Entropy and Large Deviations

Lecture Notes, Master's course, Fall 2024

Klaas Landsman

Institute for Mathematics, Astrophysics, and Particle Physics
Radboud Center for Natural Philosophy
Radboud University, Nijmegen, The Netherlands
landsman@math.ru.nl

Contents

1 Introduction

> You should call it "entropy" for two reasons: First, the function is already used in thermodynamics under that name; second, and more importantly, most people don't know what entropy really is, and if you use the word "entropy" in an argument you will win every time.[1]
>
> (John von Neumann to Claude Shannon, 1940)

The concept op *entropy* was introduced by Clausius (1865), which is one of the founding papers of thermodynamics.[2] This is done in his eq. (59), which reads $dS = dQ/T$ (where Q is heat and T absolute temperature).[3] This is preceded by the comment that dQ/T is a 'complete differential of a quantity that only depends on the state of a body at a particular instant', as opposed to a dependence on some path towards that state (like heat and work, whose sum, energy, is also path-independent and hence a 'function of state'). A few pages later, Clausius writes the following:

> Sucht man für S einen bezeichnenden Namen, so könnte man, ähnlich wie von der Grösse U gesagt ist, sie sei der Wärme - und Werkinhalt des Körpers. Da ich es aber für besser halte, die Namen derartiger für die Wissenschaft wichtiger Grössen aus den alten Sprachen zu entnehmen, damit sie unverädert in allen neuen Sprachen angewandt werden können, so schlage ich vor, die Grösse S nach dem Griechischen Worte ἡ τροπὴ, die Verwandlung, die E n t r o p i e des Körpers zu nennen. Das Wort E n t r o p i e habe ich absichtlich dem Worte E n e r g i e möglichst ähnlich gebildet, denn die beiden Grössen, welche durch diese Wörter benannt werden sollen, sind ihren physikalischen Bedeutungen nach einander so nahe verwandt, dass eine gewisse Gleichartigkeit in der Benennung mit zweckmässig zu seinn scheint. (Clausius, 1867, p. 34).[4]

The end of this paper is very famous, expressing the (main) two laws of thermodynamics:[5]

> **1)** D i e E n e r g i e d e r W e l t i s t c o n s t a n t.
>
> **2)** D i e E n t r o p i e d e r W e l t s t r e b t e i n e m M a x i m u m zu.[6]

[1] Myron Tribus got this anecdote first-hand from Shannon in 1961 (Levine & Tribus, 1978, pp. 2–3).

[2] See Brush (1974, 1976, 2003), Truesdell (1980), von Plato (1994), Müller (2007), Uffink (2007), Emch & Liu (2013), Weinberger (2013), Gaudenzi (2019), Saslow (2020), and Norton (2022) for history and analysis of thermodynamics and 19th century statistical physics. Apart from entropy, the creation of thermodynamics also included the introduction of the concept of *energy* in physics; see e.g. Smith (1998) for (especially) the British side, Wegener (2009) for (especially) the German side, and Elkana (1974), Harman (1982), Coppersmith (2010), and Pitts (2021) in general.

[3] This idea originated in the work of Carnot (1824), which (with hindsight) is often seen as the beginning of thermodynamics. Carnot described "heat engines" like a steam engine abstractly as devices in which incoming heat Q_1 is converted to work W and outgoing heat $Q_2 < Q_1$, so that $W = Q_1 - Q_2$ (this was one of the roots of the later idea of energy conservation, although Carnot himself mistakenly believed in the conservation of heat!). If Q_1 comes in from a reservoir at temperature T_1 and Q_2 leaves to a reservoir at some lower temperature T_1, then according to Carnot a maximally efficient (and hence reversible) machine has $\frac{Q_1}{T_1} = \frac{Q_2}{T_2}$ (the irreversible case is $\frac{Q_1}{T_1} \leq \frac{Q_2}{T_2}$). Defining the efficiency of the machine as W/Q_1, this is therefore at best equal to $1 - (T_2/T_1)$, so that T_1 mus be high and T_2 low.

[4] 'If one is looking for a descriptive name for S, one could, in a similar way to what has been said about the quantity U, say that it is the heat and work content of the body. However, since I think it is better to take the names of such important scientific terms from the old languages, so that they can be used without modification in all new languages, I suggest that the term S be used after the Greek words ἡ τροπὴ, the transformation, the e n t r o p y of the body. I have deliberately made the word e n t r o p y as similar as possible to the word e n e r g y , because the two quantities which are to be named by these words are so closely related in their physical meanings that a certain similarity in the naming seems appropriate.' Edited translation by DeepL.com (free version). The page number is from the 1867 reprint.

[5] See Uffink (2002) and Roberts (2022) for a penetrating analysis of the second law.

[6] 1) 'T h e e n e r g y o f t h e w o r l d i s c o n s t a n t.
2) T h e e n t r o p y o f t h e w o r l d t e n d s t o w a r d s a m a x i m u m.'

The next important contribution to the concept of entropy was made by Boltzmann (1872), which is one of the founding papers of kinetic gas theory; in particular, it contains the *Boltzmann equation*.[7] For us, the point is that already the Summary at the beginning contains the formula:

$$E = \int_0^\infty f(x,t) \left[\log \left(\frac{f(x,t)}{x} \right) - 1 \right] dx, \tag{1.1}$$

where $f(x,t)$ is the 'number of molecules having energy x at time t'. This is the first occurrence of the combination "$p_i \log p_i$" (usually with a minus sign omitted by Boltzmann) that characterizes practically all versions of entropy that deserve the name. Boltzmann introduced it as a tool for proving irreversibility: what is now know as is H-theorem (after a later renaming of a simplified version E to H) states that if f solves the Boltzmann equation, then $E(f)$ 'can never increase but must always decrease or remain constant.' Of equal (or even greater) importance to the (hi)story of entropy is Boltzmann (1877), in which he initiated the combinatorial and probabilistic approach to entropy that also pervades the large deviations approach advocated in these notes.[8] In particular, it contains a "maximum entropy principle", according to which 'the most likely state', identified with an equilibrium state, is found by minimizing a (negative) entropy functional

$$M' = \int_0^\infty f(x) \log f(x) dx, \tag{1.2}$$

subjects to constraints $n = \int_0^\infty f(x)dx$ and $L = \int_0^\infty xf(x)dx$ that determine the total particle number and energy, respectively; here, as in Boltzmann (1872), f is still regarded as a probability distribution function of singe-particle energy, but the expression has clearly been simplified compared to (1.1) and later in the paper Boltzmann defines the entropy with the "usual" sign, this time regarding $f(x,y,z,u,v,w)$ as a function of position $\vec{x} = (x,y,z)$ and velocity $\vec{v} = (u,v,w)$.

Boltzmann left a huge heritage. We will not discuss his ideas on irreversibility here; in our view these are predicated on the precise connection between the 'Stosszahlansatz' (molecular chaos hypothesis) in the derivation of the Boltzmann equation, high probability (later called 'typicality') as relied on in Boltzmann (1877), high entropy, chaos (as used both in the 19th century in gas theory and in the 20th century in dynamical systems with sensitive dependence on initial conditions), and randomness. We will take this up in a sequel to these notes in connection with Kolmogorov's ideas on both dynamical systems and algorithmic randomness.[9]

Gibbs (1902) is quite rightly seen as the beginning of modern statistical mechanics, or at least as a *new* beginning after Boltzmann; their differences still remain to be fully sorted out and understood.[10] Gibbs introduces the microcanonical, canonical, and grand canonical ensembles and, without ever writing down an expression like $\int \rho \log \rho$, defines the entropy (in a certain ensemble defined by a probability measure ρ on an N-particle phase space) as minus the average value (under ρ) of what he calls the 'index of probability' $\log \rho$ of ρ (also for other probabilities, for example of thermodynamic phases).[11] In its microscopic nature (being based on the probabilities of microscopic states, i.e., complete specifications of N-particle positions and velocities) this differs from Boltzmann's approach in that Gibbs introduced what we now call *fine-grained* entropies.

[7] Further to footnote 2, Carcignani (1998), Darrigol (2018) and Uffink (2022) specifically concern Boltzmann.

[8] The formula "$S = k \log W$" on Boltzmann's grave in Vienna is not to be found in Boltzmann (1877) or indeed in any of his writings; it was given by Planck (1906), §134, who also introduced "Boltzmann's constant" k. However, Boltzmann (1898), p. 172, gives a special case. See Hoyer (1980).

[9] For a first impression see Landsman (2023).

[10] Sklar (1993), Frigg (2008), and Myrvold (2021) are good starting points; see also Werndl & Frigg (2017), Frigg & Werndl (2018), Wallace (2018), etc.

[11] 'the average index of probability *with its sign reversed* corresponds to entropy.' (Gibbs, 1902, p. 50). The notation ρ seems of later use; Gibbs (1902) always writes down explicit expressions for his probability measures.

Boltzmann's earlier entropies, on the other hand, are *coarse-grained* entropies, defined by the probabilities of specific macroscopic states (like his distribution function f, which compresses data about an astronomical number of particles into a single function). Finally, we just mention that Gibbs (1902) proved the variational principles (1.16) and (1.18) below for the free energy and the entropy, which corresponds to Theorems III and II in his Chapter XI, respectively.

Our next hero of entropy is Einstein (1910),[12] who turned the Boltzmann–Planck formula

$$S = k \log W \tag{1.3}$$

on its head:

> Das Boltzmannsche Prinzip kann durch die Gleichung
>
> $$S = \lg W + \text{konst.} \tag{1}$$
>
> formuliert werden. Hierbei bedeutet R die Gaskonstante, N die Zahl der Moleküle in einem Grammolekül, S die Entropie, W ist die Größe, welche als die "Wahrscheinlichkeit" desjenigen Zustandes bezeichnet zu werden pflegt, welchem der Entropiewert S zukommt.
>
> Gewöhnlich wird W gleichgesetzt der Anzahl der möglichen verschiedenen Arten (Komplexionen), in welchen der ins Auge gefaßte, durch die beobachtbaren Parameter eines Systems im Sinne einer Molekulartheorie unvollständig definierte Zustand realisiert gedacht werden kann. Um W berechnen zu können, braucht man eine *vollständige* Theorie (etwa eine vollständige molekular-mechanische Theorie) des ins Auge gefaßten Systems. Deshalb erscheint es fraglich, ob bei dieser Art der Auffassung dem Boltzmannschen Prinzip *allein*, d.h. ohne eine *vollständige* molekular-mechanische oder sonstige die Elementarvorgänge vollständig darstellende Theorie (Elementartheorie) irgend ein Sinn zukommt. Gleichung (1) erscheint ohne Beigabe einer Elementartheorie oder - wie man es auch wohl ausdrücken kann - vom phänomenologischen Standpunkt aus betrachtet inhaltlos. (...)
>
> Trotzdem aber können aus (1) genaue Beziehungen über das statistische Verhalten eines Systems abgeleitet werden, und zwar in dem Falle, daß der Bereich der Zustandsvariabeln, für welchen W in Betracht kommende Werte hat, als unendlich klein angesehen werden kann. Aus Gleichung (1) folgt
>
> $$W = \text{konst.}\, e^{\frac{N}{R}S}$$
>
> Diese Gleichung gilt der Größenordnung nach, wenn man jedem Zustand Z ein kleines Gebiet, von der Größenordnung wahrnehmbarer Gebiete, zuordnet. Die Konstante bestimmt sich der Größenordnung nach durch die Erwägung, daß W für den Zustand des Entropiemaximums (Entropie S_0) von der Größenordnung Eins ist, so daß man der Größenordnung nach hat
>
> $$W = e^{\frac{N}{R}(S-S_0)}$$
>
> Daraus ist zu folgern, daß die Wahrscheinlichkeit dW dafür, daß die Größen $\lambda_1, \ldots \lambda_n$ zwischen λ_1 und $\lambda_1 + d\lambda_1 \ldots \lambda_n$ und $\lambda_n + d\lambda_n$ liegen, der Größenordnung nach gegeben ist durch die Gleichung
>
> $$dW = e^{\frac{N}{R}(S-S_0)} d\lambda_1 \ldots d\lambda_n$$
>
> und zwar in dem Falle, daß das System durch die $\lambda_1, \ldots \lambda_n$ (in phänomenologischem Sinne) nur unvollständig bestimmt ist.[13]

[12] Einstein's early work (i.e. before 1905) mainly concerned thermodynamics and statistical mechanics, and he continued to work on these topics throughout his career. The canonical reference is *The Collected Papers of Albert Einstein*, especially Volumes 1, 2, and 3, available online at `https://einsteinpapers.press.princeton.edu`. See also Stachel et al., (1990), Klein et al. (1994), and Uffink (2006).

