. Uustalu, editors, Theoretical Aspects of Computing, number 11187 in Lect. Notes Comp. Sci., pages 153-172. Springer, Berlin, 2018.
[36] V. Danos and T. Ehrhard. Probabilistic coherence spaces as a model of higherorder probabilistic computation. Information $\mathcal{E}$ Computation, 209(6):966-991, 2011.
[37] A. Darwiche. Modeling and Reasoning with Bayesian Networks. Cambridge Univ. Press, 2009.
[38] S. Dash and S. Staton. A monad for probabilistic point processes. In D. Spivak and J. Vicary, editors, Applied Category Theory Conference, Elect. Proc. in Theor. Comp. Sci., 2020.
[39] S. Dash and S. Staton. Monads for measurable queries in probabilistic databases. In A. Sokolova, editor, Math. Found. of Programming Semantics, 2021.
[40] E. Davies and J. Lewis. An operational approach to quantum probability. Communic. Math. Physics, 17:239-260, 1970.
[41] B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. Memorie della R. Accademia Nazionale dei Lincei, IV, fasc. 5:86-113, 1930. Available at www. brunodefinetti.it/Opere/funzioneCaratteristica.pdf
[42] B. de Finetti. Theory of Probability: A critical introductory treatment. Wiley, 2017.
[43] P. Diaconis and S. Zabell. Updating subjective probability. Journ. American Statistical Assoc., 77:822-830, 1982.
[44] P. Diaconis and S. Zabell. Some alternatives to Bayes' rule. Technical Report 339, Stanford Univ., Dept. of Statistics, 1983.
[45] F. Dietrich, C. List, and R. Bradley. Belief revision generalized: A joint characterization of Bayes' and Jeffrey's rules. Journ. of Economic Theory, 162:352371, 2016.
[46] E. Dijkstra. A Discipline of Programming. Prentice Hall, Englewood Cliffs, NJ, 1976.
[47] E. Dijkstra and C. Scholten. Predicate Calculus and Program Semantics. Springer, Berlin, 1990.
[48] A. Dvurečenskij and S. Pulmannová. New Trends in Quantum Structures. Kluwer Acad. Publ., Dordrecht, 2000.
[49] T. Ehrhard, C. Tasson, and M. Pagani. Probabilistic coherence spaces are fully abstract for probabilistic PCF. In S. Jagannathan and P. Sewell, editors, Princ. of Programming Languages, pages 309-320. ACM Press, 2014.
[50] M. Erwig and E. Walkingshaw. A DSL for explaining probabilistic reasoning. In W. Taha, editor, Domain-Specific Languages, number 5658 in Lect. Notes Comp. Sci., pages 335-359. Springer, Berlin, 2009.
[51] C. Reijnen et al. Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study. PLoS Medic, 17(5): e1003111, 2020.
[52] W. Feller. An Introduction to Probability Theory and Its applications, volume I. Wiley, $3^{\text {rd }}$ rev. edition, 1970.
[53] R. Fisher. On the mathematical foundations of theoretical statistics. Philos. Trans. Royal Soc., 222:594-604, 1922.
[54] B. Fong. Causal theories: A categorical perspective on Bayesian networks. Master's thesis, Univ. of Oxford, 2012. see https://arxiv.org/abs/1301. 6201
[55] D. J. Foulis and M.K. Bennett. Effect algebras and unsharp quantum logics. Found. Physics, 24(10):1331-1352, 1994.
[56] K. Friston. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2):127-138, 2010.
[57] T. Fritz. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. Advances in Math., 370:107239, 2020.
[58] T. Fritz and E. Rischel. Infinite products and zero-one laws in categorical probability. Compositionality, 2(3), 2020.
[59] S. Fujii, S. Katsumata, and P. Melliès. Towards a formal theory of graded monads. In B. Jacobs and C. Löding, editors, Foundations of Software Science and Computation Structures, number 9634 in Lect. Notes Comp. Sci., pages 513530. Springer, Berlin, 2016.
