A Review of the Discrete Boltzmann Distribution

Bart Jacobs
iHub, *Radboud University*
Nijmegen, The Netherlands
January 2, 2026

ABSTRACT

The Boltzmann distribution is an iconic probability distribution in physics, but it receives relatively little attention in probabilistic computing. This paper aims to balance this disparity. It reviews the (discrete) Boltzmann distribution from a modern (categorical) perspective. It introduces new constructions and results, such as bi-/tri-/quadri-/ *etc* nomial coefficients, for counting microstates with a certain energy, and two new sufficient statistics results involving energy distributions. It is shown that they are closed under convolution. Along the way the paper introduces multisets in the context of statistical physics, to describe indistinguishable microstates. In the end, Markov chains are defined on microstates (and on multisets), for computing equilibria. They involve some subtleties about entropy. The energy dynamics captured by Boltzmann distributions is of general interest, beyond statistical physics. This paper aims to put it in a wider perspective, demonstrating the commonality with standard probabilistic models, like coins and dices.

KEYWORDS

Boltzmann distribution, entropy, multiset, sufficient statistic, convolution, Markov chain

ACM Reference Format:
Bart Jacobs. 2026. A Review of the Discrete Boltzmann Distribution. In *Proceedings of* . ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Entropy is a concept with a technical origin that has captured the popular imagination. In general terms, entropy refers to chaos and disorder, which increases in spontaneous situations. In [11, Chap. 9] one finds descriptions like: "The universe is running down. It is a degeneratieve one-way street. The final state of maximum entropy is our destiny." Successful popular-science books have been written about the topic of entropy, such as [3; 11]. For more mathematically oriented overview books, see *e.g.* [1; 2; 5; 18; 19].

The concept of entropy emerged in statistical physics, in the 19th and early 20th century via the work of William Rankine, Rudolf Clausius, and Ludwig Boltzmann. The second law of thermodynamics — also known as the law of entropy — says that the entropy of an (isolated) system tends to increase over time until it reaches a maximum at equilibrium. That's why physicists Will try to describe this equilibrium via (a distribution with) maximum entropy. The computational experiments at the end of this paper indicate that the story of maximal entropy involves some nuances.

Entropy became a topic of study for mathematicians and computer scientists through the work of Claude Shannon [22], in the middle of the 20th century. He introduced entropy as a measure of information. Also biologists and chemists use entropy, *e.g.* to study DNA sequences from an information-theoretic perspective. Today, entropy is seen as one of the core concepts of science.

Statistical physics studies particles, in large numbers, via probability distributions. In essence, these are *discrete* distributions on large, but finite sample spaces. Since the numbers involved are big, continuous distributions are often used, arising in the limit. In this paper however, we remain firmly within the discrete world. We investigate what we can compute there, see the many plots below. Particles may have different energies and are thus studied as distributions over discrete energy levels. In this setting these energy levels are simply natural numbers, in a set $\boldsymbol{L} \coloneqq \{0, 1, \ldots, L-1\}$, for some number L. Our starting point (in Theorem 3.2) is the fundamental observation that the discrete Boltzmann distribution is the one with maximal entropy on $\boldsymbol{L}$, given a certain mean. This is an instance of the maximal entropy principle of Jaynes [16; 17].

Next, multiple particles are considered, say N many of them. The obvious sample space is then the N-fold Cartesian product $\boldsymbol{L}^N = \boldsymbol{L} \times \cdots \times \boldsymbol{L}$. Elements of this product $\boldsymbol{L}^N$ are sequences $(i_1, \ldots, i_N)$ of length N, with energies $0 \leq i_i < L$. These are called *microstates* in physics. One can use the N-fold (tensor) product distribution on $\boldsymbol{L}^N$. This product distribution can also be described in terms of maximal entropy, see Section 4

Physicists are not always happy with these microstates as sequences, since they contain too many details. Microstates are considered to be *indistinguishable* when they are permutations of each other. Physicists thus wish to abstract away from the order of the particles in a microstate and are interested only in how many particles live at each energy level (see the checkmarks in Example 8.4). This leads to multiple particles as *multisets* over energy levels. We recall that a multiset is like a set, except that elements can occur multiple times. Alternatively, a multiset is like a list, except that the order of the elements does not matter, only their multiplicities.

Intriguingly, physicists struggle with the difference between lists / sequences / microstates on the one hand, and multisets on the other. Indistinguishability of microstates is discussed *e.g.* in [6, Ex. 1.15–1.18] or [23, p.106]. In [20, §1.6]:

> Thus, the correct way of specifying a microstate of the system is through the distribution numbers $\{n_j\}$, and not through the statement as to "which particle is in which state".

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The distribution numbers in this quote are what we call the multiplicities of the elements in a multiset, see Section 5. The term multiset is rarely used in physics. The book [21, §8.1] is an exception, but 'multiset' is used there as synonym for a 'generic' (or 'true') microstate, in contrast to a 'specific' microstate, which is a lists of particles. What seems fair to say is that physicists do not make a clear distinction between lists and multisets. This paper does make the distinction and shows that doing so has conceptual and practical advantages, but means that entropy is no longer maximal: indeed, as we shall see, quotients reduce entropy, like from lists to multisets.

Mathematicians also wrestle with the concept of a multiset, but maybe more with the difference between sets and multisets. For instance when they say that a matrix has a *set* of eigen values. This should be a *multiset*, since eigenvalues may occur multiple times. They same holds for the roots of a polynomial. The prime factorisation theorem says that each non-zero natural number can be identified with a multiset of prime numbers — but it is never formulated as such.

This is where computer scientists can step in. They are trained to systematically handle different data types, with their different operations and properties. This paper offers a review of the very basics of the theory of particles at different energy levels. It takes the systematic perspective of modern (categorical) probability theory, for a thorough and precise analysis. Its goal is to explore the relevant structures and see what can be computed, in terms of energies and transitions (via Markov chains). The paper will lead to new results, in combinatorics (about trinomial, quadrinomial *etc.* coefficients) and in probability theory (about sufficient statistics). It will also introduce Markov chains on microstates and on multisets, for reaching equilibria as stationary distributions. This involves very basic theory that is relevant beyond particle physics, for instance in computer science or economics [7]. Indeed, particles at different energy levels, with their dynamics, may also be seen as individuals with certain levels of wealth, engaged in economic transactions. Many other such applications can be foreseen where valuable resources are exchanged.

This paper starts by collecting relevant background information on discrete probability distributions. This is applied in Section 3 to the (discrete) Boltzmann distribution. It is described in an original, minimal manner that helps to see what its essential properties are: maximality of its entropy, and invertibility of its mean. This is illustrated in various plots. In a follow-up section these Boltzmann distributions are put in parallel, via a tensor product of distributions on sequences (microstates), again with a maximal entropy property.

There is then another background section, this time on multisets. Next, section 6 has a combinatorial character and contains new formulations and new properties of 'nomial' numbers, generalising binomial, trinomial coefficients etc. These numbers turn out to be crucial for describing and computing energy distributions. Section 8 uncovers two new sufficient statistics situations (see [4; 9; 24]) in the theory of particle energies. Such situations are a big thing in probability theory. They provide an efficient way to summarise / compress via a function, without losing information, since the relevant information can be recovered via an associated channel — typically of the form of a probabilistic inverse or dagger. The final section introduces Markov chains to compute equilibrium distributions on microstates (and in the appendix also on multisets). An elementary example is elaborated where the entropy goes down, as the system evolves towards an equilibrium. Also, somewhat remarkably, these equilibria need not be uniform distributions. This challenges our formalisation, especially of thermal agitation, in relation to prevailing discourses in physics about entropy, and opens up avenues for further research.

The author is not a physicist, but a mathematician / computer scientist. This paper takes inspiration from particle physics, but its developments are not driven by intuitions from physics, but from probability theory and from (theoretical) computer science. Some use of category theory is made, but only superficially, without assuming prior knowledge.

2 BACKGROUND ON DISTRIBUTIONS

This section briefly introduces the basics of (finite, discrete, probability) distributions. We use ket notation $|\cdot\rangle$ to separate multiplicities and elements and write for instance $\frac{1}{2}|H\rangle + \frac{1}{2}|T\rangle$ for the fair coin distribution, with a probability $\frac{1}{2}$ both for head and tail. We write $\mathcal{D}(X)$ for the set of distributions over a set X. The elements of $\mathcal{D}(X)$ may be written in ket form as finite formal convex sums $\sum_i r_i|x_i\rangle$, where $x_i \in X$ and $r_i \in [0,1]$ with $\sum_i r_i = 1$. Equivalently, such distributions can be written in functional form as probability density functions $\omega\colon X \to [0,1]$ with finite support $\mathit{supp}(\omega) \coloneqq \{x \in X \mid \omega(x) \neq 0\}$ and with $\sum_x \omega(x) = 1$.

Functoriality

Given a function $f\colon X \to Y$ and a distribution $\omega \in \mathcal{D}(X)$, one can form an *image distribution* on Y, written as $\mathcal{D}(f)(\omega)$. Formulated in ket form:

$$\mathcal{D}(f)\left(\sum_{x\in X} \omega(x)\big|x\big\rangle\right) \coloneqq \sum_{x\in X} \omega(x)\big|f(x)\big\rangle. \tag{1}$$

This means that $\mathcal{D}$ is a functor $\mathbf{Sets} \to \mathbf{Sets}$ on the category of sets and functions. In fact, $\mathcal{D}$ is a monad on $\mathbf{Sets}$, but we use this fact only implicitly.

Product distributions

Given two distributions $\omega \in \mathcal{D}(X)$ and $\rho \in \mathcal{D}(Y)$ on different sets X, Y, one can form the (parallel) *product distribution* $\omega \otimes \rho \in \mathcal{D}\big(X \times Y\big)$, namely:

$$\omega \otimes \rho \coloneqq \sum_{x\in X,\, y\in Y} \omega(x)\cdot\rho(y)\big|x, y\big\rangle \tag{2}$$

These products can be iterated. For instance, we write $\mathit{iid}[N](\omega) \coloneqq \omega^N = \omega \otimes \cdots \otimes \omega$ for the N-fold product with itself — given as 'independent and identical distributions'.

In general, a distribution $\tau \in \mathcal{D}\big(X \times Y\big)$ on a product set is called a *joint distribution*. Using the obvious projection functions $\pi_1\colon X\times Y \to X$ and $\pi_2\colon X\times Y \to Y$ we can define the *marginals* of τ as image distributions $\mathcal{D}(\pi_1)(\tau) \in \mathcal{D}(X)$ and $\mathcal{D}(\pi_2)(\tau) \in \mathcal{D}(Y)$. One has $\mathcal{D}(\pi_1)\big(\omega\otimes\rho\big) = \omega$ and $\mathcal{D}(\pi_2)\big(\omega\otimes\rho\big) = \rho$, but an arbitrary joint distribution τ differs in general from the product $\mathcal{D}(\pi_1)(\tau) \otimes \mathcal{D}(\pi_2)(\tau)$ of its marginals. When τ happens to be equal to the product of its marginals, it is called independent, non-entwined, non-entangled, or non-correlated.

Mean and variance

For a distribution $\omega \in \mathcal{D}(\mathbb{R})$ on the (real) numbers, or on a subset, we write $\mathit{mean}(\omega) \in \mathbb{R}$ for the mean / average:

$$\mathit{mean}(\omega) \coloneqq \sum_{x \in \mathit{supp}(\omega)} \omega(x) \cdot x = \sum_{x \in \mathbb{R}} \omega(x) \cdot x. \tag{3}$$

The variance of a distribution $\omega \in \mathcal{D}(\mathbb{R})$ describes the extent to which the elements in its support differs from the mean:

$$\begin{aligned} \mathit{var}(\omega) &\coloneqq \sum_{x \in \mathbb{R}} \omega(x) \cdot \big(x^2 - \mathit{mean}(\omega)^2\big) \\ &= \left(\sum_{x \in \mathbb{R}} \omega(x) \cdot x^2\right) - \left(\sum_{x \in \mathbb{R}} \omega(x) \cdot x\right)^2. \end{aligned} \tag{4}$$

For a joint distribution $\tau \in \mathcal{D}\big(\mathbb{R}^N\big)$ the mean is defined as the N-tuple in $\mathbb{R}^N$ of means of its marginals $\mathcal{D}(\pi_i)(\tau) \in \mathcal{D}(\mathbb{R})$, as in:

$$\mathit{mean}(\tau) \coloneqq \Big(\mathit{mean}\big(\mathcal{D}(\pi_1)(\tau)\big), \ldots, \mathit{mean}\big(\mathcal{D}(\pi_N)(\tau)\big)\Big). \tag{5}$$

Channels

A *channel* from a set X to a set Y is a function of the form $c \colon X \to \mathcal{D}(Y)$. Such a channel c maps an element $x \in X$ to a distribution $c(x) \in \mathcal{D}(Y)$ on Y. Channels often occur as conditional probabilities $P(y|x)$, but are used here as probabilistic functions.