[13] Boltzmann's principle can be expressed by the equation $(1) : S = \lg W + \text{konst.}$ where R is the gas constant, N

Since he was the first to relate entropy to fluctuations, Einstein (1910) may be seen as the beginning of large deviation (= large fluctuation) theory (Touchette, 2009; Jona-Lasinio, 2015).[14]

Continuing our entropic Hall of Fame: in the founding paper of information theory Shannon (1948) asked how much "'choice" is involved in the selection of an event' that is drawn from a finite space of events with probabilities $p_1, \ldots, p_n$, or 'how uncertain we are of the outcome'. His answer is what is now called the *Shannon entropy* (which he attributes to Boltzmann):[15]

$$S_2(p) := -\sum_{a \in A} p(a) \log_2 p(a), \tag{1.4}$$

where for the moment we follow Shannon in using the base-2 logarithm; we will later switch to base-e. Apart from some heuristic motivation for the use of the logarithm (see below), Shannon also gave an axiomatic characterization of (1.4), of which we give the following slightly different and streamlined version, where $\mathrm{Prob}(A)$ denotes the set of all probability distributions on A:[16]

1. $S_2(p) \equiv S_2(p_1, \ldots, p_n)$, where $n = |A|$, $p_i = p(a_i)$, is continuous in the probabilities $p(a)$.
2. If $p(a) = 1/n$ for all $a \in A$, then $g(n) := S_2(1/n, \ldots, 1/n)$ satisfies $g(n+1) \geq g(n)$ for all n.
3. If p' is obtained from p by omitting all events with zero probability, then $S_2(p') = S_2(p)$.
4. If $r \in \mathrm{Prob}(C)$, where C is finite like A, and we coarse-grain $C = \bigsqcup_{a \in A} U_a$ into disjoint events $U_a \subset C$, with associated distribution $p \in \mathrm{Prob}(A)$ given by $p(a) = \sum_{c \in U_a} p(c)$, then

$$S_2(r) = S_2(p) + \sum_{a \in A} p(a) S_2(r|a), \tag{1.5}$$

   where the conditional probability $r|a \in \mathrm{Prob}(C)$ is given by $(r|a)(c) = r(c|U_a)$, which equals zero of $c \notin U_a$ and equals $r(c)/r(U_a)$ if $c \in U_a$.
5. $S_2(\frac{1}{2}, \frac{1}{2}) = 1$.

is the number of molecules in one gram-molecule, S is the entropy, W is the quantity customarily designated as the "probability" of the state with which the entropy value S is associated. W is commonly equated with the number of different possible ways (complexions) in which the state considered-which is incompletely defined in the sense of a molecular theory by observable parameters of a system-can conceivably be realized. In order to be able to calculate W, one needs a *complete* theory (perhaps a complete molecular-mechanical theory) of the system under consideration. Given this kind of approach, it therefore seems questionable whether Boltzmann's principle *by itself* has any meaning whatsoever, i.e., without a *complete* molecular-mechanical or other theory that completely represents the elementary processes (elementary theory). If not supplemented by an elementary theory or–to put it differently–considered from a phenomenological point of view, equation (1) appears devoid of content. (...) Nevertheless, it is possible to derive exact relationships concerning the statistical behavior of a system from equation (1) in cases where the range of the state variables for which W has values for the kind under consideration can be regarded as infinitely small. It follows from equation (1) that $W = \mathrm{konst.} \exp((N/R) \cdot S)$. This equation is valid to an order of magnitude if each state Z is assigned a small region of the order of magnitude of perceptible regions. The order of magnitude of the constant is determined by taking into account that for the state of maximum entropy (entropy S_0) W is of the order of magnitude one, so that we then have, with order-of-magnitude accuracy, $W = \exp((N/R) \cdot (S - S_0))$. From this we can conclude that the probability dW that the quantities $\lambda_1, \ldots \lambda_n$ lie between λ_1 and $\lambda_1 + d\lambda_1 \ldots \lambda_n$ and $\lambda_n + d\lambda_n$ is given, in order of magnitude, by the equation $dW = \exp((N/R) \cdot (S - S_0)) d\lambda_1 \ldots d\lambda_n$ in the case when the system is determined only incompletely (in the phenomenological sense) by $\lambda_1, \ldots \lambda_n$.

[14] Einstein's argument is not probabilistic but empirical in nature: as the above quotation makes clear, he is dissatisfied with the Botlzmann–Planck formula (1.3) for the entropy because in practice one may not know the microscopic theory from which to compute W in sufficient detail, whereas if one knows the entropy phenomenologically, it gives information about the microscopic theory.

[15] Shannon (1948) gives no reference but mentions 'Boltzmann's famous H theorem.' Guizzo (2003) and Gleick (2011) are histories of information theory.

[16] Shannon tacitly assumed that $S_2(p)$ is defined by the same formula for any A. Khinchin (1957) made this explicit.

Exercise 1 *Prove* (1.4) *satisfies conditions 1–4, and that conversely 1–4 imply* (1.4)*.*

Taking $C = A \times B$ and $U_a = \{a\} \times B$, the probability distribution $p \in \mathrm{Prob}(A)$ in (1.5) is now the marginal of $r \in \mathrm{Prob}(A \times B)$, that is, $p(a) = \sum_{b \in B} r(a,b)$, whereas $q|a \in \mathrm{Prob}(B)$ is given by

$$(q|a)(b) = r(a,b)/p(a), \tag{1.6}$$

provided that $p(a) > 0$; if $p(a) = 0$ then the term $p(a)S_2(q|a)$ in the sum over a is omitted. If A and B are probabilistically independent, i.e., $r = p \times q$ in the sense that $r(a,b) = p(a)q(b)$, then $q|a = q$, and (1.5) states that S_2 is additive on independent probabilities,[17] that is,

$$S_2(r) = S_2(p) + S_2(q). \tag{1.7}$$

For general $r \in \mathrm{Prob}(A \times B)$ with marginals $p \in \mathrm{Prob}(A)$ (given above) and $q \in \mathrm{Prob}(B)$, i.e., $q(b) = \sum_{a \in A} r(a,b)$, the Shannon entropy (1.4) is merely *subadditive*, in that

$$S_2(r) \leq S_2(p) + S_2(q). \tag{1.8}$$

Exercise 2 *Prove that Shannon's entropy* (1.4) *satisfies* (1.8)*.*

What else does (1.4) mean? Taking a flat distribution $p = f$, defined by

$$f(a) := 1/|A|, \tag{1.9}$$

one obtains

$$S_2(f) = \log_2 |A|, \tag{1.10}$$

which is a poor man's version of Planck's formula (1.3); for the rich man's version see §5 from (5.19) onwards. In the further special case where $|A| = 2^k$, so that $S_2(f) = k$, the Shannon entropy (1.4), degenerating into (1.10), states the minimal number of bits necessary to encode the elements $a \in A$ (or, equivalently, the minimal number of yes-no questions one needs to ask to figure out which element the opponent has in mind in a systematic procedure not relying on guesswork and luck). The general case is covered by Shannon's *noiseless coding theorem*, which, we will discuss as Theorem 6.3 in these notes. For now, the theorem roughly speaking states that

$$I_2(a) = \log_2(1/p(a)) \tag{1.11}$$

is close to the length $\ell(C(a))$ of the code-word $C(a)$ in some optimal binary coding $C : A \to 2^*$ of A, and hence $S_2(p)$ is the average length of such code-words.[18] Here $2^* = \bigcup_{N \in \mathbb{N}} 2^N$ is the set of all finite binary *strings*; 2^N is the set of binary strings of length N, and $\mathbb{N} = \{0,1,2,\ldots\}$. Thus

$$S_2(p) = \langle I_2 \rangle_p \tag{1.12}$$

[17] But this requirement could not replace the stronger axiom (1.5), since mere additivity on independent probabilities would allow the (base-2) *Rényi entropy* $S^{(\alpha)}(p) = (1-\alpha)^{-1} \log_2(\sum_{a \in A} p(a)^\alpha)$, where $0 < \alpha < \infty$ and $\alpha \neq 1$; in the limit $\alpha \to 1$ one recovers S. Khinchin (1957) showed that (up to a multiplicative constant) $p \mapsto S_2(p)$ is the unique function of $p \in \mathrm{Prob}(A)$, where A is a finite set, that: (i) is defined by the same formula for any A; (ii) is *symmetric* (i.e. permutation-invariant in its arguments $p(a_1), \ldots, p(a_{|A|})$; (iii) is maximal on flat distributions (1.9); (iv) does not change if probability zero events are added; and (v) satisfies (1.5) for $C = A \times B$ and $U_a = \{a\} \times B$. A third axiomatization, due to Faddeev (1956), is: (i) S_2 is symmetric; (ii) if $A = 2 = \{0,1\}$ and $p = p(0)$, then $p \mapsto S_2(p, 1-p)$ is continuous and positive at least at one point; (iii) $h^{(n+1)}(p_1, \ldots, p_{n-1}, p_n, p_{n+1}) = h^{(n)}(p_1, \ldots, p_{n-1}, p'_n) + p'_n h^{(2)}(p_n/p'_n, p_{n+1}/p'_n)$, where $p'_n := p_n + p_{n+1}$ and for clarity we have added the number of arguments of h. This is also a special case (and hence a weakening) of Shannon's coarse-graining axiom, where, if $C = \{c_1, \ldots, c_n, c_{n+1}\}$, we take $A = \{1, \ldots, n\}$ with $U_1 = \{c_1\}, \ldots, U_{n-1} = \{c_{n-1}\}, U_n = \{c_n, c_{n+1}\}$, so that in the above notation we have $p'_n = p(c_n) + p(c_{n+1})$, etc. See Aczél & Daróczy (1975) for a detailed discussion, and more recently also Klir (2006), §3.2.2.

[18] See e.g. Cover & Thomas (2006), §5.4. Of course, the equality $\ell(C(a)) = I_2(a)$ can only be satisfied if $p(a) = 2^{-m}$ for some integer $m \in \mathbb{N}$. Otherwise, one can find a code for which $\ell(C(a))$ equals the smallest integer $\geq I_2(a)$.

is the average length of code-words under an optimal code, as determined by (A,p), $p \in \mathrm{Prob}(A)$. Here and in what follows we denote averages or expectation values of functions $f : A \to \mathbb{R}$ by

$$\langle f \rangle_p := \sum_{a \in A} p(a) f(a), \tag{1.13}$$

and likewise for general probability spaces. The quantity (1.11), which as we have seen was already used by Gibbs (1902), may also be interpreted as a quantification of the *surprise* brought by an outcome $a \in A$ upon sampling a probability space (A,p). The idea is that the surprise should be great if the outcome is unlikely, suggesting a "surprise function" $s(a) \sim 1/p(a)$. But one likes this function to be additive on pairs of independent events, which rather suggests the *information function* (1.11). Thus $S_2(p)$ is also the "average surprise" inherent in p. In particular, if $p(a) = 1$ for some $a \in A$ then $S_2(p) = 0$, and indeed there is no surprise at all since the outcome is certain to be a. On the other hand, eq. (1.9) turns out to maximize the average surprise.

Before turning to our last champion of entropy (i.e. Kolmogorov), we briefly explain the connection between Shannon's entropy (1.4) and Gibbsian thermodynamics and statistical mechanics,[19] still in the finite case (we switch to the base-e logarithm, as usual in physics). For good reasons we write Ω for the finite set previously called A, change to the usual base e logarithm and hence define the entropy of a probability distribution $p \in \mathrm{Prob}(\Omega)$ as as

$$S(p) := -\sum_{\omega \in \Omega} p(\omega) \log p(\omega). \tag{1.14}$$

For any function $U : \Omega \to \mathbb{R}$ we define the *partition function* $Z(U)$ and *free energy* $F(U)$ by

$$Z(U) := \sum_{\omega \in \Omega} e^{-U(\omega)}; \qquad\qquad F(U) := -\log Z(U), \tag{1.15}$$

respectively. In its simplest form, the *variational principle* of thermodynamics then states that

$$F(U) = \inf_{p \in \mathrm{Prob}(\Omega)} \{ \langle U \rangle_p - S(p) \}, \tag{1.16}$$

where the infimum is actually achieved uniquely (so it is a minimum) by the *Gibbs distribution*

$$p_U(\omega) = \frac{1}{Z(U)} e^{-U(\omega)}. \tag{1.17}$$

Conversely, the entropy may be recovered from the free energy by

$$S(p) = \inf_U \{ \langle U \rangle_p - F(U) \}, \tag{1.18}$$

where the infimum (actually a minimum) is taken over all functions $U : \Omega \to \mathbb{R}$. Thus entropy (as a function of probability distributions p) and free energy (as a function of "energy" U) are dual to each other with respect to a ("Fenchel") transformation we will later describe in some more detail.

In physics one writes $U = \beta H$, where $\beta > 0$ is a constant, related to the temperature by $\beta = 1/T$, and *for a given Hamiltonian* H writes $F(\beta) = F(\beta H)/\beta$, and similarly $p_\beta := p_{\beta H}$, i.e.,

$$p_\beta(\omega) := \frac{1}{Z(\beta)} e^{-\beta H(\omega)}; \qquad\qquad Z(\beta) := \sum_{\omega \in \Omega} e^{-\beta H(\omega)} = e^{-\beta F(\beta)}. \tag{1.19}$$

[19] This is a special case of what is called *thermodynamic formalism* (Ruelle, 2004; Beck & Schlögl, 1993; Viana, 1997; Viana & Oliveira, 2016, Ch. 12). See §11 of these notes for a brief summary.

Computing the minimum in (1.16) using (1.13) and (1.4) by straightforward calculus gives

$$F(\beta) = \langle H \rangle_{p_\beta} - TS(p_\beta), \tag{1.20}$$

which is a version of the equation "$F = E - TS$" familiar from equilibrium thermodynamics, except for the fact that In thermodynamics entropy is a function of energy $u \in \mathbb{R}$ rather than probability $p \in \mathrm{Prob}(\Omega)$. To distinguish thermodynamics entropy from Shannon's entropy we denote the former by S_C (in honour of Clausius). From a modern perspective we may *define* S_C by

$$S_C(u) = \sup_{p \in \mathrm{Prob}(\Omega)} \{S(p) \mid \langle H \rangle_p = u\}; \tag{1.21}$$

see Theorem 8.1. The constraint $\langle H \rangle_p = u$ can be satisfied for some $p \in \mathrm{Prob}(\Omega)$ if and only if

$$E_{\min} \leq u \leq E_{\max}, \tag{1.22}$$

where $E_{\min}$ and $E_{\max}$ are the minimum and maximum values of H, respectively; if not, we put $S(u) = -\infty$ (which is justified by the property $\sup \emptyset = -\infty$). If $u = E_{\min}$ or $u = E_{\max}$, then $S_C(u) = 0$. If $E_{\min} < u < E_{\max}$, then there exists a unique $\beta \in \mathbb{R}$ for which $\langle H \rangle_{p_\beta} = u$, cf. (1.19), and hence the supremum in (1.21) is in fact a maximum. For this value of β, seen as $\beta(u)$, we then obtain

$$S_C(u) = \beta(u - F(\beta)). \tag{1.23}$$

This matches (1.20) in the sense that $S_C(u) = S(p_\beta)$ at the value of β just defined, and hence *a posteriori* justifies the definition (1.21). Eq. (1.23) and hence eq. (1.21) may then be rewritten as

$$S_C(u) = \inf_{\beta \in \mathbb{R}} \{\beta(u - F(\beta))\}. \tag{1.24}$$

It follows that the counterpart of the (Fenchel) dual expression (1.16) then reads

$$\beta F(\beta) = \inf_{u \in \mathbb{R}} \{\beta u - S_c(u)\}; \tag{1.25}$$

provided that $\beta > 0$, one could simply write this as $F(\beta) = \inf_{u \in \mathbb{R}} \{u - TS_c(u)\}$, where $T = 1/\beta$.