[60] D. Geiger, T., and J. Pearl. Identifying independence in Bayesian networks. Networks, 20:507-534, 1990.
[61] A. Gibbs and F. Su. On choosing and bounding probability metrics. Int. Statistical Review, 70(3):419-435, 2002.
[62] G. Gigerenzer and U. Hoffrage. How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102(4):684-704, 1995.
[63] M. Giry. A categorical approach to probability theory. In B. Banaschewski, editor, Categorical Aspects of Topology and Analysis, number 915 in Lect. Notes Math., pages 68-85. Springer, Berlin, 1982.
[64] A. Gordon, T. Henzinger, A. Nori, and S. Rajamani. Probabilistic programming. In Future of Software Engineering, pages 167-181. ACM, 2014.
[65] T. Griffiths, C. Kemp, and J. Tenenbaum. Bayesian models of cognition. In R. Sun, editor, Cambridge Handbook of Computational Cognitive Modeling, pages 59-100. Cambridge Univ. Press, 2008.
[66] J. Halpern. Reasoning about Uncertainty. MIT Press, Cambridge, MA, 2003.
[67] M. Hayhoe, F. Alajaji, and B. Gharesifard. A pólya urn-based model for epidemics on networks. In American Control Conference, pages 358-363, 2017.
[68] C. Heunen, O. Kammar, S. Staton, and H. Yang. A convenient category for higher-order probability theory. In Logic in Computer Science, pages 1-12. IEEE Computer Society, 2017.
[69] W. Hino, H. Kobayashi I. Hasuo, and B. Jacobs. Healthiness from duality. In Logic in Computer Science. IEEE, Computer Science Press, 2016.
[70] J. Hohwy. The Predictive Mind. Oxford Univ. Press, 2013.
[71] M. Hyland and J. Power. The category theoretic understanding of universal algebra: Lawvere theories and monads. In L. Cardelli, M. Fiore, and G. Winskel, editors, Computation, Meaning, and Logic: Articles dedicated to Gordon Plotkin, number 172 in Elect. Notes in Theor. Comp. Sci., pages 437-458. Elsevier, Amsterdam, 2007.
[72] B. Jacobs. Convexity, duality, and effects. In C. Calude and V. Sassone, editors, IFIP Theoretical Computer Science 2010, number 82(1) in IFIP Adv. in Inf. and Comm. Techn., pages 1-19. Springer, Boston, 2010.
[73] B. Jacobs. New directions in categorical logic, for classical, probabilistic and quantum logic. Logical Methods in Comp. Sci., 11(3), 2015.
[74] B. Jacobs. Introduction to Coalgebra. Towards Mathematics of States and Observations. Number 59 in Tracts in Theor. Comp. Sci. Cambridge Univ. Press, 2016.
[75] B. Jacobs. Introduction to coalgebra. Towards mathematics of states and observations. Cambridge Univ. Press, to appear, 2016.
[76] B. Jacobs. Hyper normalisation and conditioning for discrete probability distributions. Logical Methods in Comp. Sci., 13(3:17), 2017. See https: //lmcs.episciences.org/3885
[77] B. Jacobs. A note on distances between probabilistic and quantum distributions. In A. Silva, editor, Math. Found. of Programming Semantics, Elect. Notes in Theor. Comp. Sci. Elsevier, Amsterdam, 2017.
[78] B. Jacobs. A recipe for state and effect triangles. Logical Methods in Comp. Sci., 13(2), 2017. See https://lmcs.episciences.org/3660
[79] B. Jacobs. The mathematics of changing one's mind, via Jeffrey's or via Pearl's update rule. Journ. of Artif. Intelligence Research, 65:783-806, 2019.
[80] B. Jacobs. From multisets over distributions to distributions over multisets. In Logic in Computer Science. IEEE, Computer Science Press, 2021.