A particular channel that we shall use is the *probabilistic inverse* $f^{\sim 1} \colon Y \to \mathcal{D}(X)$, for a surjective function $f \colon X \twoheadrightarrow Y$ between non-empty finite sets X, Y. It is pointwise the uniform distribution:

$$f^{\sim 1}(y) \coloneqq \sum_{x \in f^{-1}(y)} \frac{1}{|f^{-1}(y)|} \,|x\rangle, \tag{6}$$

where we write $|\cdot|$ for the number of elements (the size) of a finite set, in this case of $f^{-1}(y) \coloneqq \{x \in X \mid f(x) = y\}$.

For a channel $c \colon X \to \mathcal{D}(Y)$ and a distribution $\omega \in \mathcal{D}(X)$, one forms the *pushforward distribution* $c_*(\omega) \in \mathcal{D}(Y)$, via:

$$c_*(\omega) \coloneqq \sum_{y \in Y} \left(\sum_{x \in X} \omega(x) \cdot c(x)(y)\right) |y\rangle. \tag{7}$$

In this way one can define the sequential composition $d \mathbin{\circ} c$ with a channel $d \colon Y \to \mathcal{D}(Z)$ as $\big(d \mathbin{\circ} c\big)(x) = d_*(c(x))$. This composition $\circ$ is associative and has channels $\mathit{unit} \colon X \to \mathcal{D}(X)$ as neutral element, with $\mathit{unit}(x) = 1|x\rangle$. Thus, channels form a category, which, in categorical terms, is called the Kleisli category $\mathcal{K}\ell(\mathcal{D})$ of the distribution monad $\mathcal{D}$. An ordinary function $f \colon X \to Y$ can be promoted to a channel $\mathit{unit} \circ f \colon X \to \mathcal{D}(Y)$. We often do this implicitly.

A *Markov chain* is an 'endo' channel $c \colon X \to \mathcal{D}(X)$, with the same domain and codomain. One can then define iterated compositions c, $c^2 = c \mathbin{\circ} c$, $c^3 = c \mathbin{\circ} c \mathbin{\circ} c$, *etc.* A distribution $\omega \in \mathcal{D}(X)$ is called *stationary* or an *equilibrium*, for this channel / Markov chain c, if $c_*(\omega) = \omega$.

Entropy of a distribution

For a distribution $\omega \in \mathcal{D}(X)$ we write $H(\omega) \in \mathbb{R}_{\geq 0}$ for the (Shannon) *entropy*. Intuitively, the entropy is a measure of its uncertainty. We describe it in terms of the natural logarithm ln as:

$$\begin{aligned} H(\omega) &\coloneqq -\sum_{x \in X} \omega(x) \cdot \ln\big(\omega(x)\big) \\ H_\omega(c) &\coloneqq -\sum_{x \in X} \omega(x) \cdot H\big(c(x)\big) \end{aligned} \tag{8}$$

When $\omega(x) = 0$, we understand that the element $x \in X$ does not contribute to these sums. The second definition $H_\omega(c)$ in (8) captures the conditional entropy, for a channel $c \colon X \to \mathcal{D}(Y)$, as average entropy of the distributions $c(x)$, for $x \in X$.

We shall use the following standard facts about entropy, without proof. For details, see *e.g.* [5].

Lemma 2.1. (1) *Zero entropy $H(\omega) = 0$ holds precisely when ω is a point distribution, that is, when $\omega = 1|x\rangle$, for a (unique) element x in its support.*

(2) *For a non-empty finite set X, say with $N > 0$ elements one has $H(\omega) \leq H(\upsilon_X) = \ln(N)$ for the uniform distribution $\upsilon_X = \sum_{x \in X} \frac{1}{N} |x\rangle$ on X.*

(3) *For a joint distribution $\tau \in \mathcal{D}(X \times Y)$ with marginals $\tau_1 \coloneqq \mathcal{D}(\pi_1)(\tau) \in \mathcal{D}(X)$ and $\tau_2 \coloneqq \mathcal{D}(\pi_2)(\tau) \in \mathcal{D}(Y)$,*

$$H(\tau) \leq H(\tau_1) + H(\tau_2).$$

This inequality $\leq$ is an equality $=$ if and only if $\tau = \tau_1 \otimes \tau_2$, i.e. when τ is the product of its marginals — and thus non-entwined / non-entangled.

(4) *Let $f \colon X \to Y$ be an arbitrary function, with a distribution $\omega \in \mathcal{D}(X)$ on its domain. Then:*

$$H\Big(\mathcal{D}(f)(\omega)\Big) \leq H(\omega).$$

The inequality $\leq$ is an equality when f is injective. □

In the last item (4), when f is surjective (and thus not injective), quotienting leads to a decrease of entropy. This is relevant later on, especially when we switch from a product distribution on sequences / microstates, to a multinomial distribution on multisets — in particular in Equation (14). The entropy decrease can be made precise, see Remark 8.5.

3 BACK TO BOLTZMANN

This section starts with a distribution on energy levels that is a simplified version of what is commonly called the (discrete) Boltzmann distribution. We first define this distribution and then formulate and prove its main maximum entropy property. Only then we put things in a wider perspective.

Definition 3.1. For positive numbers $L \in \mathbb{N}_{>0}$ and $t \in \mathbb{R}_{>0}$ we define the Boltzmann distribution $\mathit{bo}[L](t) \in \mathcal{D}(\boldsymbol{L})$ on energy levels $\boldsymbol{L} = \{0, 1, \ldots, L-1\}$ via iterated powers as:

$$\mathit{bo}[L](t) \coloneqq \sum_{0 \leq i < L} \frac{t^i}{Z} \,|i\rangle. \tag{9}$$

The normalisation factor Z ('*Zustandssumme*') is:

$$Z = Z[L](t) \coloneqq \sum_{0 \leq j < L} t^j = \begin{cases} L & \text{if } t = 1 \\ \dfrac{1 - t^L}{1 - t} & \text{if } t \neq 1. \end{cases}$$

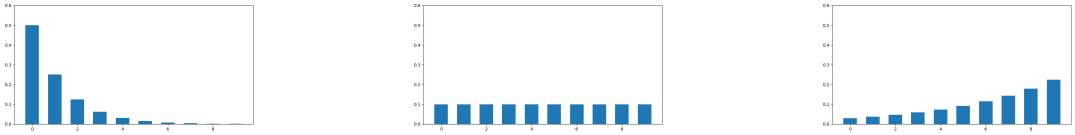


Figure 1: Three bar charts illustrating the Boltzmann distribution $bo[10](t)$ from Definition 3.1, with $L = 10$ energy levels on the horizontal axis, and with level parameter $t = \frac{1}{2}$ on the left, with $t = 1$ in the middle and with $t = \frac{5}{4}$ on the right. These level parameters t correspond, respectively, to temperatures $T \approx 1.44$, $T = \pm\infty$ and $T \approx -4.48$, via the formula $T = \frac{-1}{\ln(t)}$, see Remark 3.3.

We leave the dependence of Z on L, t implicit when it is clear from the context. When $Z[L]$ is seen as a function $\mathbb{R}_{>0} \to \mathbb{R}_{>0}$, it often called the partition function.

We call the number t the level parameter. It is related to temperature, see Remark 3.3 below. It yields a uniform distribution for $t = 1$.

Figure 1 contains three illustrations of the Boltzmann distribution. The next result contains its key property.

Theorem 3.2. *Fix a number $L \in \mathbb{N}_{>0}$. For an arbitrary distribution $\omega \in \mathcal{D}(L)$ one has:*

$$mean(\omega) = mean\big(bo[L](t)\big) \implies H(\omega) \leq H\big(bo[L](t)\big).$$

This says that the Boltzmann distribution has maximal entropy among the distributions with a fixed mean. Later on, in Remark 3.5, we show how to actually obtain a parameter t for a given mean.

Proof. We write $u = mean\big(bo[L](t)\big) = mean(\omega)$.

$$\begin{aligned}
&H\big(bo[L](t)\big) - H(\omega) \\
&\overset{(8)}{=} -\sum_{0\leq i<L} \frac{t^i}{Z} \cdot \ln\left(\frac{t^i}{Z}\right) - H(\omega) \\
&= \ln(Z) - \sum_{0\leq i<L} \frac{t^i \cdot i}{Z} \cdot \ln(t) - H(\omega) \\
&= \ln(Z) - u \cdot \ln(t) - H(\omega) \\
&= \sum_{0\leq i<L} \omega(i) \cdot \ln(Z) - \sum_{0\leq i<L} \omega(i) \cdot i \cdot \ln(t) \\
&\qquad + \sum_{0\leq i<L} \omega(i) \cdot \ln\big(\omega(i)\big) \\
&= \sum_{0\leq i<L} \omega(i) \cdot \left(\ln\big(\omega(i)\big) - \ln\left(\frac{t^i}{Z}\right)\right) \\
&= \sum_{0\leq i<L} \omega(i) \cdot \ln\left(\frac{\omega(i)}{bo[L](t)(i)}\right) = D_{KL}\Big(\omega,\, bo[L](t)\Big) \geq 0.
\end{aligned}$$

In the last line we make use of the Kullback-Leibler divergence D_{KL}, which is always non-negative, see *e.g.* [5]. □

We claim very limited originality for this maximum entropy result — at most for its formulation as an intrinsic property of the Boltzmann distribution and for its proof. In physics this distribution is usually constructed from a fixed mean, with maximal entropy as goal, via the Lagrange multiplier method, see *e.g.* [16] or [6, Chap. 10]. The distribution that then arises involves e-powers, see the discussion below. Our 'simple' formulation of the Boltzmann distribution avoids these e-powers but still satisfies maximality.

Remark 3.3. *In Definition 3.1 we use a level parameter $t \in \mathbb{R}_{>0}$. Physicists may like to read it as:*

$$t = e^{-\beta} \qquad \text{or as} \qquad t = e^{-\frac{1}{k_B \cdot T}}, \tag{10}$$

where k_B is Boltzmann's constant and $T \in \mathbb{R} \setminus 0$ stands for temperature. We prefer a simple letter t as parameter, instead of these more complicated e-powers, to keep things as simple as possible. One obtains the traditional e-power formulation of the Boltzmann distribution, written here as $\underline{bo}[L](T)$, via a substitution instance of (9)*:*

$$\underline{bo}[L](T) \coloneqq bo[L]\Big(e^{-\frac{1}{T}}\Big) = \sum_{0\leq i<L} \frac{e^{-\frac{i}{T}}}{Z} \big|i\big\rangle,$$

where now $Z = \sum_i e^{-\frac{i}{T}}$. Here we ignore the Boltzmann constant k_B — which is common in mathematical accounts.

The letter t that we use is intentionally chosen for the connection to temperature, formally as $T = \frac{-1}{\ln(t)} \in \mathbb{R} \setminus 0$. Via the equations (10) *one obtains, as suggested in Figure 1,*

$$\begin{aligned}
T > 0 &\iff t \in (0,1) \iff \textit{descending probabilities} \\
T < 0 &\iff t \in (1,\infty) \iff \textit{ascending probabilities.}
\end{aligned}$$

One may add that the bigger $T > 0$ is, the smaller $t \in (0,1)$, and the stronger the descent. Similarly for negative temperature $T < 0$. Systems with a positive temperature will absorb energy, whereas a negative temperature characterises systems that tend to give off energy. Working with $t \in \mathbb{R}_{>0}$ avoids the undefinedness for $T = 0$.

We add three side remarks.

1. *We could allow the border cases $t = 0$ and $t = \infty$ in Definition 3.1, so that the point distributions $1|0\rangle = bo[L](0)$ and $1|L-1\rangle = bo[L](\infty)$ are also Boltzmann distributions. In the current set-up these point distributions can only be approximated.*
2. *The maximum entropy property exists also beyond finite distributions: for infinite distributions (on $\mathbb{N}$) this property holds for the geometric distribution, and for continuous distributions (on $\mathbb{R}_{>0}$) it holds for the exponential distribution (see [2, §3.2.4.3]).*
3. *The maximality in Theorem 3.2 assumes equal means. This can be extended to n-ary moments, by using an adapted distribution $bo_n[L](t) \coloneqq \sum_{0\leq i<L} \frac{t^{(i^n)}}{Z_n} |i\rangle$, with obvious normaliser Z_n.*

We turn to some basic properties of the mean (energy) and variance of the Boltzmann distribution, see Figure 2 for relevant pictures. Physicists like to use derivatives in this context, especially of the partition function.