This tacitly supposed a flat prior (1.9). For an arbitrary prior $q \in \mathrm{Prob}(\Omega)$, replace (1.14) by

$$S_q(p) := S(p) + \sum_{\omega \in \Omega} p(\omega) \log q(\omega) = -D(p\|q), \tag{1.26}$$

where the right-hand side is the *relative entropy* or *Kullback–Leibler distance* (or *divergence*):[20]

$$D(p\|q) := \sum_{\omega \in \Omega} p(\omega) \log \left(\frac{p(\omega)}{q(\omega)} \right). \tag{1.27}$$

This is defined as stated provided $q(\omega) = 0$ implies $p(\omega) = 0$ (i.e., $p \ll q$, or: p is absolutely continuous w.r.t. q); if not, we put $D(p\|q) = \infty$. Note that

$$S_f(p) = -D(p\|f) = S(p) - \log|\Omega|. \tag{1.28}$$

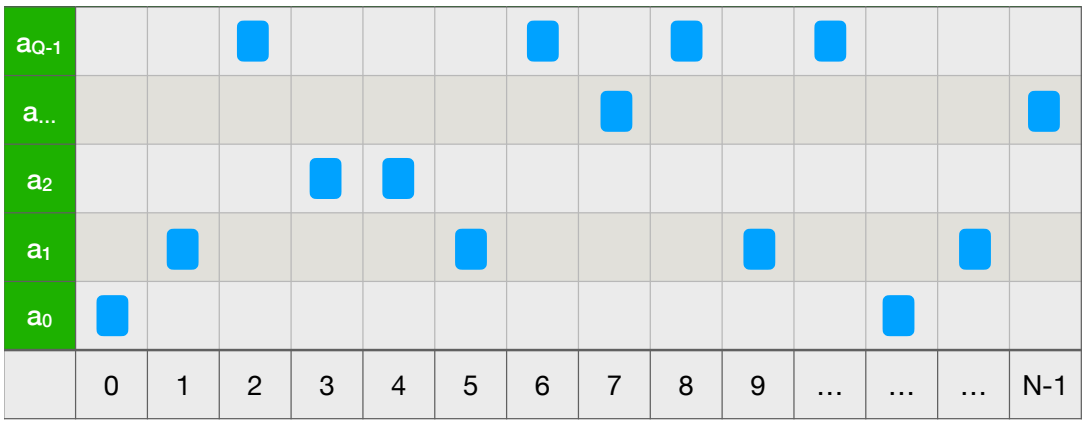
Furthermore, (1.15) and (1.19) are replaced by

$$Z_q(U) := \sum_{\omega \in \Omega} q(\omega) e^{-U(\omega)} = \langle e^{-U} \rangle_q; \qquad Z_q(U) := -\log Z_q(U); \tag{1.29}$$

$$p_\beta(\omega) := \frac{1}{Z_q(\beta)} q(\omega) e^{-\beta H(\omega)}; \qquad Z_q(\beta) := \langle e^{-\beta H} \rangle_q = e^{-\beta Z_q(\beta)}. \tag{1.30}$$

As a bridge to Kolmogorov and to this course in general, consider the following diagram:

[20] The original reference is Kullback & Leibler (1951). This entropy was rediscovered by Jauch & Baron (1972).



Here $N \in \mathbb{N}_* = \{1,2,3,\ldots\}$ is some *large* natural number (we will often consider the limit $N \to \infty$), whereas $Q = |A| \in \{2,3,\ldots\}$ is the cardinality of a given finite set $A = \{a_0,\ldots,a_{|A|-1}\}$. Unlike, N, the case of small $|A|$ is already interesting (even $|A| = 2$). As already mentioned for $A = 2 = \{0,1\}$, A^N is the set of all functions $\sigma : N \to A$, where $N = \{0,1,\ldots,N-1\}$ as usual in set theory. Such a function is also called a *string* over A, having length $\ell(\sigma) \equiv |\sigma| = N$. We write either $\sigma(n)$ or σ_n for its value at $n \in N$, and may write σ as $\sigma_0\sigma_1\cdots\sigma_{N-1}$. In particular, if $A = 2 = \{0,1\}$, then σ is a binary string. Thus a (binary) *string* σ is *finite*, whereas a (binary) *sequence* s is *infinite*. The set of all binary sequences is denoted by $2^{\mathbb{N}}$, and likewise $A^{\mathbb{N}}$ consists of all functions $s : \mathbb{N} \to A$. For $s \in A^{\mathbb{N}}$, we write $s_{|N}$ for $s_0 s_1 \cdots s_{N-1} \in A^N$, to be sharply distinguished from $s_n \equiv s(n) \in A$.

This diagram has at least four interpretations, of which we have already seen the first two:

- In *statistical mechanics* à la Boltzmann (1877), N is the number of (distinguishable) particles in a gas, and $a \in A$ labels some property a single particle may have. For example, for non-interaction particles one may think of $a \in A$ as a label of their energy $E_a \in \mathbb{R}$. Spin chains also fall under this formalism, where $a \in A$ labels some internal degree of freedom at site n. In statistical mechanics $\sigma \in A^N$ is called a *microstate* of the gas (or spin chain, etc.).

- In *information theory* as created by Shannon (1948), N is the number of letters drawn from an alphabet A by sampling a given probability distribution $p \in \mathrm{Prob}(A)$, the space of all probability distributions on A. So each "microstate" $\sigma \in A^N$ is a word with N letters.

- In *dynamical systems* à la Kolmogorov (1958), see §2 for details, consider a partition

$$X = \bigsqcup_{a \in A} U_a \tag{1.31}$$

of the phase space or configuration space X of each particle, so that $U_a \subset X$ and different subsets U_a (called *cells* of the partition) are disjoint (this is expressed by the symbol $\bigsqcup$). If (X,Σ,P) is a probability space (where $\Sigma \subset \mathscr{P}(X)$ is the σ-algebra on which the probability measure P is defined),[21] each U_a is measurable (i.e. $U_a \in \Sigma$) and typically the cells U_a need neither be mutually disjoint nor exhaust X, as long as $P(U_a \cap U_b) = 0$ whenever $a \neq b$ and $P(\bigcup_a U_a) = 1$. If we also have a measurable map $T : X \to X$, a microstate $\sigma \in A^N$ describes the coarse-grained trajectory *of a single particle* that started at time $t = n = 0$ at some point $x \in U_{\sigma(0)}$, then hopped to $Tx \in U_{\sigma(1)}$, …, on to $T^n x \in U_{\sigma(n)}$ at $t = n$, etc., till time $t = N-1$.

- In *probability theory*, our diagram displays a (truncated) sample path of a discrete-time stochastic process. This will be explained in §3. It incorporates the previous interpretations.

[21] We use the inclusion symbol $\subset$ as what is often called $\subseteq$, so that $\Sigma \subset \mathscr{P}(X)$ allows $\Sigma = \mathscr{P}(X)$.

The aim of this course is to introduce relevant concepts to entropy in the relatively simple setting of the above diagram, with special attention to the above interpretations.[22] As recognized by Boltzmann, Einstein, Shannon, and Kolmogorov, the origin of entropy lies in probability theory. From the point of view of our diagram, a natural point of entry lies in the following objects:

$$L_N : A^N \to \mathrm{Prob}(A); \qquad L_N(\sigma) = \frac{1}{N}\sum_{n=0}^{N-1} \delta_{\sigma_n}; \tag{1.32}$$

$$T_N : \mathrm{Prob}(A) \to \mathscr{P}(A^N); \qquad T_N(p) := \{\sigma \in A^N \mid L_N(\sigma) = p\}, \tag{1.33}$$

where, for any $a \in A$, the point measure $\delta_a \in \mathrm{Prob}(A)$ gives $\delta_a(B) = 1$ if $a \in B$ and $\delta_a(B) = 0$ if $a \notin B$, where $B \subset A$. Here L_N is called the *empirical measure* and T_N is called the *type class*. The former reads the configuration of squares in our diagram and computes the corresponding relative frequency (seen as a probability) that some square enters $a \in A$. The latter *starts* from a probability distribution on A and assembles all microstates whose empirical probability equals the given one. A key idea of this course lies in the specific way the Shannon entropy (1.4) appears as a limit:

$$S(p) = \lim_{N\to\infty} \frac{1}{N} \log_2 |T_N(p)|. \tag{1.34}$$

Similarly, adding a prior $q \in \mathrm{Prob}(A)$, the relative entropy (1.27) appears in a certain limit:

$$\inf_{p\in U} D(p\|q) = -\lim_{N\to\infty} \frac{1}{N} \log q^N(L_N \in U), \tag{1.35}$$

where $U \subset \mathrm{Prob}(A)$ is open, and q^N is the (Bernoulli) probability measure on A^N induced by q. If $q \in U$, then the left-hand side vanishes and this forces $q^N(L_N \in U) \to 1$, consistent with the (weak) law of large numbers $L_N \to q$. But if $q \notin U$, then $\inf_{p\in U} D(p\|q) > 0$ and (1.35) states that $q^N(L_N \in U)$ exponentially decays in N. This is a basic result in the theory of *large deviations*.

Historically, the oldest example of a large deviation result was *Cramér's theorem* (1938).[23] Let (X_n) be $\mathbb{R}$-valued i.i.d. random variables, which for simplicity we here take to be given by some function $E : A \to \mathbb{R}$ so that $X_n : A^N \to \mathbb{R}$ with $X_n(\sigma) = E(\sigma_n)$, distributed via $q^N \in \mathrm{Prob}(A^N)$ for some prior $q \in \mathrm{Prob}(A)$, that is, $P(X_n \in B) = q^N(\{\sigma \in A^N \mid X_N(\sigma) \in B)\})$, for $B \subset \mathbb{R}$. Then

$$S_N := \frac{1}{N}\sum_{n=1}^{N-1} X_n, \tag{1.36}$$

is a real random variable, defined on the same probability space A^N (we still take A finite). Then:

1. The average $S_N \to \mu \equiv \langle E\rangle_q$ is described by the weak law of large numbers:

$$\forall_{\varepsilon>0} \lim_{N\to\infty} q^N(-\varepsilon \le S_N - \mu \le \varepsilon) = 1. \tag{1.37}$$

2. The $O(N^{-1/2})$ fluctuations of $S_N - \mu$ follow from the central limit theorem:

$$\lim_{N\to\infty} q^N\left(\frac{a}{\sqrt{N}} \le S_N - \mu \le \frac{b}{\sqrt{N}}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b dx\, e^{-x^2/2\sigma^2}, \tag{1.38}$$

where $\sigma^2 = \langle E^2\rangle_q - \langle E\rangle_q^2$ is the variance of X_1 (i.e. of E) and hence of any X_n.

[22] For most results we also state more general versions for the case where A is no longer finite but compact and metrizable, or even *Polish*, that is, metrizable such that the resulting space is complete and separable (for compact spaces metrizability and separability are equivalent). Our proofs are mainly restricted to the finite case, though.

[23] The original source is Cramér (1938), but the version of Theorem 8.5 goes back to Chernoff (1952).

3. The $O(1)$ or "large" fluctuations, then, are described by Cramér's theorem:

$$\lim_{N\to\infty}\frac{1}{N}\log q^N(a\le S_N-\mu\le b)=-\inf_{x\in[\mu+a,\mu+b]}I_q(x), \tag{1.39}$$

where the rate of this exponential decay I_q is defined by a "maximum entropy principle"

$$I_q(x):=\inf\{D(p\|q)\mid p\in\mathrm{Prob}(A),\langle E\rangle_p=x\}. \tag{1.40}$$

It is interesting to quote one of the masters of modern probability theory here:

Important Note

> I want you to know that *beyond* the law of large numbers and the central limit theorem, there is *no*, so to speak *philosophically correct refinement*. (McKean, 2014, p. 51)

But Cramér's theorem (which McKean covers!), and more generally large deviation theory, seems to undermine this view. Even short of this theorem, we can prove a special case of (1.39), namely in the case where the X_n are not only i.i.d. but also Gaussian = normal:[24] In the standard case for simplicity,[25] S_N is also normal, still with mean $\mu=0$ but with variance $\sigma=1/N$.

Exercise 3 *Prove that for any $\delta>0$ we have*

$$\lim_{N\to\infty}\frac{1}{N}\log(q^N(S_N\ge\delta))=-\delta^2/2, \tag{1.41}$$

so that the 'large fluctuations' $q^N(S_N\ge\delta)$ exponentially decay like $\exp(-N\delta^2/2)$.

This course explains the concept of entropy in the context of such asymptotic results, which culminate in two key results, called *Sanov's theorem* and (as we have just seen) *Cramér's theorem*. On the basis of these results we also discuss the abstract and general theory of large deviations, covering for example key theorems by Varadhan and by Gärtner and Ellis, which we then apply to statistical mechanics (a different application, appropriate for a course in computer science, would be coding theory à la Shannon, of which we only review the most basic results). This closes the historical circle, since statistical mechanics historically speaking emerged from thermodynamics, which gave rise to the birth of entropy in the first place).[26] This mainly involves a combination of combinatorics, probability theory, and analysis. The latter in turn often relies on *convexity*, whose basic theory we review in the appendix. Working on these notes the author increasingly got the impression that a serious understanding of this field also requires introductions to Markov chains as well as to ergodic theory; indeed, ergodic theory is the real source of various laws of large numbers in probability theory, and also lies behind one of the key theorems on entropy, namely the *Shannon–McMillan–Breiman theorem*, see (5.41). This, in turn, led to the inclusion of a chapter on Kolmogorov–Sinai (also called metric) entropy of dynamical systems, which not only beautifully fits into our list of possible interpretations of our diagram above, but also is inseparable from ergodic theory. The limiting procedure used to define this particular entropy also gives a nice relationship with the so-called "thermodynamic limit" in statistical mechanics.

Finally, we relied on numerous sources, listed in the bibliography, but especially recommend: Austin (2017); Cover & Thomas (2006); Dembo & Zeitouni (1998); Dorlas (2021); Ellis (1995); Georgii (2011); McKean (2014); Rassoul-Agha & Seppäläinen (2015); Viana & Oliveira (2016).

[24] This is taken from Dembo & Zeitouni, §1.1.

[25] The general case is $P(X_n\in[a,b])=(2\pi\sigma^2)^{-1/2}\int_a^b dx\,\exp(-(x-\mu)^2/2\sigma^2)$; the standard case is $\mu=0$, $\sigma=1$.

[26] It is actually thermodynamics that "emerges" from statistical mechanics!

2 Entropy in dynamical systems

Kolmogorov (1958) introduced the third way of looking at our diagram in the Introduction:[27]

- A *dynamical system* (in the probabilistic sense we use) is a triple (X,P,T), where $(X,P) \equiv (X,P,\Sigma)$ is a probability space (we suppress the σ-algebra Σ whenever possible), and

$$T : X \to X \tag{2.1}$$

is a measurable (but not necessarily invertible) map, required to preserve P in the sense that

$$P(T^{-1}B) = P(B), \tag{2.2}$$

for any measurable $B \subset X$ ($B \in \Sigma$). Here $T^{-1}(B) := \{x \in X \mid T(x) \in B\}$; if T is invertible, this set coincides with the image $\{T^{-1}(x) \mid x \in B\}$ of B under the inverse map $T^{-1} : X \to X$.