[81] B. Jacobs. Learning from what's right and learning from what's wrong. In A. Sokolova, editor, Math. Found. of Programming Semantics, number 351 in Elect. Proc. in Theor. Comp. Sci., pages 116-133, 2021.
[82] B. Jacobs. Multinomial and hypergeometric distributions in Markov categories. In A. Sokolova, editor, Math. Found. of Programming Semantics, number 351 in Elect. Proc. in Theor. Comp. Sci., pages 98-115, 2021.
[83] B. Jacobs. Basic combinatorics and sufficient statistics for mutations on multiple datatypes. Researchers.One, https://researchers.one/articles/ 22.11.00003, 2022.
[84] B. Jacobs. Partitions and Ewens distributions in element-free probability theory. In Logic in Computer Science. IEEE, Computer Science Press, 2022. Article No. 23.
[85] B. Jacobs. A reconstruction of Ewens' sampling formula via lists of coins. In N. Jansen, M. Stoelinga, and P. van den Bos, editors, A Journey from Process Algebra via Timed Automata to Model Learning. Essays Dedicated to Frits Vaandrager on the Occasion of His 60th Birthday, number 13560 in Lect. Notes Comp. Sci., pages 339-357. Springer, Berlin, 2022.
[86] B. Jacobs. Stick breaking, in coalgebra and probability. In H. Hansen and F. Zanasi, editors, Coalgebraic Methods in Computer Science (CMCS 2022), number 13225 in Lect. Notes Comp. Sci. Springer, Berlin, 2022.
[87] B. Jacobs. Multisets and distributions, in drawing and learning. In A. Palmigiano and M. Sadrzadeh, editors, Samson Abramsky on Logic and Structure in Computer Science and Beyond. Springer, 2022, to appear.
[88] B. Jacobs. Sufficient statistics and split idempotents in discrete probability theory. In J. Hsu and Ch. Tasson, editors, Math. Found. of Programming Semantics, number 1 in Elect. Notes in Theor. Inform. \& Comp. Sci., 2023.
[89] B. Jacobs, J. Mandemaker, and R. Furber. The expectation monad in quantum foundations. Information $\mathcal{E}$ Computation, 250:87-114, 2016.
[90] B. Jacobs, A. Silva, and A. Sokolova. Trace semantics via determinization. Journ. of Computer and System Sci., 81(5):859-879, 2015.
[91] B. Jacobs and S. Staton. De Finetti's construction as a categorical limit. In D. Petrişan and J. Rot, editors, Coalgebraic Methods in Computer Science (CMCS 2020), number 12094 in Lect. Notes Comp. Sci., pages 90-111. Springer, Berlin, 2020.
[92] B. Jacobs and D. Stein. Counting and matching. In B. Klin and E. Pimentel, editors, Computer Science Logic, volume 252 of LIPIcs, pages 28:1-28:15. Schloss Dagstuhl, 2023.
[93] B. Jacobs and D. Stein. Overdrawing urns with signed probabilities. In Applied Category Th., 2023.
[94] B. Jacobs and D. Stein. Pearl's and Jeffrey's update as modes of learning in probabilistic programming. In M. Kerjean and P. Levy, editors, Math. Found. of Programming Semantics, Elect. Notes in Theor. Comp. Sci., 2023.
[95] B. Jacobs and A. Westerbaan. Distances between states and between predicates. Logical Methods in Comp. Sci., 16(1), 2020. See https://lmcs. episciences.org/6154
[96] B. Jacobs and F. Zanasi. A predicate/state transformer semantics for Bayesian learning. In L. Birkedal, editor, Math. Found. of Programming Semantics, number 325 in Elect. Notes in Theor. Comp. Sci., pages 185-200. Elsevier, Amsterdam, 2016.
[97] B. Jacobs and F. Zanasi. A formal semantics of influence in Bayesian reasoning. In K. Larsen, H. Bodlaender, and J.-F. Raskin, editors, Math. Found. of Computer Science, volume 83 of LIPIcs, pages 21:1-21:14. Schloss Dagstuhl, 2017.