Lemma 3.4. (1) *The mean energy of the Boltzmann distribution* (9) *can be expressed as derivative of the partition function $Z[L] : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, namely as:*

$$mean\big(bo[L](t)\big) = \sum_{0\leq i<L} \frac{i \cdot t^i}{Z[L](t)} = \frac{t}{Z} \cdot \frac{\partial}{\partial t} Z[L].$$

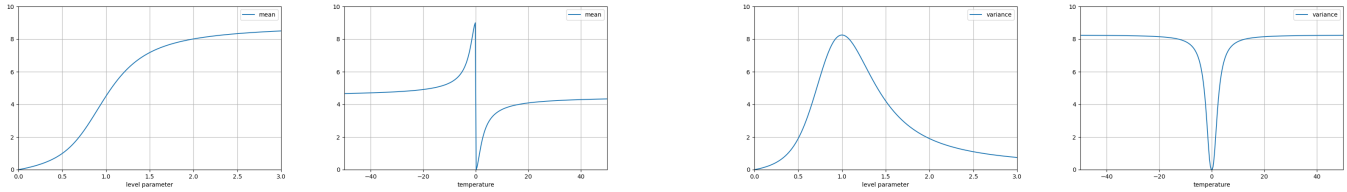


Figure 2: On the left two plots of the mean energy of the Boltzmann distributions, $mean\big(bo[10](t)\big)$ and $mean\big(\underline{bo}[10](T)\big)$, as function of the level parameter $t \in (0,3)$ and the temperature $T \in (-50,50)$, with $L = 10$ energy levels. The limit is $L-1 = 9$. The mean at $t = 1$ is $\frac{9}{2}$, for the uniform Boltzmann distribution. The two limits $T \to \pm\infty$ both go to $\frac{9}{2}$, for the uniform Boltzmann distribution, with the maximum mean reached as $T \uparrow 0$ and the minimum as $T \downarrow 0$. On the right, analogues for the energy variances $var\big(bo[10](t)\big)$ and $var\big(\underline{bo}[10](T)\big)$. The highest variance is reached for level parameter $t = 1$, corresponding to temperature $T = \pm\infty$, where the Boltzmann distribution is uniform. This highest variance value is $\frac{L^2-1}{12} = \frac{33}{4} = 8.25$. The variance becomes zero when the temperature approaches zero, from both sides.

This mean lies in the interval $\big(0, L-1\big) \subseteq \mathbb{R}_{>0}$. The mid-value $\frac{L-1}{2}$ is the mean of the uniform distribution, for $t = 1$.

(2) *The energy variance of the Boltzmann distribution can be expressed as derivative of its mean:*

$$var\big(bo[L](t)\big) \;=\; t \cdot \frac{\partial}{\partial t}\, mean\big(bo[L](t)\big).$$

This variance lies in $\left(0, \frac{L^2-1}{12}\right] \subseteq \mathbb{R}_{>0}$, where the top value is reached by the uniform distribution.

(3) *The heat capacity $\underline{hc}[L](T)$ and $hc[L](t)$, standardly defined as the derivative of the mean energy with respect to the temperature, can be expressed as:*

$$\underline{hc}[L](T) \;\coloneqq\; \frac{\partial}{\partial T}\, mean\big(\underline{bo}[L](T)\big) \;=\; \frac{var\big(\underline{bo}[L](T)\big)}{T^2}.$$

By substituting $T = \frac{-1}{\ln(t)}$ this heat capacity becomes, in terms of a level parameter $t \in \mathbb{R}_{>0}$,

$$hc[L](T) \;=\; \ln(t)^2 \cdot var\big(bo[L](t)\big). \qquad \square$$

The heat capacity describes how much heat (transfer of energy) a system can absorb/release for a temperature change.

Remark 3.5. *In the diagram on the left in Figure 2 we see a relative simple mean-of-Boltzmann function $mean\big(bo[L](-)\big) : \mathbb{R}_{>0} \to \big(0, L-1\big)$. Just looking at the graph of this function, we see that there should be an inverse. There is no known formula for this inverse, but there is a way to compute it by solving a polynomial equation. This works as follows. Given a (mean) energy value $u \in \mathbb{R}_{>0}$, we have:*

$$\begin{aligned} mean\Big(bo[L](t)\Big) = u \;&\Longleftrightarrow\; \sum_{0 \leq i < L} i \cdot t^i = \sum_{0 \leq i < L} u \cdot t^i \\ &\Longleftrightarrow\; u + (u-1) \cdot t^1 + (u-2) \cdot t^2 \\ &\qquad + \cdots + (u-L+1) \cdot t^{L-1} \;=\; 0. \end{aligned}$$

This equation can be solved computationally[1]. For instance, in the setting of Figure 2 with $L = 10$, for $u = 2$ we get as inverse $t \approx 0.6937$ and for $u = 5$ we get $t \approx 1.0629$. Doing this systematically yields the plot on the left in Figure 3. This same polynomial-solution approach is used in [6, Ex. 5.3] to obtain a (Boltzmann) die distribution (for $L = 6$) with a maximal entropy from a mean.

[1]We use Python's `nroots` library.

Once we have have recovered the level parameter t from a given mean u, we can also find the corresponding temperature T, as $T = \frac{-1}{\ln(t)}$. This yields an inverse for the second plot from the left in Figure 2, see the second picture from the left in Figure 3. Notice that as long as the mean is below the mid-value $\frac{9}{2}$ the temperature is positive, but it is negative for means above this mid-value. This corresponds to pictures used in physics, see for instance [6, Fig. 12.3, middle right]. The two plots on the right in Figure 3, for the Boltzmann entropy and heat capacity as a function of the energy mean, correspond to the picture in [6, Fig. 12.3, upper and lower right]. There they are derived for two energy levels only (i.e. for $L = 2$).

Remark 3.6. *The thermodynamic definition of temperature T happens via its reciprocal $\frac{1}{T}$, which is defined as the derivative of the entropy, with respect to energy. This is commonly written as $\frac{1}{T} = \frac{\partial S}{\partial U}$, where S is the entropy and U is the energy, see e.g. [6, Chap. 12] or any other textbook. We briefly show how this equation that defines temperature emerges in the current context with the Boltzmann distribution.*

In the previous remark we have seen how the mean-energy function $mean\big(bo[L](-)\big) : \mathbb{R}_{>0} \to \big(0, L-1\big)$ is an isomorphism. Let's write its inverse as a function $t : \big(0, L-1\big) \to \mathbb{R}_{>0}$. The associated temperature function $T : \big(0, L-1\big) \to \mathbb{R}$ is then $T(u) = \frac{-1}{\ln(t(u))}$. We now take the energy-derivative of entropy of the Boltzmann distribution:

$$\begin{aligned} &\frac{\partial}{\partial u}\, H\big(bo[L](t(u))\big) \\ &= \frac{\partial}{\partial u} - \sum_{0 \leq i < L} \frac{t(u)^i}{Z[L](t(u))} \cdot \ln\left(\frac{t(u)^i}{Z[L](t(u))}\right) \\ &= \frac{\partial}{\partial u}\, \ln\Big(Z[L](t(u))\Big) - \sum_{0 \leq i < L} \frac{t(u)^i \cdot i}{Z[L](t(u))} \cdot \ln\big(t(u)\big) \\ &= \frac{1}{Z} \cdot \sum_{0 \leq i < L} i \cdot t(u)^{i-1} \cdot t'(u) - \frac{\partial}{\partial u}\, u \cdot \ln\big(t(u)\big) \\ &= \frac{u \cdot t'(u)}{t(u)} - \ln\big(t(u)\big) - \frac{u \cdot t'(u)}{t(u)} \;=\; -\ln\big(t(u)\big) \;=\; \frac{1}{T(u)}. \end{aligned}$$

This closes the circle and concludes our account of the Boltzmann distribution, in which we have introduced it in a simple form, with a level parameter t only and no e-powers. This leads to simpler formulas, plots, and solutions of polynomial equations.

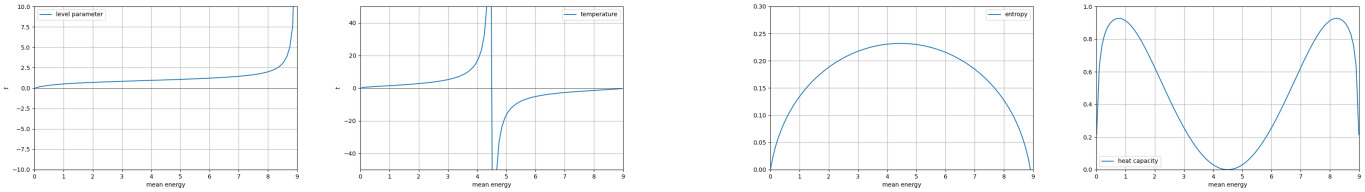


Figure 3: On the left one sees two plots of the inverses of the mean energy functions from the two pictures on the left in Figure 2, for $L = 10$ energy levels. These inverses are computed pointwise for 90 mean values $0.05, 0.15, \ldots, 8.95$ from the interval $\big(0, L-1\big)$ on the horizontal axis. The solutions are computed as root of a polynomial equation, as described in Remark 3.5, and are then drawn as continuous lines. We do not have an explicit formula for these inverses. On the right there are plots that map the mean energy to the entropy and to the heat capacity of the associated Boltzmann distribution. The standard entropy bump is obtained via the steps $u \mapsto t(u) \mapsto H\big(bo[L]\big(t(u)\big)\big)$. The entropy is highest for the mid energy mean $\frac{L-1}{2} = \frac{9}{2}$, corresponding to the uniform Boltzmann distribution (with $t = 1$). The reciprocal temperature $\frac{1}{T}$ arises as the derivative of the entropy, with respect to energy, see Remark 3.6. One can see from the bump shape that the reciprocal of the derivative in the third picture yields the second one. The right-most picture shows the typical shape of the heat (transfer) capacity as a function of the mean energy, obtained via $u \mapsto t(u) \mapsto hc[L]\big(t(u)\big)$. When the mean is low, the capacity to absorb energy is high, and when the mean is high, the capacity to release energy is also high. This is nicely symmetric.

4 MULTIPLE PARTICLES

The Boltzmann distribution $bo[L](t)$ on $\boldsymbol{L}$ introduced in the previous section describes the energy level probabilities of one particle, for a given level parameter t (or temperature T, or mean energy u). When we have N particles, it makes sense to (first) consider them as a list and use the associated tensor product probability $iid[N]\big(bo[L](t)\big) = bo[L](t)^N = bo[L](t) \otimes \cdots \otimes bo[L](t)$.

Proposition 4.1. *Fix numbers $L, N \in \mathbb{N}_{>0}$ and $t \in \mathbb{R}_{>0}$.*

(1) *The N-fold parallel tensor product is:*

$$iid[N]\big(bo[L](t)\big) \;=\; bo[L](t)^N \;=\; \sum_{\vec{i} \in \boldsymbol{L}^N} \frac{t^{sum(\vec{i})}}{Z^N} \big|\vec{i}\big\rangle,$$

with $Z = \sum_{0 \leq i < L} t^i$ as in Definition 3.1.

(2) *For a joint distribution $\omega \in \mathcal{D}\big(\boldsymbol{L}^N\big)$,*

$$mean\big(\omega\big) = (u, \ldots, u) = mean\big(bo[L](t)^N\big) \;\Longrightarrow\; H(\omega) \leq H\big(bo[L](t)^N\big).$$

One can also take the temperature version $\underline{bo}[L](T)^N$ of this parallel Boltzmann distribution, involving e-powers of the total energy $sum(\vec{i})$ of a microstate $\vec{i}$.