Now suppose we have a partition (1.31) of X. Then $\sigma \in A^N$ is a *coarse-grained path of a single particle* (rather than a configuration of N particles, as in the previous point), such that

$$T^n x \in U_{\sigma(n)} \qquad\qquad (n = 0,1,\ldots,N-1, \sigma_n \in A). \tag{2.3}$$

Hence our particle starts at $x \in U_{\sigma(0)}$ at $t = 0$, moves to $Tx \in U_{\sigma(1)} \subset X$ at $t = 1$, etc., and at time $t = N-1$ finds itself at $T^{N-1}x \in U_{\sigma(N-1)}$. Thus a partition (1.31) defines a map

$$\xi : X \to A^{\mathbb{N}}; \qquad\qquad \xi(x)_n = a \in A \;\text{ iff }\; T^n x \in U_a \;\; (n \in \mathbb{N}). \tag{2.4}$$

Hence a fine-grained path $(x, Tx, T^2x, \ldots) \in X^{\mathbb{N}}$ is coarse-grained to $\xi(x) \in A^{\mathbb{N}}$, which in turn may be truncated to give a map $\xi_N : X \to A^N$ defined by $\xi_N(x) = \xi(x)_{|N} \in A^N$, where $\xi(x)_{|N} \in A^N$ is the restriction of $\xi \in A^{\mathbb{N}}$ to $N \subset \mathbb{N}$. A key role will be played by the (unilateral) *shift*

$$S : A^{\mathbb{N}} \to A^{\mathbb{N}}; \qquad\qquad (Ss)_n := s_{n+1} \;\; (n = 0,1,\ldots), \tag{2.5}$$

because this map satisfies the intertwiner relation

$$S \circ \xi = \xi \circ T. \tag{2.6}$$

To see this, we note that by definition of S we have $S \circ \xi(x)_n = a$ iff $\xi(x)_{n+1} = a$, which by definition of ξ is the case iff $T^{n+1}x \in U_a$. On the other hand, $\xi \circ T(x)_n = a$ iff $T^n(Tx) \in U_a$. But $T^n(Tx) = T^{n+1}x$, which proves (2.6). Here is an almost trivial but instructive example.

Exercise 4 *Take $X = A^{\mathbb{N}}$ with time-evolution $T = S$ and partition $(U_a)_{a \in A}$ defined by*

$$U_a = \{s \in A^{\mathbb{N}} \mid U_0 = a\}. \tag{2.7}$$

Show that $\xi = \mathrm{id}$, that is, $\xi(s) = s$ for all $s \in A^{\mathbb{N}}$.

Thus the given triple (X,P,T) is coarse-grained by a new triple $(A^{\mathbb{N}},P',S)$, where

$$P'(B) = P(\xi^{-1}B). \tag{2.8}$$

[27] See Sinai (1989) and more generally Charpentier, Lesne, & Nikolski (2007) for Kolmogorov's closely related contributions to dynamical systems, entropy, and ergodic theory. Relevant textbooks include for example Collet & Eckmann (2006), Castiglione *et al.* (2008), and Viana & Oliveira (2016).

If ξ were injective, then, spectacularly, nothing would be lost in coarse-graining. This property is quite rare, but it is (almost by construction) the case for the *doubling map* $T_D : [0,1) \to [0,1)$,

$$T_D(x) = 2x \ \ (0 \leq x < 1/2); \qquad\qquad T_D(x) = 2x - 1 \ \ (1/2 \leq x < 1), \tag{2.9}$$

so that $T_D(x)$ takes the fractional part of $2x$. Perhaps surprisingly (because of the factor 2), this map preserves Lebesgue measure μ_L, acting as the probability measure on $X = [0,1)$ in charge. The key is that T_D is not invertible. To show this, one only needs to check (2.2) for intervals.[28]

Exercise 5 *For $0 \leq a < b < 1$, show that $\mu_L(T_D^{-1}[a,b]) = \mu_L([a,b]) = a - b$.*

In terms of the binary expansion

$$x = \sum_{n=0}^{\infty} s_n 2^{-(n+1)}, \tag{2.10}$$

the doubling map T_D corresponds to the unilateral shift (2.5) on $2^{\mathbb{N}}$. There are two cases:[29]

1. if $0 \leq x < 1/2$, which corresponds to $s_0 = 0$, then

$$T_D(x) = 2x = 2 \cdot \sum_{n=0}^{\infty} s_n 2^{-(n+1)} = \sum_{n=0}^{\infty} s_n 2^{-n} = \sum_{n=1}^{\infty} s_n 2^{-n} = \sum_{n=0}^{\infty} s_{n+1} 2^{-(n+1)} = \sum_{n=0}^{\infty} (Ss)_n 2^{-(n+1)}.$$

2. If $1/2 \leq x < 1$, which corresponds to $s_0 = 1$, then s_0 also disappears.

Exercise 6 *Show this.*

In other words, if we write $T'_D : 2^{\mathbb{N}} \to 2^{\mathbb{N}}$ for the transfer of the doubling map $T_D : [0,1) \to [0,1)$ to $2^{\mathbb{N}}$ via (2.10), then $T'_D = S$. We use this fact to show that if we use the partition $\{U_0, U_1\}$ of $[0,1)$ given by $U_0 = [0,1/2)$ and $U_1 = [1/2,1)$, then the infinite coarse-grained path $\xi(x) \in 2^{\mathbb{N}}$ equals

$$\xi : [0,1) \to 2^{\mathbb{N}}; \qquad\qquad \xi(x) = s, \tag{2.11}$$

where s is defined by (2.10). To see this, first look at the partition $\{U'_0, U'_1\}$ of $2^{\mathbb{N}}$ where, cf. (2.7)

$$U'_i = \{s \in \ 2^{\mathbb{N}} \mid s_0 = i\}; \ \ (i = 0,1). \tag{2.12}$$

Clearly, $x \in U_i$ iff $s \in U'_i$ for $i = 1,2$. Then $\xi' : 2^{\mathbb{N}} \to 2^{\mathbb{N}}$, defined as in (2.4) by the property

$$\xi'(s)_n = i \in \{0,1\} \ \text{ iff } \ S^n s \in U_i, \tag{2.13}$$

is the identity, since for $n = 1$ the condition $Ss \in U_i$ means that $(Ss)_0 = s_1 = i$, and likewise $S^n s \in U_i$ means $s_n = i$. Transporting this back to $[0,1)$ gives (2.11).[30]

A slightly more subtle example where ξ is a bijection is given by the *tent map*

$$T_t : [0,1) \to [0,1); \qquad\qquad T_t(x) = 1 - |2x - 1|, \tag{2.14}$$

that is, $T_t(x) = 2x$ for $0 \leq x < 1/2$ and $T_t(x) = 2 - 2x$ if $\frac{1}{2} \leq x < 1$. This map also preserves μ_L.

[28] Turning these special cases into a proof requires, for example, the following theorem: *Let $R \subset \Sigma$ be an* algebra*, that is, a subset $\Sigma \subset \mathscr{P}(X)$ that contains X and is closed under complementation $A \backslash B$ and unions $A \cup B$ (and hence under intersections $A \cap B$). Let R also contains an increasing sequence (B_n) such that $X = \bigcup_n B_n$. If $P(T^{-1}B) = P(B)$ for all $B \in R$, then $P(T^{-1}B) = P(B)$ for all $B \in \Sigma$.* See for example Dajani & Kalle, (2021), Theorem 1.2.1.

[29] More precisely, this is true provided dyadic rationals in $[0,1)$ are mapped to finite binary sequences (i.e., those ending with infinitely many zeros). Binary expansions are not unique; dyadic rationals (i.e. numbers in $[0,1)$ with a *finite* binary expansion, = fractions of the kind $m \cdot 2^{-n}$ for $m,n \in \mathbb{N}_*$ subject to $0 < m < 2^n$), have two binary expansions, e.g. $1/2 = 2^{-1} = \sum_{n=2}^{\infty} 2^{-n}$. If x is a dyadic rational we should take the $s \in 2^{\mathbb{N}}$ with infinitely many zeros at the end. This is clear from (2.9), since every dyadic rational eventually ends up at $x = 0$ and stays there.

[30] Continuing the previous footnote: this map ξ is not quite a bijection, but it is a bijection up to the set of dyadic rationals in $[0,1)$, which has Lebesque measure zero. Alternatively, one could omit elements of $2^{\mathbb{N}}$ that end with an infinite string of 1's (these are precisely the binary sequences that superfluously represent dyadic rationals).

Exercise 7 *Show this.*

The situation is similar to the doubling map, but with a twist: seen as a map

$$T_t' : 2^{\mathbb{N}} \to 2^{\mathbb{N}} \tag{2.15}$$

via the binary expansion (2.10) of x, for $s_0 = 0$, corresponding to $x \in [0, 1/2)$, we have

$$T_t' = T_D' = S. \tag{2.16}$$

But if $s_0 = 1$ and hence $x \in [1/2, 1)$, we have

$$T_t'(s)_n = s'_{n+1} := 1 - s_{n+1}. \tag{2.17}$$

Exercise 8 *Show this.*

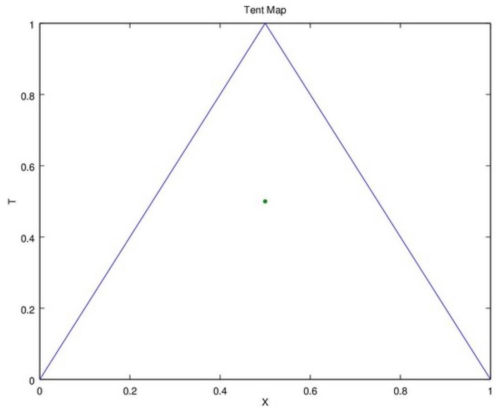
This gives a formula for the map

$$\xi_t : [0,1) \to 2^{\mathbb{N}}, \tag{2.18}$$

relative to the same partition $\{U_0, U_1\}$ as for the doubling map, defined recursively:

$$\xi_t(x)_0 = s_0; \qquad \xi_t(x)_n = s_n \ \ (\xi_t(x)_{n-1} = 0); \qquad \xi_t(x)_n = s'_n \ \ (\xi_t(x)_{n-1} = 1). \tag{2.19}$$

This is once again a bijection (up to the dyadic rationals, which may be taken out without loss).



There are *invertible* $2d$ versions of both maps, both called the *Baker's map*. The $2d$ counterpart of the doubling map is the so-called *unfolded Baker's map*.[31] Here (2.9) is extended by

$$\begin{aligned} X &= [0,1) \times [0,1); \\ T_B(x,y) = (2x, y/2) \ \ (0 \le x < 1/2); &\qquad T_B(x,y) = (2x-1, y/2+1/2) \ \ (1/2 \le x < 1). \end{aligned} \tag{2.20}$$

This map consists of three stages, visible if we divide the square into two equal columns:[32]

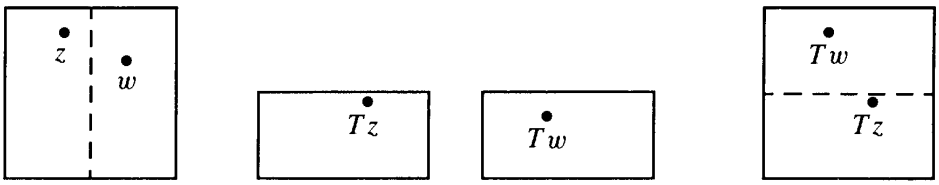
1. Each column is compressed vertically by a factor 1/2.

[31] The $2d$ extension of the tent map is given by $T(x,y) = (2x, y/2)$ for $0 \le x < 1/2$ and $T(x,y) = (2-2x, 1-y/2)$ for $1/2 \le x < 1$. Here also $\xi : X \to 2^{\mathbb{Z}}$ is an isomorphism, differing from the one in the main text for the unfolded version by a twist, as for the tent map. This is called the *folded Baker's map*.

[32] We follow Shields (1996), §I..2.b, almost *verbatim*, and also steal two of his pictures, viz. Figure I.2.4 and I.2.5.

2. Each compressed column is subsequently stretched horizontally by a factor 2;

3. The right rectangle is finally placed on top of the left one (is this really what bakers do?).



By either calculation or visualization, it follows that T_B preserves $2d$ Lebesgue measure. Short of defining chaos, this is arguably the most chaotic map known: if all points in the columns

$$U_0 = \{(0 \leq x < 1/2, 0 \leq y < 1)\}; \qquad U_1 = \{(1/2 \leq x < 1, 0 \leq y < 1)\} \tag{2.21}$$

are marked red and blue, respectively, then the two colours look completely mixed after a dozen or so steps.[33] The simplest coarse-graining already gives an isomorphism with the shift map, albeit it the two-sided or bilateral version: instead of $2^{\mathbb{N}}$ we now use $2^{\mathbb{Z}}$, carrying a shift map

$$S : 2^{\mathbb{Z}} \to 2^{\mathbb{Z}}; \qquad (Ss)_n := s_{n+1} \ \ (n \in \mathbb{Z}). \tag{2.22}$$

We now use the binary expansion (2.10) of x, and a relabeled binary expansion for y:

$$y = \sum_{n=0}^{\infty} s_{-(n+1)} 2^{-(n+1)}. \tag{2.23}$$

Thus (x, y) corresponds to a single sequence $s \in 2^{\mathbb{Z}}$, and it is easy to show that under this correspondence the map (2.20) is exactly the bilateral shift (2.22).

Exercise 9 *Show this.*

Consequently, the (unfolded) Baker's map is essentially the same as the bilateral Bernoulli shift (2.22). Using $\mathbb{Z}$ instead of $\mathbb{N}$, we replace the subset $N \equiv \{0, 1, \ldots, N-1\} \subset \mathbb{N}$ by

$$\Lambda_N := \{-N, \ldots, 0, \ldots, N\}, \tag{2.24}$$

and in its wake replace A^N by $A^{\Lambda_N} = \{\sigma : \Lambda_N \to A\}$, here with $A = \{0, 1\}$, and replace (2.4) by

$$\xi : X \to A^{\mathbb{Z}}; \qquad \xi(x)_n = a \in A \ \text{ iff } \ T^n x \in U_a \ \ (n \in \mathbb{Z}), \tag{2.25}$$

which may be truncated to Λ_N so as to obtain the finite-time coarse-grained path

$$\xi_{\Lambda_N}(x) = \xi(x)_{|\Lambda_N} \in A^{\Lambda_N}. \tag{2.26}$$

Applying this to the Baker's map (2.20) via the partition (2.21) leads to the conclusion that the map ξ in (2.25), with $A = \{0, 1\}$ and $X = [0, 1) \times [0, 1)$, is a bijection, so that coarse-graining is lossless. To see this, one again starts with the bilateral shift (2.22), with corresponding partition

$$U_0' = \{s \in 2^{\mathbb{Z}} \mid s_0 = 0\}; \qquad U_1' = \{s \in 2^{\mathbb{Z}} \mid s_0 = 1\}, \tag{2.27}$$

and repeats the arguments for the doubling map almost *verbatim*.