[98] B. Jacobs and F. Zanasi. The logical essentials of Bayesian reasoning. In G. Barthe, J.-P. Katoen, and A. Silva, editors, Foundations of Probabilistic Programming, pages 295-331. Cambridge Univ. Press, 2021.
[99] E. Jaynes. Some applications and extensions of the De Finetti representation theorem. In P. Goel and A. Zellner, editors, Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti, volume 6 of Studies in Bayesian Econometrics and Statistics, pages 31-42. North Holland, 1982.
[100] E. Jaynes. Probability Theory: the Logic of Science. Cambridge Univ. Press, 2003.
[101] R. Jeffrey. The Logic of Decision. The Univ. of Chicago Press, $2^{\text {nd }}$ rev. edition, 1983.
[102] F. Jensen and T. Nielsen. Bayesian Networks and Decision Graphs. Statistics for Engineering and Information Science. Springer, $2^{\text {nd }}$ rev. edition, 2007.
[103] N. Johnson and S. Kotz. Urn models and their application: An approach to modern discrete probability theory. John Wiley, 1977.
[104] P. Johnstone. Stone Spaces. Number 3 in Cambridge Studies in Advanced Mathematics. Cambridge Univ. Press, 1982.
[105] C. Jones. Probabilistic Non-determinism. PhD thesis, Edinburgh Univ., 1989.
[106] P. Joyce. Partition structures and sufficient statistics. Journ. of Applied Probability, 35(3):622-632, 1998
[107] A. Jung and R. Tix. The troublesome probabilistic powerdomain. In A. Edalat, A. Jung, K. Keimel, and M. Kwiatkowska, editors, Comprox III, Third Workshop on Computation and Approximation, number 13 in Elect. Notes in Theor. Comp. Sci., pages 70-91. Elsevier, Amsterdam, 1998.
[108] D. Jurafsky and J. Martin. Speech and language processing. Third Edition draft, available at https://web.stanford.edu/~jurafsky/slp3/ 2018.
[109] O. Kallenberg. Foundations of Modern Probability. Number 99 in Probability Theory and Stochastic Modelling. Springer, 2021.
[110] E. Kamenica and M. Gentzkow. Bayesian persuasion. American Economic Review, 101(6):2590-2615, 2011.
[111] K. Keimel and G. Plotkin. Mixed powerdomains for probability and nondeterminism. Logical Methods in Comp. Sci., 13(1), 2017. See https://lmcs. episciences.org/2665.
[112] A. Kock. Closed categories generated by commutative monads. Journ. Austr. Math. Soc., XII:405-424, 1971.
[113] A. Kock. Monads for which structures are adjoint to units. Journ. of Pure $\mathcal{E}$ Appl. Algebra, 104:41-59, 1995.
[114] A. Kock. Commutative monads as a theory of distributions. Theory and Appl. of Categories, 26(4):97-131, 2012.
[115] D. Koller and N. Friedman. Probabilistic Graphical Models. Principles and Techniques. MIT Press, Cambridge, MA, 2009.
[116] D. Kozen. Semantics of probabilistic programs. Journ. Comp. Syst. Sci, 22(3):328-350, 1981.
[117] D. Kozen. A probabilistic PDL. Journ. Comp. Syst. Sci, 30(2):162-178, 1985.
[118] D. Kozen and A. Silva. Multisets and distributions, 2023. See https://arxiv. org/abs/2301.10812
[119] P. Lau, T. Koo, and C. Wu. Spatial distribution of tourism activities: A Pólya urn process model of rank-size distribution. Journ. of Travel Research, 59(2):231246, 2020.
[120] S. Lauritzen. Graphical models. Oxford Univ. Press, Oxford, 1996.
[121] S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. Journ. Royal Statistical Soc., 50(2):157-224, 1988.
[122] F. Lawvere. The category of probabilistic mappings. Unpublished manuscript, seencatlab.org/nlab/files/lawvereprobability1962.pdf 1962.