Proof. The first point holds since:

$$\begin{aligned} bo[L](t)^N &\overset{(2)}{=} \sum_{\vec{i} \in \boldsymbol{L}^N} \prod_{1 \leq n \leq N} bo[L](t)(i_n) \big|\vec{i}\big\rangle \\ &= \sum_{\vec{i} \in \boldsymbol{L}^N} \prod_{1 \leq n \leq N} \frac{t^{i_n}}{Z} \big|\vec{i}\big\rangle \;=\; \sum_{\vec{i} \in \boldsymbol{L}^N} \frac{t^{sum(\vec{i})}}{Z^N} \big|\vec{i}\big\rangle. \end{aligned}$$

The assumption $mean\big(\omega\big) = (u, \ldots, u) = mean\big(bo[L](t)^N\big)$ in the second point yields $mean\big(\mathcal{D}(\pi_i)(\omega)\big) = u = mean\big(bo[L](t)\big)$ by (5). Hence $H\big(\mathcal{D}(\pi_i)(\omega)\big) \leq H\big(bo[L](t)\big)$ by Theorem 3.2. But then we are done by Lemma 2.1 (3):

$$\begin{aligned} H\big(\omega\big) &\leq \textstyle\sum_i H\big(\mathcal{D}(\pi_i)(\omega)\big) \\ &\leq \textstyle\sum_i H\big(bo[L](t)\big) \;=\; H\big(bo[L](t)^N\big). \end{aligned}$$ □

The good thing about the parallel Boltzmann distribution $bo[L](t)^N$ is that it assigns the same probability to sequences (microstates) $\vec{i}$ with the same total energy $sum(\vec{i})$, see item (1). This matches a fundamental postulate. Another good thing is that its entropy is maximal — among joint distributions whose marginals all have the same mean.

What is not so good about this product distribution is that it involves microstates as sequences. As discussed in the first two sections, from a physical perspective one likes to identify (not-distinguish) sequences up-to-permutation, that is, when they accumulate to the same multiset.

5 BACKGROUND ON MULTISETS

For an arbitrary set X, a multiset over X is an expression of the form $n_1|x_1\rangle + \cdots + n_k|x_k\rangle = \sum_i n_i|x_i\rangle$. It involves elements $x_i \in X$ with associated multiplicities $n_i \in \mathbb{N}$. One can equivalently write such a multiset as a function $\varphi \colon X \to \mathbb{N}$ with finite support: the set $supp(\varphi) \coloneqq \{x \in X \mid \varphi(x) \neq 0\}$ is required to be finite. Thus we can write $\varphi = \sum_x \varphi(x)\big|x\big\rangle$.

The *size* of a multiset is the total number of its elements, including multiplicities. In general, we write $\|\varphi\| \coloneqq \sum_x \varphi(x)$ for the size of a multiset φ. We shall also write $\mathcal{M}(X)$ for the set of all multisets over X, and $\mathcal{M}[N](X) \subseteq \mathcal{M}(X)$ for the subset of multisets of size $N \in \mathbb{N}$. For $N = 0$, the set $\mathcal{M}[N](X)$ has precisely one member, namely the empty multiset $\mathbf{0}$ with zero elements.

Accumulation, sums and totals

There is an obvious way to turn a list of elements into a multiset, simply by forgetting the order, but counting the multiplicities. This operation is called accumulation and written as $acc \colon X^N \to \mathcal{M}[N](X)$. For instance $acc(a, b, c, c, b, b) = 1|a\rangle + 3|b\rangle + 2|c\rangle$.

We shall often use sequences and multisets over a set of numbers $\boldsymbol{L} = \{0, 1, \ldots, L-1\}$, for $L \in \mathbb{N}_{>0}$. We have already seen the addition operation $sum \colon \boldsymbol{L}^N \to \{0, 1, \ldots, (L-1) \cdot N\}$ that takes the sum of a sequence of N numbers in $\boldsymbol{L}$. There is an analogue $tot \colon \mathcal{M}[N](\boldsymbol{L}) \to \{0, 1, \ldots, (L-1) \cdot N\}$ that takes the 'total' amount of a multiset, via $tot(\varphi) \coloneqq \sum_{i \in \boldsymbol{L}} \varphi(i) \cdot i$. For instance,

$tot\big(1|0\rangle + 2|2\rangle + 1|3\rangle + 3|4\rangle\big) = 19$. There is a commuting diagram:

$$\begin{array}{ccc} \boldsymbol{L}^N & \xrightarrow{\ acc\ } & \mathcal{M}[N](\boldsymbol{L}) \\ & \searrow^{sum} & \downarrow{\scriptstyle tot} \\ & & \{0, 1, \ldots, (L-1)\cdot N\} \end{array} \tag{11}$$

We collect some basic (combinatorial) properties of multisets.

Lemma 5.1. *Let X be a finite set of size $L \coloneqq |X| \geq 1$.*

(1) *The set $\mathcal{M}(X)$ is the free commutative monoid on X, with pointwise addition of multisets: $(\varphi+\psi)(x) = \varphi(x)+\psi(x)$, and with the empty multiset $\mathbf{0} \in \mathcal{M}(X)$ as neutral element.*

(2) *The number of multisets over X of size N is given by the multichoose coefficient $\left(\binom{L}{N}\right)$, that is:*

$$\Big|\mathcal{M}[N](X)\Big| = \left(\!\!\binom{L}{N}\!\!\right) = \binom{L+N-1}{N}.$$

(3) *For an arbitrary multiset $\varphi \in \mathcal{M}[N](X)$, the number of sequences / microstates $\vec{x} \in X^N$ with $acc(\vec{x}) = \varphi$ is equal to the multiset coefficient $\multinom{\varphi}$, defined as:*

$$\multinom{\varphi} \coloneqq \frac{\|\varphi\|!}{\prod_x \varphi(x)!} = \frac{N!}{\prod_x \varphi(x)!}. \tag{12}$$

(4) *For the sum of these multiset coefficients one has:*

$$\sum_{\varphi\in\mathcal{M}[N](X)} \multinom{\varphi} = L^N.$$

(5) *For $X = \boldsymbol{L}$, the sum of totals is:*

$$\sum_{\varphi\in\mathcal{M}[N](\boldsymbol{L})} tot(\varphi) = \frac{(L-1)\cdot N}{2} \cdot \left(\!\!\binom{L}{N}\!\!\right).$$

(6) *Using multisets there is a snappy formulation of the Multinomial Theorem, namely as:*

$$\Big(x_0 + \cdots + x_{L-1}\Big)^N = \sum_{\varphi\in\mathcal{M}[N](\boldsymbol{L})} \multinom{\varphi} \cdot \prod_{0\leq i<L} x_i^{\varphi(i)}. \qquad \square$$

The multiset coefficient $\multinom{\varphi}$ occurs frequently in statistical mechanics and is then often written as W. This coefficient can be used to describe the probabilistic inverse (6) of the (surjective) accumulation function $acc\colon X^N \to \mathcal{M}[N](X)$. We call it *arrangement*, written as *arr*, and define it, via Lemma 5.1 (3), on $\varphi \in \mathcal{M}[N](X)$ as:

$$arr(\varphi) \coloneqq acc^{-1}(\varphi) = \sum_{\vec{x}\in acc^{-1}(\varphi)} \frac{1}{\multinom{\varphi}} \big|\vec{x}\big\rangle. \tag{13}$$

Thus, $arr(\varphi)$ is the uniform distribution of all microstates that accumulate to the multiset φ. They should not be distinguished. By Lemma 2.1 (2) we have as entropy $H\big(arr(\varphi)\big) = \ln\big(\multinom{\varphi}\big)$. Here one may recognise an instance of Boltzmann's famous entropy formula $S = \ln(W)$, where S is the entropy and W is the number of microstates (accumulating to φ).

The multinomial distribution

One can see a distribution $\omega \in \mathcal{D}(X)$ as an abstract urn, where X is the set of colours and $\omega(x) \in [0,1]$ gives the probability of drawing a ball of colour $x \in X$. The product distribution $\omega^N = \omega \otimes \cdots \otimes \omega \in \mathcal{D}\big(X^N\big)$ captures the probabilities associated with a sequence $\vec{x} \in X^N$ of draws, where the order matters. This is the probability of $\vec{x}$ as a microstate. If however, we wish to draw multisets — microstate up-to indistinguishability — from ω, we need to use the image distribution, along the accumulation function $acc\colon X^N \to \mathcal{M}[N](X)$. This yields the so-called *multinomial distribution* $mn[N](\omega) \in \mathcal{D}\big(\mathcal{M}[N](X)\big)$, described as:

$$\begin{aligned} mn[N](\omega) &\coloneqq \mathcal{D}(acc)\big(\omega^N\big) \\ &= \sum_{\varphi\in\mathcal{M}[N](X)} \multinom{\varphi} \cdot \prod_{x\in X} \omega(x)^{\varphi(x)} \big|\varphi\big\rangle. \end{aligned} \tag{14}$$

There is one property that we wish to make explicit, as background for subsequent analogous results.

Theorem 5.2. *The accumulation function $acc\colon X^N \to \mathcal{M}[N](X)$ is a sufficient statistic for the identical and independent distribution, as described by the string diagram on the left in Figure 4. As equation it amounts to:*

$$\langle acc, id\rangle_*\Big(\omega^N\Big) = \langle id, arr\rangle_*\Big(mn[N](\omega)\Big). \qquad \square$$

The fact that a map is a sufficient statistic is a fundamental property in probability theory. It means that the identifications introduced by this map can be undone, for a particular distribution. This undoing for *acc* happens via its probabilistic inverse $acc^{\sim 1} = arr$. The general description of sufficient statistics situations in terms of string diagrams comes from [10]. The fact that accumulation is such a sufficient statistic captures a fundamental relationship between lists, multisets and distributions, see [13] for more details.

We can apply the multinomial distribution to a Boltzmann distribution, as 'urn', from which one draws N particles in the form of a multiset of size N. This gives, basically as in [21, Eqn. (8.46)],

$$mn[N]\Big(bo[L](t)\Big) = \sum_{\varphi\in\mathcal{M}[N](\boldsymbol{L})} \frac{\multinom{\varphi}\cdot t^{tot(\varphi)}}{Z^N} \big|\varphi\big\rangle \tag{15}$$

This distribution does not assign the same probability to multisets with the same totals, since there is a factor $\multinom{\varphi}$ involved. Also, this multinomial distribution (15) does not have maximal entropy — in a certain class of distributions — since the multinomial is an image distribution (14) and images reduce entropy, see Lemma 2.1 (4). Remark 8.5 contains the precise entropy reduction with respect to the product distribution. Still, (15) is the obvious distribution if one wishes to use N-ary microstates up-to indistinguishability.

6 COMBINATORIAL INTERMEZZO

In the previous section we have introduced multisets and counted how many sequences accumulate to a specific multiset φ, namely $\multinom{\varphi}$, see Lemma 5.1 (3). In this section we wish to count sequences and multisets with a given sum / total. This leads to new combinatorial results.

Definition 6.1. For numbers $L, N \in \mathbb{N}_{>0}$ with $0 \leq u \leq (L-1)\cdot N$ we define *L-nomials* as:

$$\begin{aligned} C_L(N,u) &\coloneqq \Big|\big\{\vec{i} \in \boldsymbol{L}^N \;\big|\; sum(\vec{i}) = u\big\}\Big| \\ &= \sum_{\varphi\in\mathcal{M}[N](\boldsymbol{L}),\, tot(\varphi)=u} \multinom{\varphi}. \end{aligned} \tag{16}$$

These numbers $C_L(N,u)$ generalise binomial coefficients to trinomial, quadrinomial, *etc.* For $L = 2$ one has $C_2(N,u) = \binom{N}{u}$. We

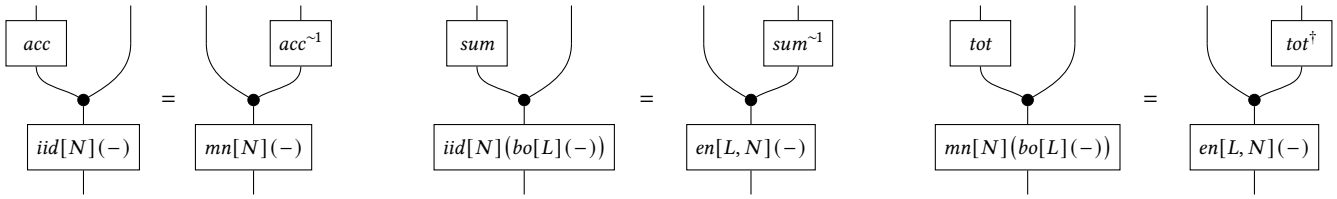


Figure 4: Three string-diagrammatic descriptions of sufficient statistics situations, in Theorem 5.2, 8.1 and 8.2. The first one is known, the other two about Boltzmann and energy distributions are new.

recall that the binomial coefficient $\binom{N}{u} \in \mathbb{N}$ gives the number of subsets of size u, of a (finite) set of size N. Such a subset of size u, say of a set $\{x_1, \ldots, x_N\}$ of N elements, can be identified with sequence of binary numbers $(b_1, \ldots, b_N) \in \{0,1\}^N$ of length N with numbers $b_i \in \{0,1\}$ satisfying $u = sum(b_1, \ldots, b_N) = b_1 + \cdots + b_N$. Each number b_i then tells if the element x_i is in the subset (when $b_i = 1$) or not (when $b_i = 0$). This description is generalised above, from $L = 2$ with $\boldsymbol{L} = \{0,1\}$ to sequences $\vec{i} \in \boldsymbol{L}^N$ for arbitrary L. In physical terms, this number $C_L(N,u)$ counts the number of microstates with N particles at energy levels from $\boldsymbol{L} = \{0,1,\ldots,L-1\}$, with combined energy equal to u. This number $C_L(N,u)$ can also be determined via multiset coefficients, as in the second line of (16), using Lemma 5.1 (3).