[33] See for example the movie on `https://en.wikipedia.org/wiki/Baker's_map`.

Returning to the general story: Kolmogorov's idea was to refine the partition (1.31), written as

$$\pi = \{U_a, a \in A\}; \qquad X = \bigcup_a U_a; \qquad P(U_a \cap U_b) = 0 \ (a \neq b); \qquad P(\cup_a U_a) = 1, \tag{2.28}$$

to a finer partition π^N, defined for all $N \geq 1$, by

$$\pi^N = \{U_\sigma, \sigma \in A^N\}; \tag{2.29}$$

$$U_\sigma := \bigcap_{n=0}^{N-1} T^{-n} U_{\sigma_n} = U_{\sigma_0} \cap T^{-1} U_{\sigma_1} \cap \cdots \cap T^{-(N-1)} U_{\sigma_{N-1}}$$

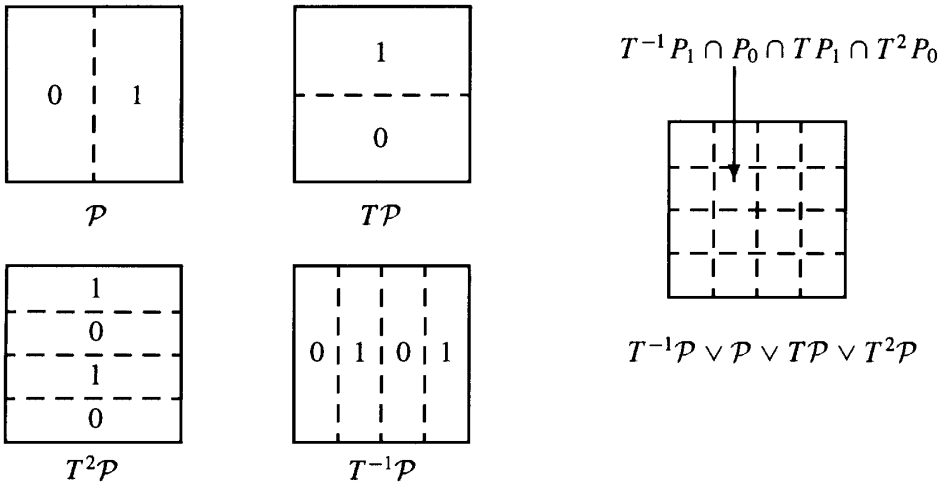
$$= \{x \in X \mid x \in U_{\sigma_0}, Tx \in U_{\sigma_1}, \ldots T^{N-1} x \in U_{\sigma_{N-1}}\}, \tag{2.30}$$

i.e. the set of all x whose coarse-grained path $\xi_N(x)$ up to time $N-1$ corresponds to $\sigma \in A^N$; here $T^0 = \mathrm{id}_X$ so that $\pi^1 = \pi$. Note that $\sigma \in A^N$ is just a label identifying such a coarse-grained path, which itself is given by the sequence $(U_{\sigma_0}, U_{\sigma_1}, \ldots U_{\sigma_{N-1}})$ of elements of the partition π in which the particle resides at time $n \in \{0, 1, \ldots, N-1\}$. The set U_σ is empty for those $\sigma \in A^N$ that (for given map T) are not in the image of the map $\xi_N : X \to A^N$. Such labels σ should be omitted in A^N, so that in general π^N corresponds to some subset of A^N. Here are (*the*) two extreme cases:

1. If $T = \mathrm{id}_X$, i.e., $T(x) = x$, then $U_\sigma \neq \emptyset$ iff $\sigma_n = a$ for all $n \in \{0, 1, \ldots, N-1\}$, for some $a \in A$.
2. If $T = S$ for $X = A^{\mathbb{N}}$, with partition $U_a = \{s \in A^{\mathbb{N}} \mid s_0 = a\}$, then all σ occur, since

$$U_\sigma = [\sigma] = \{s \in A^{\mathbb{N}} \mid s_{|N} = \sigma\}. \tag{2.31}$$

We also display a few partitions for the Baker's map (also forward in time), starting with (2.27):



Within the image of ξ_N, $\xi_N(x) \in A^N$ bijectively corresponds to the subset $U_{\xi_N(x)} \in \pi^N$. If for any partition γ we write $\gamma(x)$ for the element of γ containing x, then our notation implies that

$$\pi^N(x) = U_{\xi_N(x)}. \tag{2.32}$$

Like ξ regarding $A^{\mathbb{N}}$, each ξ_N defines a probability $P_N \in \mathrm{Prob}(A^N)$ induced by $P \in \mathrm{Prob}(X)$, i.e.

$$P_N(\sigma) = P(U_\sigma), \tag{2.33}$$

so that $P_N(\xi_N(x))$ is the probability of the truncated coarse-grained path $\xi_N(x)$. This suggests (base-e)notions of information and entropy, as follows. First, for any finite measurable partition γ of (X,P), suppressing our fixed X in the notation we define information and entropy by, cf. (1.14),

$$I_P(\gamma) := \sum_{C\in\gamma} 1_C \log\left(1/P(C)\right) = -\sum_{C\in\gamma} 1_C \log P(C); \tag{2.34}$$

$$H_P(\gamma) := \langle I_P(\gamma)\rangle_P = -\sum_{C\in\gamma} P(C)\log P(C). \tag{2.35}$$

Note that $I_P(\gamma)$ is a (measurable) function on X, so that γ is not its argument but a label, and

$$I_P(\gamma)(x) = -\log P(\gamma(x)). \tag{2.36}$$

In particular, we have

$$I_P(\pi^N)(x) = -\log P(\pi^N(x)) = -\log P(U_{\xi_N(x)}) = -\log P_N(\xi_N(x)); \tag{2.37}$$

$$H_P(\pi^N) = \langle I_P(\pi^N)\rangle_P = -\sum_{\sigma\in A^N} P(U_\sigma)\log P(U_\sigma) = -\sum_{\sigma\in A^N} P_N(\sigma)\log P_N(\sigma) = S(P_N), \tag{2.38}$$

where $S(P_N)$ is the Shannon–Boltzmann entropy (1.14) of the probability space $(A^N, \mathscr{P}(A^N), P_N)$.

For any two finite partitions γ and β of X we similarly introduce the *conditional information* and *conditional entropy* of γ given β by

$$I_P(\gamma|\beta) := -\sum_{C\in\gamma, B\in\beta} 1_{C\cap B}\log P(C|B); \tag{2.39}$$

$$H_P(\gamma|\beta) := \langle I_P(\gamma|\beta)\rangle_P = -\sum_{C\in\gamma, B\in\beta} P(C\cap B)\log P(C|B), \tag{2.40}$$

where $P(C|B) = P(C\cap B)/P(B)$ as usual, provided $P(B) > 0$. Clearly, for all $x \in X$,

$$I_P(\gamma)(x) \geq 0; \qquad I_P(\gamma|\beta)(x) \geq 0; \qquad H_P(\gamma) \geq 0; \qquad H_P(\gamma|\beta) \geq 0. \tag{2.41}$$

The fact that the set of all partitions of some given set (or measure space) X form a lattice plays a major role in the study of entropy. If γ and β are partitions of X, we define $\gamma \leq \beta$ if for all $C \in \gamma$ there is $B \in \beta$ such that $C \subset B$. In that case γ is called a *refinement* of β (for example, the partition γ of a unit square into four equal squares of size 1/2 is a refinement of its partition β into two equal columns).[34] This partial order has suprema $\vee$ and infima $\wedge$ (making it a lattice); we only need

$$\gamma\wedge\beta = \{C\cap B \mid C\in\gamma, B\in\beta\}. \tag{2.42}$$

For example, in our dynamical system (X,P,T) we have $T^{-n}\pi = \{B\subset X \mid T^n(B)\in\pi\}$ and

$$\pi^N = \wedge_{n=0}^{N-1} T^{-n}\pi = \bigcap_{n=0}^{N-1} T^{-n}\pi = \{U_\sigma, \sigma\in A^N\}, \tag{2.43}$$

cf. (2.30), where $N \geq 1$; we omit those $\sigma \in A^N$ for which $U_\sigma = \emptyset$. If $M \geq N$, then $\pi^M \leq \pi^N$.

Lemma 2.1 *1. For any two partitions γ,β of X we have*

$$H_P(\beta|\gamma) \leq H_P(\beta); \tag{2.44}$$

$$H_P(\gamma\wedge\beta) = H_P(\gamma) + H_P(\beta|\gamma), \tag{2.45}$$

and hence

$$H_P(\gamma\wedge\beta) \leq H_P(\gamma) + H_P(\beta). \tag{2.46}$$

[34] Most literature writes $\beta \leq \gamma$ in this case, but our notation makes (2.43) look better since $\wedge$ resonates with $\cap$. On the other hand, we will have $H_P(\gamma) \geq H_P(\beta)$ if $\gamma \leq \beta$, which would look better with the opposite ordering convention.

2. *If $\gamma \leq \beta$, then $H_P(\gamma) \geq H_P(\beta)$ as well as $H_P(\beta|\gamma) = 0$; and for any third partition α,*

$$H_P(\alpha|\gamma) \leq H_P(\alpha|\beta). \tag{2.47}$$

3. *For the information functions we have*

$$I_P(\gamma \wedge \beta) = I_P(\gamma) + I_P(\beta|\gamma). \tag{2.48}$$

Exercise 10 *Prove parts 1 and 2. Hint: for* (2.44) *use Jensen's inequality* (A.4). □

For example, in the unit square example (with Lebesgue measure) where β consists of two columns and γ of four equal "quarter-squares", we have $H_P(\gamma) = 2$ whilst $H_P(\beta) = 1$. Intuitively, if partition γ refines partition β and hence has more blocks, then revealing the location of a particle by specifying the block of γ it is in provides more information than revealing the corresponding black of β; and hence γ should have more entropy than β. Similarly, in dynamical systems, if $M > N$ then knowing a coarse-grained path from time $t = 0$ till time $t = M-1$ gives more information than knowing it only until time $t = N-1$, so that $H_P(\pi^M) \geq H_P(\pi^N)$.

We now show that $H_P(\pi^N)$ as defined in (2.38) has a limit

$$h_P(\pi) := \lim_{N\to\infty} \frac{1}{N} H_P(\pi^N). \tag{2.49}$$

The key to its existence is the subadditivity property

$$H_P(\pi^{M+N}) \leq H_P(\pi^N) + H_P(\pi^M), \tag{2.50}$$

which follows from Lemma 2.1. To see this, for $0 \leq k \leq N-1$ extend the notation (2.43) to

$$\pi_k^N := \wedge_{n=k}^{N-1} (T^{-n}\pi) = \bigcap_{n=k}^{N-1} (T^{-n}\pi), \tag{2.51}$$

so that e.g. $\pi_0^N = \pi^N$ and $\pi_1^{N+1} = T^{-1}\pi^N$. Eqs. (2.35) and (2.2), and subsequently (2.46), give

$$H_P(\pi_M^{N+M}) = H_P(\pi^N); \qquad \Rightarrow \tag{2.52}$$

$$H_P(\pi^{M+N}) = H_P(\pi^N \wedge \pi_N^{M+N}) \leq H_P(\pi^N) + H_P(\pi_N^{M+N}) = H_P(\pi^N) + H_P(\pi^M). \tag{2.53}$$

Lemma 2.2 (Fekete) *If (a_N) is a subadditive sequence in $\mathbb{R}^+$ in the sense that $a_{M+N} \leq a_M + a_N$, then $\lim_{N\to\infty}(a_N/N)$ exists (its value may be ∞) and equals $\inf_N(a_N/N)$.*

Exercise 11 *Prove this lemma (you may look at the internet or elsewhere!).*

There is an interesting alternative formule for (2.49), namely, using the notation (2.51),

$$H_P(\pi) = \lim_{N\to\infty} H_P(\pi|\pi_1^N). \tag{2.54}$$

First, the limit exists, because for $M > N$ the partition π_1^M is obviously finer than π_1^N (i.e. $\pi_1^M \leq \pi_1^N$), so that $H_P(\pi|\pi_1^M) \leq H_P(\pi|\pi_1^N)$ by (2.47). Moreover, by definition (2.39) we have $I_P(\gamma|\beta) \geq 0$ and hence $H_P(\gamma|\beta) \geq 0$ by (2.40), so that the sequence $(H_P(\pi|\pi_1^N))_N$ is non-increasing and bounded below, so that it converges. This implies (nontrivially) that its limit equals its Césaro limit, that is,

$$\lim_{N\to\infty} H_P(\pi|\pi_1^N) = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} H_P(\pi|\pi_1^j). \tag{2.55}$$

Eq. (2.54) then follows immediately from another nontrivial formula, viz.

$$H_P(\pi) + \sum_{j=1}^{N-1} H_P(\pi|\pi_1^{j+1}) = H_P(\pi^N). \tag{2.56}$$

Exercise 12 *Prove this by induction.*

Definition 2.3 *The* Kolmogorov–Sinai *(or* metric*)* entropy *of the dynamical system* (X,P,T) *is*

$$h(X,P,T) := \sup_{\pi} h_P(\pi), \tag{2.57}$$

where the supremum (which may be infinite) is taken over all finite measurable partitions of X*.*

Here the right-hand side tacitly depends on T via (2.49) and (2.30). The reason this lofty quantity is often accessible lies in the crucial *Kolmogorov–Sinai theorem*:

Theorem 2.4 *If* π *is a finite partition of* X *whose refinements* π^N *generate the* σ*-algebra* Σ*, i.e.,*

$$\Sigma = \sigma\left(\bigcup_{n=0}^{\infty} T^{-n}\pi\right) = \sigma(\{\pi^N, N \in \mathbb{N}_*\}), \tag{2.58}$$

on which the probability measure P *was defined in the first place, then*

$$h(X,P,T) = h(\pi). \tag{2.59}$$

We omit the proof, which is very technical, see e.g. Viana & Oliveira, Theorem 9.2.1. But the idea is simple: the assumption in the theorem means that as $N \to \infty$ the refinements π^N of π generate a maximally refined *measurable* partition of X: the cells of π^N could not eventually go beyond the elements of the given σ-algebra Σ. By Lemma 2.1.2, entropy increases under refinement and hence a maximally refined partition should have maximal entropy, reaching the supremum in (2.57).