[123] T. Leinster. Basic Category Theory. Cambridge Studies in Advanced Mathematics. Cambridge Univ. Press, 2014. Available online via https://arxiv.org/ abs/1612.09375
[124] L. Libkin and L. Wong. Some properties of query languages for bags. In C. Beeri, A. Ohori, and D. Shasha, editors, Database Programming Languages, Workshops in Computing, pages 97-114. Springer, Berlin, 1993.
[125] D. Mackay. Information Theory, Inference, and Learning Algorithms. Cambridge Univ. Press, 2003.
[126] H. Mahmoud. Pólya Urn Models. Chapman and Hall, 2008.
[127] E. Manes. Algebraic Theories. Springer, Berlin, 1974.
[128] R. Mardare, P. Panangaden, and G. Plotkin. Quantitative algebraic reasoning. In Logic in Computer Science. IEEE, Computer Science Press, 2016.
[129] S. Mac Lane. Categories for the Working Mathematician. Springer, Berlin, 1971.
[130] S. Mac Lane. Mathematics: Form and Function. Springer, Berlin, 1986.
[131] A. McIver and C. Morgan. Abstraction, refinement and proof for probabilistic systems. Monographs in Comp. Sci. Springer, 2004.
[132] A. McIver, C. Morgan, and T. Rabehaja. Abstract hidden Markov models: A monadic account of quantitative information flow. In Logic in Computer Science, pages 597-608. IEEE, Computer Science Press, 2015.
[133] A. McIver, C. Morgan, G. Smith, B. Espinoza, and L. Meinicke. Abstract channels and their robust information-leakage ordering. In M. Abadi and S. Kremer, editors, Princ. of Security and Trust, number 8414 in Lect. Notes Comp. Sci., pages 83-102. Springer, Berlin, 2014.
[134] S. Milius, D. Pattinson, and L. Schröder. Generic trace semantics and graded monads. In L. Moss and P. Sobocinski, editors, Conference on Algebra and Coalgebra in Computer Science (CALCO 2015), volume 35 of LIPIcs, pages 253-269. Schloss Dagstuhl, 2015.
[135] T. Mitchell. Machine Learning. McGraw-Hill, 1997.
[136] E. Moggi. Notions of computation and monads. Information $\mathcal{E}$ Computation, 93(1):55-92, 1991.
[137] K. Murphy. Machine Learning. A Probabilistic Perspective. MIT Press, Cambridge, MA, 2012.
[138] R. Nagel. Order unit and base norm spaces. In A. Hartkämper and H. Neumann, editors, Foundations of Quantum Mechanics and Ordered Linear Spaces, number 29 in Lect. Notes Physics, pages 23-29. Springer, Berlin, 1974.
[139] M. Nielsen and I. Chuang. Quantum Computation and Quantum Information. Cambridge Univ. Press, 2000.
[140] F. Olmedo, F. Gretz, B. Lucien Kaminski, J-P. Katoen, and A. McIver. Conditioning in probabilistic programming. ACM Trans. on Prog. Lang. $\mathcal{E}$ Syst., 40(1):4:1-4:50, 2018.
[141] M. Ozawa. Quantum measuring processes of continuous observables. Journ. Math. Physics, 25:79-87, 1984.
[142] P. Panangaden. The category of Markov kernels. In C. Baier, M. Huth, M. Kwiatkowska, and M. Ryan, editors, Workshop on Probabilistic Methods in Verification, number 21 in Elect. Notes in Theor. Comp. Sci., pages 171-187. Elsevier, Amsterdam, 1998.
[143] R. Parikh. On context-free languages. Journ. ACM, 13(4):570-581, 1966.
[144] V. Paulsen and M. Tomforde. Vector spaces with an order unit. Indiana Univ. Math. Journ., 58-3:1319-1359, 2009.
[145] J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Graduate Texts in Mathematics 118. Morgan Kaufmann, 1988.