Lemma 6.2. *Let numbers $L \geq 1$, $N \geq 0$ be given.*

(1) *One has $C_L(1,u) = 1$ and a recursion relation:*

$$C_L(N{+}1,u) \;=\; \sum_{0 \leq v \leq u} C_L(N, u{-}v),$$

which is useful to compute nomial coefficients efficiently.

(2) *Nomial coefficients are closed under reversal:*

$$C_L(N,u) \;=\; C_L\Big(N,\; (L{-}1)\cdot N - u\Big).$$

This generalises $\binom{N}{u} = \binom{N}{N-u}$.

(3) *Nomials $C_L(N,-)$ add up in the following way.*

$$\begin{aligned} \sum_{0 \leq u \leq (L-1)\cdot N} C_L(N,u) \;&=\; L^N \\ \sum_{0 \leq u \leq (L-1)\cdot N} C_L(N,u)\cdot u \;&=\; \frac{(L-1)\cdot N}{2} \cdot L^N \\ &=\; \sum_{\vec{i} \in \boldsymbol{L}^N} sum(\vec{i}). \end{aligned}$$

This generalises $\sum_{0 \leq u \leq N} \binom{N}{u} = 2^N$ in the binary case.

(4) *These nomials satisfy a Vandermonde property: for each $0 \leq u \leq (L-1)\cdot N$, if $N = N_1 + N_2$, then:*

$$C_L(N,u) \;=\; \sum_{\substack{0 \leq u_1 \leq (L-1)\cdot N_1,\\ 0 \leq u_2 \leq (L-1)\cdot N_2,\\ u_1 + u_2 = u}} C_L(N_1,u_1)\cdot C_L(N_2,u_2).$$

(5) *When $u < L$ the nomial formula simplifies to the multichoose coefficient:*

$$C_L(N,u) \;=\; \left(\!\!\binom{N}{u}\!\!\right) \;=\; \binom{N-u+1}{u}. \qquad \square$$

What we call nomials in (16) is a new implementation. The next result shows that it satisfies a specification for bi / tri / etc. nomials, in terms of polynomial expressions, occurring on the OEIS website [25]. These nomial coefficients are not well-known and studied in the literature, but they are very relevant and useful in (the current setting inspired by) statistical physics. They generalise both binomial coefficients (for $L = 2$) and multichoose coefficients (for suitably large L).

Theorem 6.3. *For $N \geq 1$ and $K \geq 0$ one has, for an arbitrary variable x,*

$$\left(\sum_{0 \leq i < L} x^i\right)^N \;=\; \sum_{0 \leq u \leq (L-1)\cdot N} C_L(N,u)\cdot x^u.$$

Proof. We use multiset formulation of the Multinomial Theorem from Lemma 5.1 (6) in the first step:

$$\begin{aligned} \left(\sum_{0 \leq i < L} x^i\right)^N \;&=\; \sum_{\varphi \in \mathcal{M}[N](\boldsymbol{L})} \binom{\varphi}{} \cdot \prod_{0 \leq i < L} \big(x^i\big)^{\varphi(i)} \\ &=\; \sum_{\varphi \in \mathcal{M}[N](\boldsymbol{L})} \binom{\varphi}{} \cdot x^{tot(\varphi)} \\ &=\; \sum_{0 \leq u \leq (L-1)\cdot N} \;\sum_{\varphi \in \mathcal{M}[N](\boldsymbol{L}),\, tot(\varphi) = u} \binom{\varphi}{} \cdot x^u \\ &\overset{(16)}{=}\; \sum_{0 \leq u \leq (L-1)\cdot N} C_L(N,u)\cdot x^u. \end{aligned}$$

□

Corollary 6.4.

$$\begin{aligned} &\sum_{0 \leq u \leq (N-1)\cdot K} C_L(N,u)\cdot u \cdot x^u \\ &\quad=\; N \cdot \left(\sum_{0 \leq i < L} x^i\right)^{N-1} \cdot \left(\sum_{0 \leq i < L} i \cdot x^i\right). \end{aligned}$$

Proof. Take the derivative $\frac{\partial}{\partial x}$ on both sides of the equation in Theorem 6.3 and multiply with x. □

7 THE CANONICAL ENERGY DISTRIBUTION

We now introduce the energy distribution in three different but equivalent ways. It turns out that the nomial coefficients introduced in the previous section can be put to good use for what is called the canonical distribution in physics.

Definition 7.1. For numbers $L, N \in \mathbb{N}_{>0}$ and $t \in \mathbb{R}_{>0}$ we use the sum and total functions from (11) to define the energy distribution $en[L,N](t) \in \mathcal{D}\big(\{0,\ldots,(L-1)\cdot N\}\big)$ as:

$$\begin{aligned} en[L,N](t) \;&\coloneqq\; \mathcal{D}\big(sum\big)\Big(bo[L](t)^N\Big) \\ &=\; \mathcal{D}\big(tot\big)\Big(mn[N]\big(bo[L](t)\big)\Big) \\ &\overset{(*)}{=}\; bo[L](t) + \cdots + bo[L](t). \end{aligned}$$

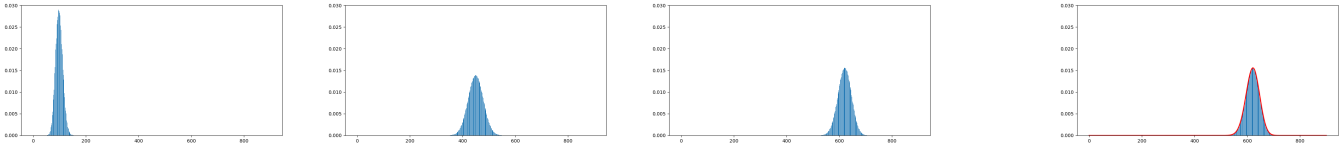


Figure 5: On the left three energy distributions are plotted, each with $L = 10$ energy levels and $N = 100$ particles, with energies $0, 1, \ldots, 900$ on the horizontal axis, where $900 = (L-1) \cdot N$. The level parameter values are respectively $t = \frac{1}{2}$, $t = 1$ and $t = \frac{5}{4}$, like in Figure 1. Accordingly, the means are $N = 100$ times higher. In the uniform case $t = 1$ the energy is nicely centred around the mid energy value 450. On the right, the last energy distribution is combined with a continuous Beta distribution, stretched to the interval $[0, 900]$, with the mean and variance matching with the energy distribution. This turns out to give a good match. The Beta parameters are $\alpha \approx 183$ and $\beta \approx 82$, in this case.

This last line describes the energy distribution as an N-fold convolution of the Boltzmann distribution. We postpone this perspective until after Theorem 7.4.

Figure 5 contains several plots for this energy distribution.

Proposition 7.2. *In the context, of Definition 7.1,*

(1) *Concretely, the energy distribution is:*

$$en[L,N](t) \;=\; \sum_{0\leq u\leq (L-1)\cdot N} \frac{C_L(N,u)\cdot t^u}{\sum_v C_L(N,v)\cdot t^v}\,\big|u\big\rangle.$$

(2) *The mean and variance of the energy distribution are a multiple of the mean and variance of Boltzmann:*

$$\begin{aligned} mean\big(en[L,N](t)\big) &= N\cdot mean\big(bo[L](t)\big) \\ var\big(en[L,N](t)\big) &= N\cdot var\big(bo[L](t)\big). \end{aligned}$$

Proof. First, by Proposition 4.1 (1),

$$\begin{aligned} en[L,N](t) &= \mathcal{D}\big(sum\big)\Big(bo[L](t)^N\Big) \\ &\overset{(1)}{=} \sum_{\vec{i}\in L^N} \frac{t^{sum(\vec{i})}}{Z^N}\big|sum(\vec{i})\big\rangle \\ &= \sum_{0\leq u\leq (L-1)\cdot N}\; \sum_{\vec{i}\in sum^{-1}(u)} \frac{t^u}{(\sum_{0\leq i<L} t^i)^N}\,\big|u\big\rangle \\ &\overset{(16)}{=} \sum_{0\leq u\leq (L-1)\cdot N} \frac{C_L(N,u)\cdot t^u}{\sum_v C_L(N,v)\cdot t^v}\,\big|u\big\rangle. \end{aligned}$$

The final step uses Theorem 6.3 in the denominator.

For point (2) we do only the mean, using Corollary 6.4:

$$\begin{aligned} mean\big(en[L,N](t)\big) &= \sum_{0\leq u\leq (L-1)\cdot N} \frac{C_L(N,u)\cdot u\cdot t^u}{Z^N} \\ &= N\cdot \sum_{0\leq i<L} \frac{t^i\cdot i}{Z} \\ &= N\cdot mean\big(bo[L](t)\big). \end{aligned}$$

□

Remark 7.3. *In Remark 3.5 we have described how we can go from a mean energy level, for the Boltzmann distribution, to a level parameter t — or equivalently, to a temperature T — by solving a polynomial equation. The same can now be done for the new energy distribution, since its mean is a multiple of the Boltzmann mean, see Proposition 7.2 (2).*

Explicitly, when the level and particle numbers $L, N \in \mathbb{N}_{>0}$ are fixed, then, for a mean energy level $u \in \big(1, (L-1)\cdot N\big)$ we obtain $\frac{u}{N} \in (0, L-1)$. Then we can find a level parameter $t \in \mathbb{R}_{>0}$ — or temperature T — with $mean\big(bo[L](t)\big) = \frac{u}{N}$. In this way we obtain an energy-mean equal to u, in $mean\big(en[L,N](t)\big) = u$.

In this way the temperature T (and level parameter t) is proportional to the average energy-per-particle.

We turn to convolution as a way of combining systems, with special relevance in this setting. We recall the general construction. Let $M = (M, +, 0)$ be a commutative monoid. The set $\mathcal{D}(M)$ of distributions on M is then also a commutative monoid, see *e.g.* [14]. For $\omega, \rho \in \mathcal{D}(M)$ their convolution sum is defined via tensors and functoriality as:

$$\omega + \rho \;\coloneqq\; \mathcal{D}(+)\big(\omega\otimes\rho\big) \;\in\; \mathcal{D}(M).$$

If $f\colon M \to M'$ is a homomorphism between monoids, then the image map $\mathcal{D}(f)\colon \mathcal{D}(M) \to \mathcal{D}(M')$ is a homomorphism of the convolution monoid on distributions.

The next result uses the (additive) commutative monoid structures on $\mathcal{M}(X)$, see Lemma 5.1 (1), and on $\mathbb{N}$.

Theorem 7.4. *Multinomial and energy distributions are closed under convolutions, as in:*

$$\begin{aligned} mn[N_1](\omega) \;+\; mn[N_2](\omega) &= mn[N_1{+}N_2](\omega) \\ en[L,N_1](t) \;+\; en[L,N_2](t) &= en[L,N_1{+}N_2](t). \end{aligned}$$

The total map $tot\colon \mathcal{M}(\mathbb{N}) \to \mathbb{N}$ is a map of monoids, so that $\mathcal{D}(tot)$ preserves these convolutions, for $\omega \in \mathcal{D}(\mathbb{N})$.

The energy of one particle is given by the Boltzmann distribution itself, as in: $en[L,1]\big(bo[L](t)\big) = bo[L](t)$. This explains the marked equation $\overset{(*)}{=}$ in Definition 7.1.