The first nontrivial case where this entropy could be computed from Theorem 2.4 was the unilateral shift (2.5), which gives the same result as the *Bernoulli shift*, i.e. the same map S but now defined on double-sided sequences $s \in A^{\mathbb{Z}}$ instead of $A^{\mathbb{N}}$. In the latter case, the Bernoulli measure p^N on A^N with prior $p \in \mathrm{Prob}(A)$ is extended to a probability measure $p^{\mathbb{N}}$ on $A^{\mathbb{N}}$. The latter measure is defined on the σ-algebra generated by the *cylinder sets*

$$[\sigma]_N := \{s \in A^{\mathbb{N}} \mid s_{|N} = \sigma\} = \sigma A^{\mathbb{N}}, \tag{2.60}$$

where $\sigma \in A^N$ and $N \in \mathbb{N}$, so that $[\sigma]_N \subset A^{\mathbb{N}}$ is the set of all half-sided sequences with initial segment σ, and $s_{|N} = \sigma$ means $s(n) = \sigma(n)$ for all $n \in N = \{0,\ldots,N-1\}$. We then define $p^{\mathbb{N}}$ by

$$p^{\mathbb{N}}([\sigma]_N) := p^N(\sigma), \tag{2.61}$$

uniquely extended to the σ-algebra Σ generated by these cylinder sets. See chapter 3 for more information. Similarly, $\sigma \in A^{[-N,N]}$ defines $[\sigma]_N = A^{-\mathbb{N}}\sigma A^{\mathbb{N}}$, the set of all sequences in $A^{\mathbb{Z}}$ with middle segment σ. The probability measure $p^{\mathbb{Z}}$ is then defined analogously on the σ-algebra generated by these cylinder sets, where $N \in \mathbb{N}_*$ and $\sigma \in A^{[-N,N]}$. Sinai (1959) found that:

$$h(A^{\mathbb{N}}, p^{\mathbb{N}}, S) = h(A^{\mathbb{Z}}, p^{\mathbb{Z}}, S) = S(p) := -\sum_{a\in A} p(a)\log p(a). \tag{2.62}$$

To see this, let $X = A^{\mathbb{N}}$, $P = p^{\mathbb{N}}$, and $T = S$, cf. (2.5); the double-sided case is similar. Define

$$U_a := [a]_1 = \{s \in A^{\mathbb{N}} \mid s_0 = a\}, \tag{2.63}$$

which gives the starting partition $\pi = \{U_a\}_{a\in A}$ of $A^{\mathbb{N}}$, cf. (2.29) - (2.28). It follows that

$$U_\sigma = [\sigma]_N; \qquad\qquad \pi^N = \{[\sigma]_N, \sigma \in A^N\}, \tag{2.64}$$

cf. (2.30). Therefore, π is such that (2.58) and hence (2.59) holds. Now finish:

Exercise 13 *Prove* (2.64) *and on that basis derive the key step in the proof of* (2.62)*, namely*

$$H_{p^{\mathbb{N}}}(\pi^N) = NS(p). \tag{2.65}$$

Finally, use this result with (2.49) *and* (2.59) *to derive* (2.62)*.*

3 Markov chains

The fourth and most general interpretation of our diagram is that it provides a representation of discrete-time stochastic processes. This approach again goes back to Kolmogorov (1933).

A *stochastic process* is a function X of two variables $t \in T$ and $\omega \in \Omega$, where (Ω, Σ, P) is some probability space and $X_t(\omega) \in A_t$, where (A_t, Σ'_t) is an appropriate measure space and each

$$X_t : \Omega \to A_t \tag{3.1}$$

is measurable. In what follows we assume $(A_t, \Sigma'_t) = (A, \Sigma')$ to be independent of t. The simplest cases have discrete time, that is, $T = \mathbb{N}$ (the one-sided case) or $n \in \mathbb{Z}$ (two-sided), and $X_n : \Omega \to A$, where (Ω, Σ, P) is a probability space and A is finite. In those cases, the *Kolmogorov representation theorem* discussed below states that without loss of generality one may assume that:

$$\Omega = A^{\mathbb{N}} \text{ or } \Omega = A^{\mathbb{Z}}; \qquad X_n(s) = s_n; \qquad s : \mathbb{N} \to A \text{ or } s : \mathbb{Z} \to A. \tag{3.2}$$

Our diagram then displays a (truncated) sample path of a discrete-time stochastic process. This theorem relies on *Kolmogorov's extension lemma* (and is often conflated with it). It takes no extra effort to discuss these theorems for arbitrary index sets T and quite general state spaces (A, Σ'), which for us are "at worst" Polish spaces.[35] The reader may just think of finite A with $\Sigma' = \mathscr{P}(A)$. As usual, A^T consists of all functions $s : T \to A$. We define first the "right" σ-algebra:[36]

Definition 3.1 *The* cylindrical σ-algebra $\mathscr{F}$ *on* A^T *is the smallest* σ*-algebra that makes the coordinate functions (= evaluation maps)* $s \mapsto s(t)$ *from* A^T *to* A *measurable for all* $t \in T$.

Equivalently (as follows from the definition of measurability of functions between measure spaces),[37] for each $B \in \Sigma'$ and $t \in T$ the so-called *cylinder set* $[B]_t := \{s \in A^T \mid s(t) \in B\}$ must be measurable, so that $\mathscr{F}$ is the σ-algebra generated by these cylinder sets.[38] Alternatively, for any *finite* subset $F \subset T$ we define $\mathscr{F}_F \subset \mathscr{P}(A^T)$ as the σ-algebra generated by the ("rectangular") cylinder sets

$$\left[\prod_{t\in F} B_t\right]_F := \{s \in A^T \mid \forall_{t\in F} s_t \in B_t\} = \left\{ s \in A^T \mid s_{|F} \in \prod_{t\in F} B_t \right\} \qquad (B_t \in \Sigma', t \in F). \tag{3.3}$$

Then $\mathscr{F}$ is the σ-algebra generated by all $\mathscr{F}_F$ ($F \subset T$ finite). This is useful, if only because of:

Theorem 3.2 *Any probability measure* P *on* $\mathscr{F}$ *is uniquely determined by its values on* $\bigcup_F \mathscr{F}_F$*, and even by its values on all cylinder sets* (3.3)*, where* $F \subset$ *is finite and* $B_t \in \Sigma'$ $(t \in F)$.

[35] See e.g. Dudley (1989), §8.2 and 12.1, or Klenke (2020), §14.1. In the most general setting (Dudley, 1989, Theorem 12.1.2; Klenke, 2020, Theorem 14.39) each measure space (A_t, Σ'_t) is merely supposed to be isomorphic to some Borel set in $\mathbb{R}$. This includes all so-called *standard* measure spaces (this means isomorphic to: $[0,1]$ with Lebesgue measure, or a finite or countable set of atoms, or a disjoint union of both) and hence also all Polish spaces.

[36] Likewise, the canonical (= product) topology on A^T, where $(A, \mathscr{O})$ is a topological space and T is just a set, is the coarsest (= smallest) topology that makes all coordinate functions continuous. This is the topology in which, famously, by Tychonoff's theorem A^T is compact whenever A is compact (for any T!).

[37] If (X_1, Σ_1) and (X_1, Σ_2) are measure spaces, $f : X_1 \to X_2$ is measurable if for each $B \in \Sigma_2$ we have $f^{-1}(\Sigma_2) \in \Sigma_1$.

[38] Any collection of subsets $S \subset \mathscr{P}(X)$ of X generates a σ-algebra $\Sigma(S)$ on X, namely the smallest σ-algebra on X (i.e., in $\mathscr{P}(X)$) that contains S; thus $\Sigma(S)$ is just the intersection of all σ-algebras on X that contain S, and this is non-empty because $\mathscr{P}(X)$ obviously contains S and is a σ-algebra. We say that $\Sigma(S)$ is the σ-algebra *generated by* S.

The best-known example is the *Borel* σ-algebra, which is generated by some topology on X. In fact, the cylindrical σ-algebra $\mathscr{F}$ on A^T is a special case of this, using the product topology (which has the same basis).

Since $\bigcup_F \mathscr{F}_F$ is an algebra, the first part follows from the Carathéodory extension theorem,[39] but the very useful second part does not.[40] If we (tacitly) define the appropriate σ-algebra Σ'_F on A^F as the one from Definition 3.1, with $T \rightsquigarrow F$, we may also define more general cylinder sets in A^T by replacing the rectangular sets $\prod_{t\in F} B_t$ in (3.3), by arbitrary sets $\mathscr{B} \in \Sigma'_F$. Then $\mathscr{F}_F$ simply consists of all $[\mathscr{B}]_F$, where $\mathscr{B} \in \Sigma'_F$; in particular, this is already a σ-algebra.

This simplifies if A is finite with $\Sigma' = \mathscr{P}(A)$: for any $\sigma \in A^F$, define

$$[\sigma]_F := \{s \in A^T \mid \forall_{t\in F}(s(t) = \sigma(t))\} = \{s \in A^T \mid s_{|F} = \sigma\}. \tag{3.4}$$

Then it is easy to see that $\mathscr{F}_F$ is the σ-algebra generated by all such sets, and, invoking Theorem 3.2, any probability measure P on $\mathscr{F}$ is uniquely determined by its values on the sets $[\sigma]_F$.

If also $T = \mathbb{N}$, one may even restrict attention to the sets $F = N = \{0,1,\ldots,N-1\} \subset \mathbb{N}$ and ensuing cylinder sets (2.60), with σ-algebras $\mathscr{F}_N$ generated by these $[\sigma]_N$, since both $\bigcup_F \mathscr{F}_F$ and $\bigcup_N \mathscr{F}_N$ contain the following elementary cylinder sets which by Definition 3.1 also generate $\mathscr{F}$:

$$[a]_n := \{s \in A^{\mathbb{N}} \mid s_n = a\}, \;\; n \in \mathbb{N}, a \in A. \tag{3.5}$$

In fact, $\bigcup_N \mathscr{F}_N$ consists of all finite unions of the sets $[\sigma]_F$ (and is an algebra).[41] The simplest example of this construction is $A = 2 = \{0,1\}$ and $P = p^{\mathbb{N}}$ for any $p \in \mathrm{Prob}(2)$, cf. (2.61).

Exercise 14 *For $f(0) = f(1) = 1/2$, show that the binary expansion* (2.10) *induces an isomorphism* of probability spaces *(where μ_L is Lebesgue measure defined on the usual Borel sets $\mathscr{B}$):*[42]

$$(2^{\mathbb{N}}, \mathscr{F}, f^{\mathbb{N}}) \cong ([0,1], \mathscr{B}, \mu_L). \tag{3.6}$$

Let us return to Theorem 3.2. Note that A^F is not a subset of A^T, but that there are natural maps

$$\pi_F : A^T \to A^F; \qquad \pi_F(s) = s_{|F}; \tag{3.7}$$

$$\pi_{GF} : A^G \to A^F; \qquad \pi_{GF}(\sigma) = \sigma_{|F}, \tag{3.8}$$

where $F \subset G \subset T$, *in which F and G are always finite* in what follows, although the maps π_{GF} are defined more generally. It is clear from (3.7) that any $P \in \mathrm{Prob}(A^T)$ induces $p_F \in \mathrm{Prob}(A^F)$ via

$$p_F := \pi_F^{-1} P, \tag{3.9}$$

for each $F \subset T$. For $\mathscr{B} \in \Sigma'_F$, e.g. the rectangular set $\prod_{t\in F} B_t$ as in (3.3), eq. (3.9) gives

$$p_F(\mathscr{B}) = P([\mathscr{B}]_F) = P(s_{|F} \in \mathscr{B}). \tag{3.10}$$

For example, for A finite, $T = \mathbb{N}$, $F = N$ and $\mathscr{B} = \{\sigma\}$ with $\sigma \in A^N$, eq. (3.10) simply reads

$$p_N(\sigma) \equiv p_N(\{\sigma\}) = P([\sigma]_N). \tag{3.11}$$

[39] See e.g. Klenke (2020), Theorem 1.41. Carathéodory's theorem states that if a σ-algebra Σ is generated by an algebra $R \subset \Sigma$ (see footnote 28), then any countably additive set function on R taking values in $[0,\infty]$ (i.e. any premeasure) uniquely extends to a measure on Σ; this trivially implies that P is uniquely determined by its values on R.

[40] See Klenke (2020), Theorem 14.12.

[41] See footnotete 28 for the definition of al algebra of subsets.

[42] This is false in topology, where $2^{\mathbb{N}}$ is a Cantor space! The proof requires a generalization of the theorem in footnote 28. Let (X_1,Σ_1,P_1) and (X_2,Σ_2,P_2) be probability spaces (or more generally just measure spaces), let $R_2 \subset \Sigma_2$ be a *generating semi-algebra* for Σ_2, that is, $\emptyset \in R_2$; if $A,B \in R_2$ then $A \cap B \in R_2$; and if $B \in R_2$ then $X \backslash B$ is a *countable* disjoint union of elements of R_2; and R_2 generates Σ_2. Also assume that R_2 contains an increasing sequence (B_n) such that $X_2 = \bigcup_n B_n$. Then: if $f : X_1 \to X_2$ a map that satisfies $f^{-1}(B) \in \Sigma_1$ and $P_1(f^{-1}(B)) = P_2(B)$ for each $B \in R_2$, then f is measurable and $P_1(f^{-1}(B)) = P_2(B)$ for each $B \in \Sigma_2$. See e.g. Dajani, & Kalle (2021), Theorem 12.3.1.