[146] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In R. Brachman, H. Levesque, and R. Reiter, editors, First Intl. Conf. on Principles of Knowledge Representation and Reasoning, pages 505-516. Morgan Kaufmann, 1989.
[147] J. Pearl. Causality. Models, Reasoning, and Inference. Cambridge Univ. Press, $2^{\text {nd }}$ ed. edition, 2009.
[148] J. Pearl and D. Mackenzie. The Book of Why. Penguin Books, 2019.
[149] B. Pierce. Basic Category Theory for Computer Scientists. MIT Press, Cambridge, MA, 1991.
[150] H. Pishro-Nik. Introduction to probability, statistics, and random processes. Kappa Research LLC, 2014. Available at https://www. probabilitycourse.com
[151] J. Pitman. Exchangeable and partially exchangeable random partitions. Probability Th. and related fields, 102(2):145-158, 1995.
[152] J. Pitman. Combinatorial Stochastic Processes: École d'Été de Probabilités de Saint-Flour XXXII - 2002. Number 1875 in Lect. Notes Math. Springer, 2006.
[153] G. Plotkin. Domains. The "Pisa" notes. Available from https://homepages. inf.ed.ac.uk/gdp/publications/Domains_a4.ps. 1983.
[154] D. Poole and A. Mackworth. Artificial Intelligence. Foundations of Computational Agents. Cambridge Univ. Press, $2^{\text {nd }}$ edition, 2017. Publicly available via https://www.cs.ubc.ca/~poole/aibook/2e/html/ArtInt2e.html
[155] S. Pulmannová and S. Gudder. Representation theorem for convex effect algebras. Commentationes Mathematicae Universitatis Carolinae, 39(4):645-659, 1998.
[156] J. Quinlan. Induction of decision trees. Machine Learning, 1(1):81-106, 1986.
[157] R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2:79-87, 1999.
[158] S. Ross. Introduction to Probability Models. Academic Press, $9^{\text {th }}$ edition, 2007.
[159] S. Ross. A first course in probability. Pearson Education, $10^{\text {th }}$ edition, 2018.
[160] S. Russell and P. Norvig. Artificial Intelligence. A Modern Approach. Prentice Hall, Englewood Cliffs, NJ, 2003.
[161] A. Ścibior, O. Kammar, M. Vákár, S. Staton, H. Yang, Y. Cai, K. Ostermann, S. Moss, C. Heunen, and Z. Ghahramani. Denotational validation of higherorder Bayesian inference. In Princ. of Programming Languages, pages 60:160:29. ACM Press, 2018.
[162] P. Scozzafava. Uniform distribution and sum modulo $m$ of independent random variables. Statistics EG Probability Letters, 18(4):313-314, 1993.
[163] P. Selinger. A survey of graphical languages for monoidal categories. In B. Coecke, editor, New Structures in Physics, number 813 in Lect. Notes Physics, pages 289-355. Springer, Berlin, 2011.
[164] S. Selvin. A problem in probability (letter to the editor). Amer. Statistician, 29(1):67, 1975.
[165] J. Sethuraman. A constructive definition of Dirichlet priors. Statistica Sinica, 4:639-650, 1994.
[166] S. Sloman. Causal Models. How People Think about the World and Its Alternatives. Oxford Univ. Press, 2005.
[167] A. Sokolova. Probabilistic systems coalgebraically: A survey. Theor. Comp. Sci., 412(38):5095-5110, 2011.
[168] S. Staton. Probabilistic programs as measures. In G. Barthe, J.-P. Katoen, and A. Silva, editors, Foundations of Probabilistic Programming, pages 43-74. Cambridge Univ. Press, 2021.
[169] S. Staton and N. Summer. Quantum de finetti theorems as categorical limits, and limits of state spaces of $\mathrm{c}^{*}$-algebras. In Quantum Programming Languages (QPL 2022), 2022. to appear.