Proof. The first equation is reasonable standard, so we concentrate on the second one. It follows from the Vandermonde property

of nomial coefficients, see Lemma 6.2 (4).

$$\begin{aligned}
&en[L,N_1](t) \;+\; en[L,N_2](t) \\
&= \sum_{0\leq u_1\leq (L-1)\cdot N_1} \;\sum_{0\leq u_1\leq (L-1)\cdot N_1} \\
&\qquad en[L,N_1](t)(u_1)\cdot en[L,N_2](t)(u_2)\,\big|u_1+u_2\big\rangle \\
&= \sum_{0\leq u_1\leq (L-1)\cdot N_1} \;\sum_{0\leq u_1\leq (L-1)\cdot N_1} \\
&\qquad \frac{C_L(N_1,u_1)\cdot C_L(N_2,u_2)\cdot t^{u_1+u_2}}{Z^{N_1}\cdot Z^{N_2}}\,\big|u_1+u_2\big\rangle \\
&= \sum_{0\leq u\leq (L-1)\cdot (N_1+N_2)} \;\sum_{u_1,u_2,u_1+u_2=u} \\
&\qquad \frac{C_L(N_1,u_1)\cdot C_L(N_2,u_2)\cdot t^{u}}{Z^{N_1+N_2}}\,\big|u\big\rangle \\
&= \sum_{0\leq u\leq (L-1)\cdot (N_1+N_2)} \frac{C_L(N_1{+}N_2,u)\cdot t^{u}}{Z^{N_1+N_2}}\,\big|u\big\rangle \\
&= en[L,N_1{+}N_2](t). \qquad \square
\end{aligned}$$

Example 7.5. Let's write $die = \sum_{1\leq i\leq 6} \frac{1}{6}|\,i\,\rangle$ for the uniform die distribution. When we throw three such dice simultaneously and are interested in the distribution of the sum of the three outcomes, we can describe it equivalently in three different ways — as in Definition 7.1.

(1) As sum of a parallel product: $\mathcal{D}(sum)\big(die\otimes die\otimes die\big)$
(2) As convolution sum: $die + die + die$
(3) As total of a multinomial distribution:

$$\mathcal{D}(tot)\Big(mn[3](die)\Big) \;=\; \mathcal{D}(tot)\left(\sum_{\varphi\in\mathcal{M}[3](\{1,\ldots,6\})} \frac{\binom{\varphi}{}}{3^6}\,|\,\varphi\,\rangle\right).$$

This distribution is discussed in the book [2, §1.2], but without the vocabulary that we use here. There, the multiset coefficients $\binom{\varphi}{}$ are used as 'weights', but they are not explicitly defined. Indistinguishability of microstates is a prominent topic in this book, but the concept of a multiset does not occur.

8 SUFFICIENT STATISTICS VIA ENERGY

In Theorem 5.2 we have seen that accumulation (of sequences to multisets) forms a sufficient statistic. This section will describe two new, but related, examples of sufficient statistics, namely the sum and total maps from (11).

By construction, the nomial from Definition 6.1 is used to count the number of sequences (microstates) with a given energy u. This can be used to define a probabilistic inverse $sum^{\sim 1}\colon \{0,\ldots,(N{-}1)\cdot N\} \to \mathcal{D}\big(\boldsymbol{L}^N\big)$ in the style of (6):

$$sum^{\sim 1}(u) \;\coloneqq\; \sum_{\vec{i}\in sum^{-1}(u)} \frac{1}{C_L(N,u)}\,\big|\vec{i}\,\big\rangle. \tag{17}$$

Via this map we get another instance of Boltzmann's entropy formula, namely $H\big(sum^{\sim 1}(u)\big) = \ln\big(C_L(N,u)\big)$, since the distribution $sum^{\sim 1}(u)$ is uniform; it thus has the highest entropy among all distributions on $\boldsymbol{L}^N$ with energy sum u.

This probabilistic inverse $sum^{\sim 1}$ in (17) makes it possible to undo a sum, in the following sufficient statistic situation.

Theorem 8.1. *The addition of sequences function* $sum\colon \boldsymbol{L}^N \to \{0,\ldots,(L{-}1)\cdot N\}$ *is a sufficient statistic for the parallel Boltzmann distribution* $iid[N]\big(bo[L](t)\big) = bo[L](t)^N$*, as described in the middle of Figure 4.*

Proof. We prove the equation in the middle of Figure 4:

$$\begin{aligned}
&\langle sum, id\rangle_*\Big(bo[L](t)^N\Big) \\
&= \sum_{\vec{i}\in\boldsymbol{L}^N} bo[L](t)^N(\vec{i})\,\big|sum(\vec{i}),\vec{i}\,\big\rangle \\
&= \sum_{\vec{i}\in\boldsymbol{L}^N} \frac{t^{sum(\vec{i})}}{Z^N}\,\big|sum(\vec{i}),\vec{i}\,\big\rangle \qquad \text{by Proposition 4.1 (1)} \\
&= \sum_{0\leq u\leq (L-1)\cdot N} \;\sum_{\vec{i}\in sum^{-1}(u)} \frac{t^u}{Z^N}\,\big|u,\vec{i}\,\big\rangle \\
&= \sum_{0\leq u\leq (L-1)\cdot N} \;\sum_{\vec{i}\in sum^{-1}(u)} \frac{1}{C_L(N,u)}\cdot\frac{C_L(N,u)\cdot t^u}{Z^N}\,\big|u,\vec{i}\,\big\rangle \\
&= \sum_{0\leq u\leq (L-1)\cdot N} \;\sum_{\vec{i}\in\boldsymbol{L}^N} sum^{\sim 1}(u)(\vec{i})\cdot en[L,N](t)(u)\,\big|u,\vec{i}\,\big\rangle \\
&= \langle id, sum^{\sim 1}\rangle_*\Big(en[L,N](t)\Big). \qquad \square
\end{aligned}$$

There is a similar sufficient statistic situation for the total map on multinomials. It does not have a probabilistic inverse, but a suitable 'dagger' channel $tot^\dagger\colon \{0,\ldots,(L{-}1)\cdot N\} \to \mathcal{D}\big(\mathcal{M}[N](\boldsymbol{L})\big)$, of the form:

$$tot^\dagger(u) \;\coloneqq\; \mathcal{D}(acc)\big(sum^{\sim 1}(u)\big) \;=\; \sum_{\varphi\in tot^{-1}(u)} \frac{\binom{\varphi}{}}{C_L(N,u)}\,\big|\varphi\,\big\rangle. \tag{18}$$

There is an 'inverse' of the commuting triangle (11), in terms of composition of channels: $acc^{\sim 1}\mathbin{\circ} tot^\dagger = sum^{\sim 1}$.

We then get a similar sufficient statistics situation, now with multisets instead of sequences (microstates). The proof is like for the previous theorem and is left to the interested reader. In fact, Theorem 8.1 follows from the next result via (11).

Theorem 8.2. *The total of multisets function* $tot\colon \mathcal{M}[N](\boldsymbol{L}) \to \{0,\ldots,(L{-}1)\cdot N\}$ *is a sufficient statistic for the multinomial of the Boltzmann distribution* $mn[N]\big(bo[L](t)\big)$*, as on the right in Figure 4.* $\square$

Remark 8.3. *We have used a modern formulation the different sufficient statistics situation in terms of string diagrams, see Figure 4. There is a more traditional formulation in terms of updating / conditioning that captures more concretely how a parameter disappears in sufficient statistic situation.*

(1) *Theorem 8.1 says that if we condition a product distribution* $bo[L](t)^N$ *with respect to 'microstates with energy* u*', the parameter* t *disappears and the distribution* $sum^{\sim 1}(u)$ *remains.*
(2) *Theorem 8.2 says that conditioning* $mn[N]\big(bo[L](t)\big)$ *on 'multisets with energy* u*' yields the distribution* $tot^\dagger(u)$ *that does not depend on* t*.*

This is in line with the Fisher-Neyman factorisation theorem, see [10, Thm. 14.5] and [4, Prop 4.10] or [24, §3.3].

Example 8.4. In [8, Appendix C] an illustration is given with $L = 5$ energy levels, with $N = 4$ particles, and with total energy $u = 3$. There are three multisets in $\varphi \in \mathcal{M}[4](5)$ with $tot(\varphi) = 3$.

The following table uses the particle configurations from [8, Fig. C-1] in the column on the left. These configurations are interpreted in the current setting with multisets and their coefficients. The checkmarks ✓ indicate how many particles are at which energy level.

configuration on $0,\ldots,4$	**multiset** φ	$\binom{\varphi}{}$
✓✓✓ | | ✓ |	$\varphi_1 = 3\|0\rangle + 1\|3\rangle$	4
✓✓ | ✓ | ✓ | |	$\varphi_2 = 2\|0\rangle + 1\|1\rangle + 1\|2\rangle$	12
✓ | ✓✓✓ | | |	$\varphi_3 = 1\|0\rangle + 3\|1\rangle$	4

The associated distribution builds on the last column using that $C_5(4,3) = \binom{\varphi_1}{} + \binom{\varphi_2}{} + \binom{\varphi_3}{} = 20$. Then:

$$tot^\dagger(3) \;=\; \sum_{\varphi\in\mathcal{M}[4](5)} \frac{\binom{\varphi}{}}{C_5(4,3)}\big|\varphi\big\rangle \;=\; \tfrac{1}{5}\big|\varphi_1\big\rangle + \tfrac{3}{5}\big|\varphi_2\big\rangle + \tfrac{1}{5}\big|\varphi_3\big\rangle.$$

The book [8] describes this situation as an illustration, with the numbers $\binom{\varphi}{}$ and their sum, suggesting the general distribution (18), but without the nomial coefficients (16) needed for normalisation. The book does not mention multisets at all.

Here is another result in which the probabilistic inverse $sum^{\sim 1}$ is useful.

Remark 8.5. *In general, associated with a sufficient statistics situation, there is a (little- / un-known) entropy equation. It uses the conditional entropy notation $H_\omega(c)$ from* (8)*. Consider a function $f\colon X \to Y$ and a distribution $\omega \in \mathcal{D}(X)$ for which there is a channel $f^\dagger\colon Y \to \mathcal{D}(X)$ such that $\langle f, id\rangle_*(\omega) = \langle id, f^\dagger\rangle_*\big(\mathcal{D}(f)(\omega)\big)$. Then:*

$$H\big(\omega\big) \;=\; H\big(\mathcal{D}(f)(\omega)\big) + H_{\mathcal{D}(f)(\omega)}\big(f^\dagger\big).$$

This makes the entropy loss in Lemma 2.1 (4) *precise.*

For the cases occurring in this paper this becomes:

$$\begin{aligned} &H\Big(iid[N]\big(bo[L](t)\big)\Big) \\ &\quad= H\Big(mn[N]\big(bo[L](t)\big)\Big) + H_{mn[N](bo[L](t))}\Big(acc^{\sim 1}\Big). \\ &\quad= H\Big(en[L,N](t)\Big) + H_{en[L,N](t)}\Big(sum^{\sim 1}\Big) \\ &H\Big(mn[N]\big(bo[L](t)\big)\Big) \\ &\quad= H\Big(en[L,N](t)\Big) + H_{en[L,N](t)}\Big(tot^{\sim 1}\Big). \end{aligned}$$

9 MARKOV CHAINS ON SEQUENCES / MICROSTATES

After all the maths in the previous sections it is time for some experiments, not physical but computational. It is a fundamental idea that ensembles of particles in a stable environment undergo random interactions towards an equilibrium. We will describe these ensembles of particles as sequences / microstates, in a set $\boldsymbol{L}^N$ over a fixed set $\boldsymbol{L} = \{0,\ldots,L-1\}$ of energy levels. The transformations of these multisets will be described as a Markov on $\boldsymbol{L}^N$, that is, as a channel $\boldsymbol{L}^N \to \mathcal{D}\big(\boldsymbol{L}^N\big)$. The equilibrium then appears as stationary distribution (on microstates), for this Markov chain, that may be reached after multiple (channel) compositions. The appendix contains essentially the same Markov chain, but then on multisets. We start with the microstate version, because it is a bit easier to see what happens there.

The Markov chain that we define below is a combination of three separate, more elementary channel, called *heat*, *cool* and *agit* (for agitate). The heat channel adds one unit of energy at a random position in the microstate. Similarly, the cool channel randomly removes one unit energy. The agitate channel randomly moves one energy unit to another position. This does not change the energy of the whole microstate.