In general, since $\pi_F = \pi_{GF} \circ \pi_G$ if $F \subset G \subset T$, the (p_F) satisfy the *consistency condition*

$$p_F = \pi_{GF}^{-1} p_G. \tag{3.12}$$

The *Kolmogorov extension lemma* turns the consistency condition (3.12) on its head:

Lemma 3.3 (Kolmogorov) *If a family of probabilities (p_F) satisfies* (3.12)*, $p_F \in \mathrm{Prob}(A^F, \Sigma_F')$, there is a unique probability measure P on $(A^T, \mathscr{F})$ that induces the given family (p_F) via* (3.9)*.*

The idea of the proof is deceptively simple: the cylinder sets $[\mathscr{B}]_F$, where $\mathscr{B} \in \Sigma_F'$, form an algebra $\bigcup_F \mathscr{F}_F$ in $\mathscr{P}(A^T)$, and one defines P on this algebra by reading (3.10) from right to left; this move is of course only valid if the consistency condition (3.12) holds. Since $\mathscr{F}$ is by definition equal to the σ-algebra generated by $\bigcup_F \mathscr{F}_F$, the Carathéodory extension theorem does the rest *provided P is countably additive on $\bigcup_F \mathscr{F}_F$*. This is the difficult technical part we omit; some extra assumptions on (A, Σ'), which so far was completely arbitrary, are needed for this step (e.g., being Polish).[43]

Exercise 15 *1. Show that for finite A and $T = \mathbb{N}$ the consistency condition* (3.12) *implies*

$$p_N(\sigma) = \sum_{a \in A} p_{N+1}(\sigma a), \tag{3.13}$$

for all $N \in \mathbb{N}$, $\sigma \in A^N$ and $a \in A$, where $\sigma a \in A^{N+1}$ is the string defined by

$$(\sigma a)_n = \sigma(n) \equiv \sigma_n \ \ (n = 0, \ldots, N-1; \qquad\qquad (\sigma a)_N = a. \tag{3.14}$$

2. Show that a family (p_N) that satisfies (3.13)*, is equivalent to a single function $p : A^* \to [0,1]$, where $A^* = \bigcup_{N \in \mathbb{N}} A^N$ is the set of all finite A-valued strings, including A^0 consisting of the empty string ε, that satisfies $p(\sigma) = \sum_{a \in A} p(\sigma a)$ and $p(\varepsilon) = 1$.*

3. Using Lemma 3.3, show that such a function p (or the equivalent (p_N) in part 1) already uniquely determines a probability measure $P \in \mathrm{Prob}(A^{\mathbb{N}})$ that returns the p_N via (3.11)*.*

Theorem 3.4 (Kolmogorov) *Let $(X_t)_{t \in T}$ be a stochastic process taking values in (for example) a Polish state space (A, Σ'), so that $X_t : \Omega \to A$ for some probability space (Ω, Σ, P). Then without loss of generality one may realize the process as evaluation maps X_t' on $(A^T, \mathscr{F})$, that is,*

$$X_t' : A^T \to A; \qquad\qquad X_t'(s) = s_t; \qquad\qquad s : T \to A, \tag{3.15}$$

with a probability measure $P' \in \mathrm{Prob}(A^T, \mathscr{F})$ that returns the joint probabilities of the (X_t) via

$$P(X_t \in B_t, t \in F) = P\left(\left[\prod_{t \in F} B_t \right] \right) = P'(X_t' \in B_t, t \in F), \tag{3.16}$$

for each finite $F \subset T$ and $(B_t)_{t \in F}$, ($B_t \in \Sigma'$), with ensuing $\prod_{t \in F} B_t \in \Sigma_F'$, cf. (3.3)*.*

Here the symbolic notation $P(X_t \in B_t, t \in F)$ stands for $P(\{\omega \in \Omega \mid \forall_{t \in F}(X_t(\omega) \in B_t)\})$, and likewise $P'(X_t' \in B_t, t \in F)$ abbreviates $P'(\{s \in A^T \mid \forall_{t \in F}(s_t \in B_t)\})$.

Proof. For each finite $F \subset T$, define $p_F \in \mathrm{Prob}(A^F, \Sigma_F')$ via a little variation on (3.10),[44] namely

$$p_F\left(\prod_{t \in F} B_t \right) := P(X_t \in B_t, t \in F). \tag{3.17}$$

A technical argument then shows that this has a unique extension to $p_F \in \mathrm{Prob}(A^T, \Sigma_F')$.[45]

[43] Since each p_F is given as a probability measure, we know that P is countably additive on each $\mathscr{F}_F$, which as we already mentioned is a σ-algebra, but this is not enough for countable additivity on the in ite union $\bigcup_F \mathscr{F}_F$. The proof is even difficult for A finite and $T = \mathbb{N}$, which case is spelled out by Calude (2002), Theorem 1.7.

[44] The difference between (3.17) and (3.10) lies in the fact that the latter defines p_F via $P \in \mathrm{Prob}(A^T)$, whereas in the former we are given $P \in \mathrm{Prob}(\Omega)$ and attempt to construct $P' \in \mathrm{Prob}(A^T)$ from this P, namely via (3.17).

[45] See Klenke (2020), Theorem 14.12.(iii).

Exercise 16 *Show that the ensuing family* (p_F) *satisfies the consistency condition* (3.12).

Lemma 3.3 then gives us the required distribution P', upon which (3.16) follows from (3.10). □

For finite A and $T = \mathbb{N}$, eqs. (3.16) - (3.17) simply comes down to, for each $N \in \mathbb{N}$ and $\sigma \in A^N$,

$$P(X_0 = \sigma_0, \ldots, X_{N-1} = \sigma_{N-1}) = P(s_{|N} = \sigma) = P([\sigma]_N) = p_F(\sigma). \tag{3.18}$$

We look at the following special cases (with A always finite), where as usual, provided $P(C) > 0$, conditional probabilities are defined

$$P(B|C) = \frac{P(B,C)}{P(C)}. \tag{3.19}$$

- In the original description, the (X_n) are *independent* iff any finite subset is independent (with respect to P). In particular, for each $n \in \mathbb{N}$, $a \in A$ and $\sigma = (\sigma_0, \ldots, \sigma_{N-1}) \in A^N$,

  $$P(X_N = a | X_0 = \sigma_0, \ldots, X_{N-1} = \sigma_{N-1}) = P(X_N = a). \tag{3.20}$$

  In the Kolmogorov model (3.20) means that

  $$P(s_N = a | s_{|N} = \sigma) \equiv P(s_N = a | s_0 = \sigma_0, \ldots, s_{N-1} = \sigma_{n-1}) = P(s_N = a). \tag{3.21}$$

- The (X_n) are *identically distributed* (i.d.) if $P(X_N = a)$ is independent of N and hence equal, for all $N \in \mathbb{N}$, to $p(a)$ for some $p \in \mathrm{Prob}(A)$, and *i.i.d.* if they are both independent and identically distributed. In that case,[46]

  $$P = p^{\mathbb{N}}. \tag{3.22}$$

- In the next step of generality, the (X_n) form a *Markov chain* if, in the original model,

  $$P(X_N = a | X_0 = a_0, \ldots, X_{N-1} = a_{N-1}) = P(X_N = a \mid X_{N-1} = a_{N-1}), \tag{3.23}$$

  or, similarly to (3.21), in the Kolmogorov model, for all $N \in \mathbb{N}$, $a \in A$, and $\sigma \in A^N$,

  $$P(s_N = a | s_{|N} = \sigma) = P(s_N = a | s_{N-1} = \sigma_{N-1}), \tag{3.24}$$

- A Markov chain is *stationary* iff $P(X_N = a \mid X_{N-1} = b)$ does not depend on N, so that

  $$P(X_0 = a_0, X_1 = a_1, \ldots, X_N = a_N) = P(X_M = a_0, X_{M+1} = a_1, \ldots, X_{M+N} = a_N) \tag{3.25}$$

  for all $M > 0$. We may then conveniently introduce *transition probabilities* by

  $$P_{ab} := P(X_{N+1} = b \mid X_N = a) = P(X_1 = b \mid X_0 = a). \tag{3.26}$$

  This is a *stochastic matrix*, in that all entries are non-negative and

  $$\sum_b P_{ab} = 1 \tag{3.27}$$

  for each a. In a stationary Markov chain $P(X_n = a)$ is independent of n, and we write

  $$p(a) \equiv p_a := P(X_n = a) = P(X_0 = a). \tag{3.28}$$

  For example, for $P = p^{\mathbb{N}}$ we have $P_{ab} = p_b$ for all a.

[46] This may be seen as a repeated sampling of (A, p), where X_n is the n'th try. If (X, P, T) is a dynamical system and $f : X \to \mathbb{R}$ or $f : X \to A$ is a random variable, then the process $X_n = f \circ T^n$ is i.d. (since T preserves P) but not i.i.d.

Exercise 17 *1. Show that in a stationary Markov chain (with finite state space A):*[47]

$$P(X_N = b \mid X_0 = a) = (P^N)_{ab}; \tag{3.29}$$

$$P(X_0 = a_0, X_1 = a_1, \ldots, X_{N-1} = a_{N-1}, X_N = a_N) = p_{a_0} P_{a_0 a_1} \cdots P_{a_{N-1} a_N}; \tag{3.30}$$

$$P(X_1 = b, X_2 = c) = p_b P_{bc}; \tag{3.31}$$

$$\sum_a p_a P_{ab} = p_b. \tag{3.32}$$

2. Conversely, show that any stochastic matrix (P_{ab}) plus some probability distribution $p \in \mathrm{Prob}(A)$ that satisfies (3.32) *defines a stationary Markov chain via* (3.30)*.*

Stationarity also implies that if $p(a) = 0$ for some $a \in A$, then any occurrence $X_n = a$ has zero probability, so that we may remove a from A. In what follows we therefore assume that $p(a) > 0$ for each $a \in A$. As a check on (3.32), we already noted that for the i.i.d. case $P = p^{\mathbb{N}}$ we have $P_{ab} = p_b$, so that (3.32) reduces to $\sum_a p_a p_b = p_b$.

- The existence of a left-eigenvector p of a given stochastic matrix (P_{ab}) is guaranteed by the *Perron–Frobenius theorem*, but this is not enough to conclude $p \in \mathrm{Prob}(A)$. In the cleanest case, which is central to both ergodic theory and large deviation theory, (P_{ab}) is *irreducible*:

Definition 3.5 *1. A stationary Markov chain is called* irreducible *if for all $a, b \in A$ there is $N \in \mathbb{N}$ such that $P(X_N = b | X_0 = a) > 0$.*

2. Equivalently, its stochastic matrix (P_{ab}) (and generally a matrix with nonnegative real entries) is irreducible *if for all $a, b \in A$ there is $N \in \mathbb{N}$ such that $(P^N)_{ab} > 0$.*

3. And this is equivalent to existence, for all $a, b \in A$, of indices $(a_0, \ldots, a_N)$ with $a_0 = a$ and $a_N = b$ such that $P_{a_i a_{i+1}} > 0$ for each $i = 0, \ldots, N-1$.

Exercise 18 *1. Prove that these criteria are indeed equivalent.*

2. Is a Markov chain defined by the i.i.d. case $P = p^{\mathbb{N}}$ irreducible?

In studying the relationship between irreducibility, ergodicity and large deviations later on the following well-known linear algebra theorem will be used; but it also clarifies Definition 3.5.

Theorem 3.6 (Perron–Frobenius) *Let (P_{ab}) be an irreducible real matrix with each $P_{ab} \geq 0$.*

1. The matrix (P_{ab}) has a unique nondegenerate real eigenvalue ρ with the property that any corresponding left eigenvector p has strictly positive entries $p_a > 0$.

2. If the matrix (P_{ab}) is stochastic, then $\rho = 1$, and we may normalize its eigenvectors p to achieve $p \in \mathrm{Prob}(A)$; the stationarity condition (3.32) *then holds by construction.*

3. For every strictly positive vector $v \in \mathbb{R}^{|A|}$ (i.e. $v_a > 0$ for each a), and arbitrary $b \in A$,

$$\lim_{N \to \infty} \frac{1}{N} \log \left(\sum_{a \in A} v_a (P^N)_{ab} \right) = \log \rho. \tag{3.33}$$

[47] Physicists: be warned that repeated indices are *not* summed unless there is an explicit sumation sign!

4. *The corresponding (one-dimensional) spectral projection $E^{(\rho)}$ for ρ is given by*

$$E^{(\rho)}_{ab} = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \rho^{-n} P^n_{ab} \tag{3.34}$$

In particular, in the irreducible stochastic case $p \in \mathrm{Prob}(A)$ is given for arbitrary $a \in A$ by

$$p_b = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} P^n_{ab}. \tag{3.35}$$

5. *If there exists $M > 0$ such that $P^M_{ab} > 0$ for all $a, b \in A$ (in which case (P_{ab}) is irreducible),*[48] *then the Césaro limit in* (3.35) *may be replaced by the ordinary limit*

$$p_b = \lim_{N\to\infty} (P^N)_{ab}. \tag{3.36}$$

6. *Conversely, if (P_{ab}) is stochastic then the limit* (3.35) *exists; if the resulting vector p: (i) lies in* $\mathrm{Prob}(A)$ *and (ii) solves* (3.32)*, then (P_{ab}) is irreducible.*

For example, in the i.i.d. case $P = p^{\mathbb{N}}$ we have $P_{ab} = p_b$ for all a, so that $P^2 = P$ and hence the limit in (3.35) equals p. The ensuing Markov chain is irreducible provided $p_a > 0$ for each $a \in A$.

An interesting stationary irreducible Markov chain that is not i.i.d. is the *Ehrenfest urn model*.[49] This model has $A = \{0, 1, \ldots, M\}$, so that $a \in A$ is a number $0 \le m \le M$, seen as the number of balls (or fleas) on one of two urns (or dogs), labeled U_1 (the other is called U_0). The random variable X_n describes the number of balls in U_1 a time n (so that if $X_n = m$, then $M - m$ is the number of balls in U_0). The transition probabilities are given by:

$$P_{m,m-1} := \frac{m}{M} \ (m \ge 1); \qquad P_{m,m+1} := \frac{M-m}{M} \ (m < M); \qquad P_{mn} = 0 \ (n \ne m \pm 1). \tag{3.37}$$

The idea is that at each time step n some ball is randomly and fairly chosen from the set of M balls, and is then moved to the other urn. If U_1 contains m balls the probability that the chosen ball is in U_1 is m/M, in which case the total number of balls in U_1 after the jump equals $m-1$. Conversely, the probability that the chosen ball is in U_0 equals $1-(m/M)$, and after its jump to U_1 the total number of balls in U_1 is $m+1$. For example, for $M = 4$ the transition probability matrix equals

$$(P_{mn}) = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 3/4 & 0 & 1/4 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{3.38}$$

Check that this matrix is indeed stochastic: each row adds up to unity, cf. (3.27). Irreducibility is easy to see without any computation from part 1 of Definition 3.5: one simply imagines possible transfers needed to get from any number of balls on U_1 to any other number. The key result is that the stationary probability $p \in \mathrm{Prob}(M)$ is given by the binomial distribution

$$p_m = 2^{-M} \begin{pmatrix} M \\ m \end{pmatrix}, \tag{3.39}$$

[48] We have $P^M_{ab} > 0$ for all a, b for some $M > 0$ iff (P_{ab}) is irreducible and *aperiodic*, i.e., for each $a \in A$ we have $d(a) := \gcd\{k \in \mathbb{N}_* \mid P^k_{aa} > 0\} = 1$. In the presence of irreducibility, $d(a) = 1$ needs to be checked just for a single a.

[49] This model was introduced by Ehrenfest & Ehrenfest (1907ab) in order to explain Boltzmann's views on the second law of thermodynamics, a topic to which we shall return. The literature on the Ehrenfest model (not to speak of its wider context) is huge; Bricmont (2022), §8.7.1, is a good starting point.

which is the probability distribution if the balls were randomly distributed over the two urns in the first place. If M is even, the value $m = M/2$ has the highest probability and for large M all values far away are increasingly unlikely.