[170] S. Staton, H. Yang, C. Heunen, O. Kammar, and F. Wood. Semantics for probabilistic programming: higher-order functions, continuous distributions, and soft constraints. In Logic in Computer Science. IEEE, Computer Science Press, 2016.
[171] M. Stone. Postulates for the barycentric calculus. Ann. Math., 29:25-30, 1949.
[172] Y. Suhov and M. Kelbert. Probability and Statistics by Example: Volume 1, Basic Probability and Statistics. Cambridge Univ. Press, 2005.
[173] L. Tierney. Markov chains for exploring posterior distributions. Ann. Stat., 22(4):1701-1728, 1994.
[174] H. Tijms. Understanding Probability: Chance Rules in Everyday Life. Cambridge Univ. Press, $2^{\text {nd }}$ edition, 2007.
[175] A. Tversky and D. Kahneman. Evidential impact of base rates. In D. Kahneman, P. Slovic, and A. Tversky, editors, Judgement under uncertainty: Heuristics and biases, pages 153-160. Cambridge Univ. Press, 1982.
[176] M. Valtorta, Y.-G. Kim, and J. Vomlel. Soft evidential update for probabilistic multiagent systems. Int. Journ. of Approximate Reasoning, 29(1):71-106, 2002.
[177] D. Varacca. Probability, Nondeterminism and Concurrency: Two Denotational Models for Probabilistic Computation. PhD thesis, Univ. Aarhus, 2003. BRICS Dissertation Series, DS-03-14.
[178] D. Varacca and G. Winskel. Distributing probability over non-determinism. Math. Struct. in Comp. Sci., 16:87-113, 2006.
[179] T. Verma and J. Pearl. Causal networks: semantics and expressiveness. In R. Shachter, T. Levitt, L. Kanal, and J. Lemmer, editors, Uncertainty in Artif. Intelligence, pages 69-78. North-Holland, 1988.
[180] C. Villani. Optimal Transport - Old and New. Springer, Berlin Heidelberg, 2009.
[181] I. Witten, E. Frank, and M. Hall. Data Mining - Practical Machine Learning Tools and Techniques. Elsevier, Amsterdam, 2011.


[^0]:    ${ }^{1}$ Informally, a logic is non-monotonic if adding assumptions may make a conclusion less true.

[^1]:    For instance, I may think that scientists are civilised people, until, at some conference dinner, a heated scientific debate ends in a fist fight.

[^2]:    ${ }^{1}$ Our bind notation $c \gg=\omega$ differs from the one used in the functional programming language Haskell; there one writes the state first, as in $\omega \gg=c$. For us, the channel $c$ acts on the state $\omega$ and is thus written before the argument. This is in line with standard notation $f(x)$ in mathematics, for a function $f$ acting on an argument $x$. Later on, we shall use predicate transformation $c=\ll p$ along a channel, where we also write the channel first, since it acts on the predicate $p$. Similarly, in categorical logic the corresponding pullback (or substitution) is written as $c^{*}(p)$ with the channel before the predicate. The operation 》= works forwardly, in the direction of the channel, whereas $=\ll$ works backwardly, against the direction of the channel.

[^3]:    ${ }^{2}$ See the online overview https://wiki.haskell.org/Monad_tutorials_timeline

[^4]:    2 With thanks to Bas and Bram Westerbaan for help.

[^5]:    ${ }^{1}$ Moreover, in continuous probability there are no inclusions of states in predicates.

[^6]:    ${ }^{2}$ This is implemented for instance in Python's ScyPi library.

[^7]:    ${ }^{2}$ See Section ?? and the dice explanations at Wolfram

[^8]:    1 https://en.wikipedia.org/wiki/German_tank_problem

[^9]:    2 The distribution $\rho$ 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 ??.

[^10]:    1 See e.g.https://www.machinelearningplus.com/predictive-modeling/ how-naive-bayes-algorithm-works-with-example-and-full-code/