For a sequence $\vec{i} = (i_0,\ldots,i_{N-1}) \in \boldsymbol{L}^N$ we form the two subsets $\uparrow\!\vec{i}, \downarrow\!\vec{i} \subseteq \boldsymbol{N} = \{0,1,\ldots,N-1\}$ of positions where a unit of energy can be added or removed. Thus:

$$\uparrow\!\vec{i} \coloneqq \{n \in \boldsymbol{N} \mid i_n < L-1\} \qquad \downarrow\!\vec{i} \coloneqq \{n \in \boldsymbol{N} \mid i_n > 0\}.$$

We then define a Markov chain channel $heat\colon \boldsymbol{L}^N \to \mathcal{D}\big(\boldsymbol{L}^N\big)$ that randomly adds a unit of energy, if possible:

$$heat\big(\vec{i}\big) \;\coloneqq\; \begin{cases} 1\Big|\,\vec{i}\,\Big\rangle & \text{if } \uparrow\!\vec{i} = \emptyset \\ \displaystyle\sum_{m\in\uparrow\vec{i}} \frac{1}{M}\Big|\,\vec{i}\,[i_m{+}]\Big\rangle & \text{for } M \coloneqq \big|\uparrow\!\vec{i}\,\big| \end{cases} \tag{19}$$

In the first case occurs all the entries in $\vec{i}$ are at maximum energy $L-1$, so nothing can be added. In the second case, $\vec{i}\,[i_m{+}]$ describes the updated sequence $(i_0,\ldots,i_{m-1},i_m{+}1,i_{m+1},\ldots,i_{N-1})$ with an extra unit of energy at position m. For instance, for $L = 3$ and $N = 5$,

$$\begin{aligned} heat\big(0,2,1,1,0\big) \;=\; &\tfrac{1}{4}\big|1,2,1,1,0\big\rangle + \tfrac{1}{4}\big|0,2,2,1,0\big\rangle \\ &+ \tfrac{1}{4}\big|0,2,1,2,0\big\rangle + \tfrac{1}{4}\big|0,2,1,1,1\big\rangle. \end{aligned}$$

There is also a Markov chain channel $cool\colon \boldsymbol{L}^N \to \mathcal{D}\big(\boldsymbol{L}^N\big)$.

$$cool\big(\vec{i}\big) \;\coloneqq\; \begin{cases} 1\Big|\,\vec{i}\,\Big\rangle & \text{if } \downarrow\!\vec{i} = \emptyset \\ \displaystyle\sum_{k\in\downarrow\vec{i}} \frac{1}{K}\Big|\,\vec{i}\,[i_k{-}]\Big\rangle & \text{for } K \coloneqq \big|\downarrow\!\vec{i}\,\big|. \end{cases} \tag{20}$$

We introduce another such channel $agit\colon \boldsymbol{L}^N \to \mathcal{D}\big(\boldsymbol{L}^N\big)$, for thermal agitation. This channel randomly moves a unit of energy from one position to another.

$$agit\big(\vec{i}\big) \;\coloneqq\; \begin{cases} 1\Big|\,\vec{i}\,\Big\rangle & \text{if } \uparrow\!\vec{i} = \emptyset \text{ or } \downarrow\!\vec{i} = \emptyset \\ heat_*\big(cool(\vec{i})\big) & \text{otherwise.} \end{cases} \tag{21}$$

Here is a simple illustration, for $L = N = 3$.

$$\begin{aligned} agit\big(0,2,1\big) \;=\; &\tfrac{1}{4}\big|0,1,2\big\rangle + \tfrac{1}{4}\big|0,2,1\big\rangle \\ &+ \tfrac{1}{4}\big|1,1,1\big\rangle + \tfrac{1}{4}\big|1,2,0\big\rangle. \end{aligned}$$

We now combine the above three channels into a single 'adjust' Markov chain $adj(u)\colon \boldsymbol{L}^N \to \mathcal{D}\big(\boldsymbol{L}^N\big)$, for an energy level u. It makes single step, so that the energy of a microstate moves towards u. For $\vec{i} \in \boldsymbol{L}^N$ with sum $s \coloneqq sum(\vec{i})$,

$$adj(u)\big(\vec{i}\big) \coloneqq \begin{cases} r\cdot agit(\vec{i}) + (1{-}r)\cdot heat(\vec{i}) & \text{if } s < u,\ r \coloneqq \frac{s}{u} \\ agit(\vec{i}) & \text{if } s = u \\ r\cdot agit(\vec{i}) + (1{-}r)\cdot cool(\vec{i}) & \text{if } s > u,\ r \coloneqq \frac{u}{s} \end{cases}$$

Thus, if the energy s of the sequence differs from the goal u, then the *heat* or *cool* channel is applied, in a convex combination that gives a higher probability to a *heat* or *cool* step if the difference between s and u is higher.

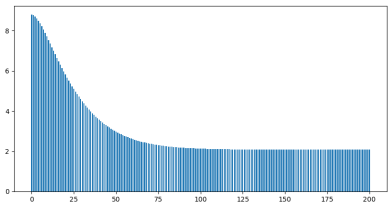
We do a computational experiment with a Python implementation of the adjust Markov chain, for relatively small numbers: we

use $L = 3$ energy levels and $N = 8$ particles. We start with the uniform distribution υ on $\boldsymbol{L}^N$, which involves $3^8 = 6561$ microstates, with entropy $H(\upsilon) = \ln\binom{3^8}{} \approx 8.79$. The average energy of υ is $\frac{(L-1)\cdot N}{2} = 8$, via Lemma 6.2 (3). We take as 15 as target energy. Physically, one can think that we put our system υ in a heat bath with constant energy 15. Thus we compute the channel composite:

$$\Big(adj(15)^n\Big)_*(\upsilon) \;=\; \Big(adj(15) \circ \cdots \circ adj(15)\Big)_*(\upsilon).$$

After $n = 200$ iterations, we end up with a fairly stable distribution on 3^8 in which (essentially) only the microstates with energy 15 remain. We have two observations.

(1) The final entropy is approximately 2.08, which is much lower than the initial entropy of 8.79. We kept track of the (Shannon) entropies during the 200 Markov chain iterations, giving the following plot.



This raises a question: the second law of thermodynamics prescribes that the entropy goes up when a system evolves to an equilibrium, but here the entropy goes down. How is this possible? A physicist will probably say that the second law holds for isolated systems only, and here we are not dealing with an isolated system, since there is a heat bath, with target energy 15. However, we can remove the heat bath by setting the target energy to 8, which is the energy that is in the original (uniform) system. Then, by rerunning the Markov iterations, we still get a descreasing entropy diagram, as above, although with a different slope.

(2) All depends on the fact that we started from a uniform distribution with a high entropy. We could have started from a singleton distribution, containing only one microstate, with entropy zero. The entropy will then go up towards an equilibrium. Such dependence of the second law of thermodynamics on the initial distribution is often not made explicit.

(3) We expected and indeed obtained the uniform distribution as equilibrium, after these 200 iterations, with all the $8 = C_3(8, 15)$ microstates in 3^8 with energy 15. This equilibrium is the distribution $sum^{\sim 1}(15)$ from (17). It has Shannon / Boltzmann entropy $H\big(sum^{\sim 1}(15)\big) = \ln\big(C_3(8, 15)\big) = \ln\binom{8}{} \approx 2.08$.
One might think at this point that the uniform distributions on microstates with the same energy u, of the form $sum^{\sim 1}(u)$, are stationary for the thermal agitation channel (21). This is *not* the case. Here is a very simple example, for $L = 3$, $N = 2$ and $u = 2$. Then $sum^{\sim 1}(2) = \frac{1}{3}\big|0, 2\big\rangle + \frac{1}{3}\big|1, 1\big\rangle + \frac{1}{3}\big|2, 0\big\rangle$ is indeed uniform, but the stationary distribution for the *agit* channel is $\frac{1}{4}\big|0, 2\big\rangle + \frac{1}{2}\big|1, 1\big\rangle + \frac{1}{4}\big|2, 0\big\rangle$, which is not uniform. A general understanding is missing, see the appendix for some more details.

In our set-up we can also form combinations of systems, involving an exchange of heat or of particles, using tensors and convolutions. This will be left to a follow-up paper.

10 CONCLUSIONS

The (discrete) Boltzmann distribution is not part of the standard repertoire in probability theory. It should be. This paper demonstrates that the Boltzmann and resulting energy distributions can be seen as generalisation of the coin / Bernouilli and of die distributions (see Example 7.5). For instance, the biased coin distribution $flip(r) = r|1\rangle + (1-r)|0\rangle$ for $r \in (0, 1)$ is an instance of $bo[2](t)$, for a suitable translation between the parameters r and t. The N-fold convolution sum of flips is the binomial distribution $bn[N](r)$, like in Definition 7.1:

$$\begin{aligned} bn[N](r) \;&=\; flip(r) + \cdots + flip(r) \\ &=\; \mathcal{D}(sum)\Big(flip(r) \otimes \cdots \otimes flip(r)\Big) \\ &=\; \mathcal{D}(tot)\Big(mn[N]\big(flip(r)\big)\Big). \end{aligned}$$

Moreover, the sum and total functions are sufficient statistics for (products / multinomials) of flips. These are the essential new properties that we proved for the energy distribution. Hence we have uncovered familiar properties in a different situation.

The Boltzmann distribution is very much part of the repertoire in statistical physics and thermodynamics. However, in those fields, the concept of a multiset has not (yet) landed and some of the probabilistic properties (like sufficient statistics) have not appeared. We have sketched how Markov chains on microstates (or multisets) can be used to model and study energy dynamics and how they fit well in the probabilistic setting that is developed here.

Statistical physics formed the basis for neural networks and can still be a rich inspiration for computing. Hopefully this article will draw closer connections. The links can become tighter, for instance by including volume, pressure, or chemical potential in probabilistic models, or by getting more clarity about fixed points and maximal entropy. This is left to future work.

REFERENCES

[1] J. Baez. *What is Entropy?* 2024. doi:10.48550/arXiv.2409.09232.
[2] A. Ben-Naim. *A Farewell to Entropy: Statistical Thermodynamics Based on Information.* World Scientific, 2008.
[3] A. Ben-Naim. *Information, entropy, life, and the universe: what we know and what we do not know.* World Scientific, 2015.
[4] J. Bernardo and A. Smith. *Bayesian Theory.* John Wiley & Sons, 2000. Available via https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316870. doi:10.1002/9780470316870.
[5] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley-Interscience, 2nd edition, 2006.
[6] K. Dill and S. Bromberg. *Molecular Driving Forces. Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience.* Garland Science, New York, 2nd rev. edition, 2010.
[7] A. Dragulescu and V. Yakovenko. Statistical mechanics of money. *The European Physical Journal B-Condensed Matter and Complex Systems*, 17:723–729, 2000. doi:10.1007/s100510070114.
[8] R. Eisberg and R. Resnick. *Quantum Physics.* Wiley, New York, 1985.
[9] R. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. Royal Soc.*, 222:594–604, 1922. doi:10.1098/rsta.1922.0009.
[10] T. Fritz. A synthetic approach to Markov kernels, conditional independence, and theorems on sufficient statistics. *Advances in Math.*, 370:107239, 2020. doi:10.1016/J.AIM.2020.107239.
[11] J. Gleick. *The Information. A History, A Theory, A Flood.* Fourth Estate, London, 2011.

[12] B. Jacobs. Partitions and Ewens distributions in element-free probability theory. In *Logic in Computer Science*. Computer Science Press, 2022. doi:10.1145/3531130.3532419.

[13] B. Jacobs. Sufficient statistics and split idempotents in discrete probability theory. In J. Hsu and Ch. Tasson, editors, *Math. Found. of Programming Semantics*, volume 1 of *Elect. Notes in Theor. Inform. & Comp. Sci.*, 2023. doi:10.46298/entics.10520.

[14] B. Jacobs. *Structured Probabilitistic Reasoning*. Cambridge Univ. Press, 2026, to appear. Preliminary version at: http://www.cs.ru.nl/B.Jacobs/PAPERS/ProbabilisticReasoning.pdf.

[15] 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. doi:10.1007/978-3-030-57201-3_6.

[16] E. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620–630, 1957. doi:10.1103/PhysRev.106.620.

[17] E. Jaynes. *Probability Theory: the Logic of Science*. Cambridge Univ. Press, 2003.

[18] T. Leinster. *Entropy and Diversity: The Axiomatic Approach*. Cambridge Univ. Press, 2021. Available online via https://arxiv.org/abs/2012.02113.

[19] D. Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge Univ. Press, 2003.

[20] R. Pathria and P. Beale. *Statistical Mechanics*. Academic Press, 3rd rev. edition, 2011. doi:10.1016/C2009-0-62310-2.

[21] J. Ramshaw. *The Statistical Foundations of Entropy*. World Scientific, Singapore, 2018. doi:10.1142/10823.

[22] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 (Part I) and 623–656 (Part II), 1948.

[23] K. Stowe. *An Introduction to Thermodynamics and Statistical Mechanics*. Cambridge Univ. Press, 2nd rev. edition, 2007. doi:10.1017/CBO9780511801570.

[24] Y. Suhov and M. Kelbert. *Probability and Statistics by Example: Volume 1, Basic Probability and Statistics*. Cambridge Univ. Press, 2005.

[25] Online Encyclopedia of Integer Sequences. Multinomial coefficients. oeis.org/wiki/Multinomial_coefficients, accessed: 2025-06-25.

APPENDIX

Additional background information will be provided about the Markov chain computations in Section 9. First, a more detailed description will be given for the agitation channel on microstates. Next, multiset analogues of the Markov chains on microstates will be defined.

An elaborated agitation example

We consider another application of the agitate channel 21 in detail and show that it can have multiple stationary distributions as fixed points, some of them uniform, and some not. We keep things simple, with level and particle parameters $L = 4$ and $N = 2$.

For energy $u = 1$ there are two sequences $(0, 1)$ and $(1, 0)$ in 4^2 with energy one. They are combined in the uniform distribution $sum^{\sim 1}(1) = \frac{1}{2}|0, 1\rangle + \frac{1}{2}|1, 0\rangle$. Applying a the *cool* channel (20) removes one unit of energy from a random position. In these to cases there is no choice:

$$cool(0, 1) = 1\big|0, 0\big\rangle = cool(1, 0).$$

When we subsequently add one unit via the *heat* channel (19) there is a choice:

$$heat(0, 0) = \tfrac{1}{2}|0, 1\rangle + \tfrac{1}{2}|1, 0\rangle\,.$$

Thermal agitation involves putting these together:

$$\begin{aligned} agit_*\Big(sum^{\sim 1}(1)\Big) &= heat_*\Big(cool_*\big(\tfrac{1}{2}|0, 1\rangle + \tfrac{1}{2}|1, 0\rangle\big)\Big) \\ &= heat_*\Big(1\big|0, 0\big\rangle\Big) \\ &= \tfrac{1}{2}|0, 1\rangle + \tfrac{1}{2}|1, 0\rangle = sum^{\sim 1}(1). \end{aligned}$$

We now do the same for $u = 3$. The uniform distribution is:

$$sum^{\sim 1}(3) = \tfrac{1}{4}|0, 3\rangle + \tfrac{1}{4}|1, 2\rangle + \tfrac{1}{4}|2, 1\rangle + \tfrac{1}{4}|3, 0\rangle.$$

Cooling the sequences involves yields:

$$\begin{aligned} cool(0, 3) &= 1|0, 2\rangle \\ cool(1, 2) &= \tfrac{1}{2}|0, 2\rangle + \tfrac{1}{2}|1, 1\rangle \\ cool(2, 1) &= \tfrac{1}{2}|1, 1\rangle + \tfrac{1}{2}|2, 0\rangle \\ cool(3, 0) &= 1|2, 0\rangle. \end{aligned}$$

As a result,

$$cool_*\Big(sum^{\sim 1}(3)\Big) = \tfrac{3}{8}|0, 2\rangle + \tfrac{1}{4}|1, 1\rangle + \tfrac{3}{8}|2, 0\rangle.$$

Heating the sequences in this distributions gives:

$$\begin{aligned} heat(0, 2) &= \tfrac{1}{2}|1, 2\rangle + \tfrac{1}{2}|0, 3\rangle \\ heat(1, 1) &= \tfrac{1}{2}|2, 1\rangle + \tfrac{1}{2}|1, 2\rangle \\ heat(2, 1) &= \tfrac{1}{2}|3, 1\rangle + \tfrac{1}{2}|2, 2\rangle. \end{aligned}$$

The resulting agitations are:

$$\begin{aligned} agit(0, 3) &= \tfrac{1}{2}|0, 3\rangle + \tfrac{1}{2}|1, 2\rangle \\ agit(1, 2) &= \tfrac{1}{4}|0, 3\rangle + \tfrac{1}{2}|1, 2\rangle + \tfrac{1}{4}|2, 1\rangle \\ agit(2, 1) &= \tfrac{1}{4}|1, 2\rangle + \tfrac{1}{2}|2, 1\rangle + \tfrac{1}{4}|3, 0\rangle \\ agit(3, 0) &= \tfrac{1}{2}|2, 1\rangle + \tfrac{1}{2}|3, 0\rangle. \end{aligned}$$

Notice that agitation includes 'identity hops' of a unit of energy from one position to the same position, leaving the sequence / microstate unchanged. The question remains if this really captures thermal agitation for microstates.

In this case thermal agitation does *not* preserve the uniform distribution:

$$\begin{aligned} agit_*\Big(sum^{\sim 1}(3)\Big) &= agit_*\Big(\tfrac{1}{4}|0, 3\rangle + \tfrac{1}{4}|1, 2\rangle + \tfrac{1}{4}|2, 1\rangle + \tfrac{1}{4}|3, 0\rangle\Big) \\ &= \tfrac{3}{16}|0, 3\rangle + \tfrac{5}{16}|1, 2\rangle + \tfrac{5}{16}|2, 1\rangle + \tfrac{3}{16}|3, 0\rangle. \end{aligned}$$

The distribution of microstates with energy 3 that does form a fixed point of *agit* is $\frac{1}{6}|0, 3\rangle + \frac{1}{3}|1, 2\rangle + \frac{1}{3}|2, 1\rangle + \frac{1}{6}|3, 0\rangle$. This is then an illustration where the equilibrium does *not* have maximal entropy. These seems at odds with the second law of thermodynamics, more specifically, with the Principle of Equal a Priori Probabilities. Before drawing any drastic conclusions, it may be good to first reach agreement on how to capture thermal agitation via a Markov chain. We have described it as random hops of units of energy between particles. It seems that the occurrence of zero energies in microstates leads to non-uniform fixed points, since at those positions with zero energy no identity hops can happen. One could redefine agitation so that it does not involve identity hops. This seems *ad hoc* and does not extend to multisets (see below).

One suggestion is that agitation can possibly be described systematically in a setting with equalisers and coequalisers, in analogy with accumulation and arrangement. Indeed, accumulation arises as the coequaliser $acc\colon X^K \to \mathcal{M}[K](X)$ of all permutation maps $X^K \to X^K$. Arrangement $arr\colon \mathcal{M}[K](X) \to \mathcal{D}(X)$ is the equaliser of these permutation maps, in the Kleisli category $\mathcal{K}\ell(\mathcal{D})$. One can see in a similar way the addition map $sum\colon \boldsymbol{L}^N \to \{0, \ldots, (L–1)\cdot N\}$ as equaliser of all energy hops.

Markov chains on multisets

In Section 9 we have elaborated an example involving distribution on microstates (sequences / lists). Specifically we used $L = 3$ levels and $N = 8$ particles, yielding $3^8 = 6561$ microstates. When we

switch from microstates to multisets, the numbers go down dramatically, since there are only $45 = \binom{\binom{3}{8}}{}$ multisets of size $N = 8$ over $L = 3$ energy levels, see Lemma 5.1 (2). We decided to introduce Markov chains on microstates first, since the transitions involved (like *heat*, *cool*, *agit*) are a bit easier to understand in terms of sequences. Here, we briefly describe the corresponding multiset versions. This means one no longer has to think concretely in terms of positions with energies, but more abstractly in terms of numbers of occurrences of energies.

Thus, for general energy and particle levels L, N we wish to describe Markov chains $Heat, Cool, Agit \colon \mathcal{M}[N](\boldsymbol{L}) \to \mathcal{D}\big(\mathcal{M}[N](\boldsymbol{L})\big)$. We use a capital for the multiset versions. We can define them via the accumulate and arrange maps, for instance as channel composition:

$$Heat \coloneqq \Big(\mathcal{M}[N](\boldsymbol{L}) \xrightarrow{\;arr\;}\!\!\circ\; \boldsymbol{L}^N \xrightarrow{\;heat\;}\!\!\circ\; \boldsymbol{L}^N \xrightarrow{\;acc\;}\!\!\circ\; \mathcal{M}[N](\boldsymbol{L})\Big).$$

Here we write $X \mathrel{-\!\!\circ\!\!\to} Y$ for a channel $X \to \mathcal{D}(Y)$.

This description is mathematically nice, but not computationally, since the large powers $\boldsymbol{L}^N$ still occur. Here is a more direct description, on a multiset $\varphi \in \mathcal{M}[N](\boldsymbol{L})$.

$$Heat(\varphi) \coloneqq \begin{cases} 1\big|\,\varphi\,\big\rangle & \text{if } \varphi(L{-}1) = N, \text{ else:} \\ \displaystyle\sum_{0 \le i < L-1} \frac{\varphi(i)}{N - \varphi(L{-}1)} \,\Big|\,\varphi - 1|\,i\,\rangle + 1|\,i{+}1\,\rangle\Big\rangle & \end{cases}$$

$$Cool(\varphi) \coloneqq \begin{cases} 1\big|\,\varphi\,\big\rangle & \text{if } \varphi(0) = N, \text{ else:} \\ \displaystyle\sum_{0 < i < L} \frac{\varphi(i)}{N - \varphi(0)} \,\Big|\,\varphi - 1|\,i\,\rangle + 1|\,i{-}1\,\rangle\Big\rangle\,. & \end{cases}$$

Both for *Heat* and *Cool* the first cases deal with the situation with maximal and minimal energy, where no single unit of energy can be added or removed. The second cases randomly add one unit of energy at those levels which are not yet at maximum $L-1$ or at minimum 0.

For instance, for $L = 4$ and $N = 10$, the above definition gives:

$$\begin{aligned} &Cool\Big(3|\,0\,\rangle + 5|\,1\,\rangle + 2|\,3\,\rangle\Big) \\ &= \tfrac{5}{7}\Big|\,4|\,0\,\rangle + 4|\,1\,\rangle + 2|\,3\,\rangle\Big\rangle + \tfrac{2}{7}\Big|\,3|\,0\,\rangle + 5|\,1\,\rangle + 1|\,2\,\rangle + 1|\,3\,\rangle\Big\rangle\,. \end{aligned}$$

These *Cool* and *Heat* maps resemble the draw-delete and draw-add maps that play a fundamental role elsewhere, for instance in population genetics [12] and in De Finetti limit results [15].

We can now define the thermal agitation channel on multisets, essentially in the same way as on microstates in (21):

$$Agit(\varphi) \coloneqq \begin{cases} 1\,\Big|\,\varphi\,\Big\rangle & \text{if } \varphi(L{-}1) = N \text{ or } \varphi(0) = N \\ Heat_*(Cool(\varphi)) & \text{otherwise.} \end{cases}$$

One can again ask what the stationary distributions are for these *Agit* channels on multisets. One might think that these stationaries are of the form $tot^\dagger(u)$, as accumulations of the uniform distributions $sum^{\sim 1}(u)$, see (18). This is not the case, as the following example shows, for $L = 3$, $N = 6$ and $u = 4$. First,

$$\begin{aligned} tot^\dagger(4) \;=\; &\tfrac{1}{6}\Big|\,2|\,0\,\rangle + 4|\,1\,\rangle\Big\rangle \\ &+ \tfrac{2}{3}\Big|\,3|\,0\,\rangle + 2|\,1\,\rangle + 1|\,2\,\rangle\Big\rangle + \tfrac{1}{6}\Big|\,4|\,0\,\rangle + 2|\,2\,\rangle\Big\rangle\,. \end{aligned}$$

Applying agitation yields a different outcome:

$$\begin{aligned} Agit_*\Big(tot^\dagger(3)\Big) \;=\; &\tfrac{7}{36}\Big|\,2|\,0\,\rangle + 4|\,1\,\rangle\Big\rangle \\ &+ \tfrac{41}{60}\Big|\,3|\,0\,\rangle + 2|\,1\,\rangle + 1|\,2\,\rangle\Big\rangle + \tfrac{11}{90}\Big|\,4|\,0\,\rangle + 2|\,2\,\rangle\Big\rangle\,. \end{aligned}$$

The actual stationary distribution for thermal agitation on multisets is in this case:

$$\tfrac{2}{9}\big|\,2|\,0\,\rangle + 4|\,1\,\rangle\,\big\rangle + \tfrac{2}{3}\big|\,3|\,0\,\rangle + 2|\,1\,\rangle + 1|\,2\,\rangle\,\big\rangle + \tfrac{1}{9}\big|\,4|\,0\,\rangle + 2|\,2\,\rangle\,\big\rangle.$$

The illustration in Section 9 involved $L = 3$, $N = 8$ with energy $u = 15$. In that case, the total-dagger does give a fixed point:

$$tot^\dagger(15) \;=\; 1\Big|\,1|\,1\,\rangle + 7|\,2\,\rangle\Big\rangle \;=\; Agit_*\Big(tot^\dagger(15)\Big).$$

Notice that this fixed point does not involve zero energies.

It remains an open question to characterise such stationary distributions for agitate, with arbitrary L, N, u.