Exercise 19 *Prove* (3.39).

There is also a finer version of the Ehrenfest model, in which $A = 2^M$ and each $a \in A$ gives the detailed distribution of the balls (i.e., its microstate): a state a now specifies $a_k \in \{0,1\}$ for each $k = 0, \dots M-1$, which means that ball number k is in urn a_k. We now have $P_{ab} > 0$ iff $b_k = a_k$ for all k except for one value k' (so that the jump flips the bit $a_{k'}$), in which case $P_{ab} = 1/M$. Since for each b there are M microstates a that differ from b by exactly one bit, this matrix duly satisfies (3.27). For the same reason the stationary probability $p \in \mathrm{Prob}(2^M)$ is given by the flat one, i.e.,

$$p_a = 2^{-M} \text{ for each } a \in 2^M. \tag{3.40}$$

Exercise 20 *Show that* (3.40) *satisfies* (3.32).

Finally, in the Kolmogorov model stationarity is equivalent to shift-invariance of P, that is,

$$P(S^{-1}(B)) = P(B) \tag{3.41}$$

for any cylinder (or measurable) set $B \subset A^{\mathbb{N}}$, where the shift map was defined in (2.5). In the one-sided case $\Omega = A^{\mathbb{N}}$ the shift S is not invertible, but in the two-sided case $\Omega = A^{\mathbb{Z}}$ it is, and the shift map extends to a $\mathbb{Z}$-action on $A^{\mathbb{Z}}$, where $\mathbb{Z}$ is seen as a group in the usual way (that is, for $n \in \mathbb{Z}$ and $s \in A^{\mathbb{Z}}$ we define $n \cdot s := S^n(s)$). Either way, we enter the realm of *ergodic theory*, which is the study dynamical systems (X,P,T) that satisfy (2.2) and the measure-preserving map $T : X \to X$ is *ergodic* (relative to P). For us, ergodic theory is useful both as a tool for deriving the strong (and hence the weak) law of large numbers (which is the beginning of large deviation theory), and as a method for studying chaotic dynamical systems (in connection with entropy).

4 Ergodic theory

Ergodic theory goes back to Boltzmann in his work on statistical mechanics, but the corresponding mathematical theory was primarily developed by Birkhoff (senior) and von Neumann.[50] From a probabilistic point of view, we already saw that if (X,P,T) is a dynamical system and $f : X \to \mathbb{R}$ s a random variable, then the process $X_n = f \circ T^n$ is i.d. but not i.i.d. The pointwise ergodic theorem below then generalizes the strong law of large numbers from i.i.d. to i.d. variables, at some price:

Definition 4.1 *Let (X,P,T) be a dynamical systems, i.e. a probability space (X,P) with measure-preserving map $T : X \to X$. This map is* ergodic *if $T^{-1}(B) = B$ implies $P(B) = 0$ or $P(B) = 1$.*

Note that the logical structure is: $\forall_{B \in \Sigma}((T^{-1}B = B) \to (P(B) = 0 \vee P(B) = 1))$, and hence non-ergodicity is equivalent to the existence of some $B \in \Sigma$ for which $T^{-1}(B) = B$ and $0 < P(B) < 1$. Up to measure zero sets, $P(B) = 0$ means that $B = \emptyset$, whereas $P(B) = 1$ means that $B = X$, so that ergodicity essentially means that the only nontrivial T-invariant subset $B \subset X$ is X.

By iteration of the antecedent, Definition 4.1 implies that (X,P,T) is ergodic iff $T^{-n}B = B$ for all $n \in \mathbb{N}$ implies $P(B) = 0$ or $P(B) = 1$. In T is invertible, it gives rise to a $\mathbb{Z}$-action on X via $n \cdot x := T^n x$ (where $\mathbb{Z}$ is seen as an additive group). From that point of view, ergodicity is equivalent to: All $\mathbb{Z}$-invariant sets B satsisfy $P(B) = 0$ or $P(B) = 1$.

[50] See Halmos (1958), von Plato (1994) and Moore (2015) for some history of ergodic theory. For the theory itself we recommend Tao (2007) and in more detail and polish Dajani & Kalle (2021) for student-friendly first introductions, and Viana & Oliveira (2016) for a full meal. Walters (1982) is intermediate and still useful.

Exercise 21 *Show that ergodicity is equivalent to any of the following properties:*[51]

1. $T^{-1}(B) \subset B$ *implies* $P(B) = 0$ *or* $P(B) = 1$.
2. $T^{-1}(B) \supset B$ *implies* $P(B) = 0$ *or* $P(B) = 1$.
3. $P(T^{-1}(B) \Delta B) = 0$ *implies* $P(B) = 0$ *or* $P(B) = 1$.

The simplest nontrivial examples of ergodic systems are given by certain operations on the torus

$$\mathbb{T} := \{z \in \mathbb{C} \mid |z| = 1\}, \tag{4.1}$$

parametrized by $z = \exp(ix)$, and equipped with normalized Lebesgue measure (i.e. Haar measure)

$$dP(z) = \frac{dz}{2\pi i z}; \qquad\qquad dP(x) = \frac{dx}{2\pi}, \tag{4.2}$$

so that $P(\mathbb{T}) = 1$. The following result is similar to the doubling map and the tent map.

Exercise 22 *Show that Lebesgue measure on* $\mathbb{T}$ *(which is the group-theoretic Haar measure) is invariant under rescaling by an integer* $p \neq 0$.

The cases of interest are *irrational rotations* and *scalings*, where $\alpha \in [0, 2\pi)$ and $p \in \mathbb{Z}_*$:

$$R_\alpha : \mathbb{T} \to \mathbb{T}; \qquad z \mapsto e^{i\alpha} z; \qquad x \mapsto x + \alpha \mod 2\pi \ \ (x \in [0, 2\pi); \tag{4.3}$$

$$M_p : \mathbb{T} \to \mathbb{T}; \qquad z \mapsto z^p; \qquad x \mapsto px \mod 2\pi \tag{4.4}$$

Proposition 4.2 $(\mathbb{T}, P, R_\alpha)$ *is ergodic iff* $\alpha/(2\pi)$ *is irrational, and* $(\mathbb{T}, P, M_p)$ *is ergodic iff* $|p| \geq 2$.

Although this can be proved directly from Definition 4.1,[52] it is instructive to give an analytic proof based on a reformulation of ergodicity that will often be used.

Proposition 4.3 *Let* (X, P, T) *be a dynamical system. The following conditions are equivalent, and hence* (X, P, T) *is ergodic iff any of these conditions hold:*

1. *For all measurable* $B \subset X$*: If* $T^{-1}(B) = B$*, then* $P(B) = 0$ *or* $P(B) = 1$.
2. *For all* $f \in L^2(X, P)$*: If* $f(Tx) = f(x)$ *P-a.e., then* f *is constant P-a.e.*
3. *For all* $f \in L^1(X, P)$*: If* $f(Tx) = f(x)$ *P-a.e., then* f *is constant P-a.e.*
4. *For all measurable* $f : X \to \mathbb{R}$*: If* $f(Tx) = f(x)$ *P-a.e., then* f *is constant P-a.e.*
5. *There exists no nontrivial convex decomposition (the trivial case being* $P_1 = P_2 = P$*)*

$$P = tP_1 + (1-t)P_2, \tag{4.5}$$

where $t \in (0, 1)$ *and* P_1 *and* P_2 *are also* T*-invariant probability measures.*[53] *That is,*

$$P \in \partial_e \mathrm{Prob}^T(X). \tag{4.6}$$

[51] Recall that for us $\subset$ means $\subseteq$, and that the *symmetric difference* of $B, C \subset X$ is $B \Delta C := (B \backslash C) \cup (C \backslash B)$.

[52] See e.g. Shields, Prop. I.2.14. His proof relies on *Kronecker's theorem*: the orbit $(R_\alpha^N)_{N \in \mathbb{N}}$ is dense in $\mathbb{T}$ iff α is irrational. This was also Boltzmann's original intuition about ergodicity.

[53] Here the space $\mathrm{Prob}(X)$ of all probability measures on X is seen as a convex set, see Definition A.20. If X is compact, the so is $\mathrm{Prob}(X)$ (in the weak or vague topology), so that its closed subspace $\mathrm{Prob}^T(X)$ of all T-invariant probability measures is also a compact convex set. Within the triple (X, P, T), we here vary P rather than T.

Here $\mathrm{Prob}^T(X)$ is the set of T-invariant probability measures on X, and we take the L^p-spaces to consist of real-valued, but could equally well use $\mathbb{C}$ instead of $\mathbb{R}$.

Proof. The implication $4 \to 3$ is trivial, and $3 \to 2$ follows from easy functional analysis:

Exercise 23 *Prove* $3 \to 2$ *by showing that* $L^2(X,P) \subset L^1(X,P)$ *for probability measures* P.

The real idea is in the implications $2 \to 1$ and $1 \to 4$, which rely on three simple facts:

1. $1_B \circ T = 1_{T^{-1}B}$.
2. $1_{T^{-1}B} = 1_B$ P-a.e. iff $T^{-1}B = B$ (up to P-measure zero).
3. 1_B is constant P-a.e. iff $P(B) = 0$ or $P(B) = 1$.

This gives the implication $2 \to 1$ (see exercise). The implication $1 \to 4$ contains a technical part we omit, and an easy part based on the above facts. The easy part concerns the case $f = \sum_i c_i 1_{B_i}$. The technical part (see e.g. Walters, 1982, Theorem 1.6) is that since f is measurable, it may be approximated by such (finite) sums, and the easy part survives this approximation procedure.

Exercise 24 *Prove the above facts, and from these, prove* $2 \to 1$ *as well as the easy part of* $1 \to 4$.

We prove $5 \to 1$ contrapositively (so we prove $\neg 1 \to \neg 5$). If P is not ergodic ($\neg 1$), by Definition 4.1 there is is $B \subset X$ such that $0 < P(C) < 1$. But this implies $\neg 5$, since we may put:

$$P = tP_1 + (1-t)P_2; \qquad t := P(B); \qquad P_1(C) := P(C|B); \qquad P_2(C) := P(C|B^c). \tag{4.7}$$

The converse $1 \to 5$ relies on a lemma of independent interest (with a luxury bonus part).[54]

Lemma 4.4 *If* $P, Q \in \mathrm{Prob}^T(X)$ *with* Q *ergodic, then* $P \ll Q$ *iff* $P = Q$. *If* P *is ergodic, too, then either* $P = Q$ *or* $P \perp Q$ *in that there is a measurable* $B \subset X$ *such that* $P(B) = 1$ *and* $Q(B) = 0$*).*

Exercise 25 *Prove this lemma.*

Since (4.5) implies that $P_1 \ll P$ and $P_2 \ll P$, the first part of the lemma gives $1 \to 5$. □

The equivalence $1 \leftrightarrow 5$ just proved is interesting even in the trivial case $T = \mathrm{id}_X$, i.e., $T(x) = x$. Since every $B \in \Sigma$ then satisfies $T^{-1}B = B$, a probability measure P is ergodic iff $P(B) = 0$ or 1 for any $B \in \Sigma$. According to criterion 5, this should be equivalent to $P \in \partial_e \mathrm{Prob}(X)$. For example:[55]

Exercise 26 *Let* X *be finite with* $\Sigma = \mathscr{P}(X)$. *Show that* $\partial_e \mathrm{Prob}(X) \cong X$ *via* $\delta_x \leftrightarrow x$ *(i.e., the map* $x \mapsto \delta_x$ *from* X *to* $\mathrm{Prob}(X)$ *is injective, takes values in* $\partial_e \mathrm{Prob}(X)$*, and is surjective onto this image).*

We now prove condition 2 of Proposition 4.3 for the irrational rotation (4.3).[56] Any $f \in L^2(\mathbb{T})$ has a Fourier series, so that, realizing $L^2(\mathbb{T})$ as $L^2([0,2\pi], dx)$, we have

$$f(x) = \sum_{n \in \mathbb{Z}} c_n e^{inx}; \qquad\qquad f(R_\alpha x) = \sum_{n \in \mathbb{Z}} c_n e^{in(x+\alpha)}. \tag{4.8}$$

[54] Recall that for any two measures P, Q on the same σ-algebra $\Sigma \subset \mathscr{P}(X)$ we say that P is *absolutely continuous* w.r.t. Q, written $P \ll Q$, iff $Q(B) = 0$ implies $P(B) = 0$ for $B \in \Sigma$. In that case the *Radon–Nikodym derivative* $\rho = dP/dQ$ exists, with the defining property $\int_X dP\, f = \int_X dQ\, \rho f$ for any Σ-measurable function $f : X \to \mathbb{R}$ (or even $\mathbb{R} \cup \{\infty\}$).

[55] Here δ_x is the measure defined by $\delta_x(B) = 1$ if $x \in B$ and $\delta_x(B) = 0$ if $x \notin B$. This is true more generally for locally compact Hausdorff spaces X with Borel structure, but it cannot be true for general measure spaces: for example, if $\Sigma = \{\emptyset, X\}$ there is just one element $P \in \mathrm{Prob}(X)$, so that $\partial_e \mathrm{Prob}(X) = \mathrm{Prob}(X)$ is also a point, irrespective of X.

[56] We follow Shields, end of §I.2.b and Viana & Oliveira, §4.2.1.

Then $f(R_\alpha x) = f(x)$ a.e. implies that $\exp(in\alpha) = 1$ for all $n \in \mathbb{Z}$ for which $c_n \neq 0$. If $\alpha \notin 2\pi\mathbb{Q}$, then $\exp(in\alpha) = 1$ implies $n = 0$ Hence the Fourier series only has a constant term $f(x) = c_0$ a.e. (note that the equality signs in (4.8) are a.e.), so that $(\mathbb{T}, P, R_\alpha)$ is ergodic. On the other hand, if $\alpha \in 2\pi\mathbb{Q}$, say $\alpha = 2\pi(p/q)$ for $p, q \in \mathbb{Z}_*$, then clearly there are many $n \neq 0$ such that $\exp(in\alpha) = 1$, namely all integral multiples of q/p. In that case f is *not* constant, and hence $(\mathbb{T}, P, R_\alpha)$ is *not* ergodic. □

Exercise 27 *Prove (similarly) that condition 2 holds for M_p iff $|p| \geq 2$.*

For $P = p^{\mathbb{N}}$ or $P = p^{\mathbb{Z}}$ (i.e. the i.i.d. case) the shift map is ergodic. This follows from from a more general result for stationary Markov chains (Theorem 4.7), but it is instructive to also prove it from a different stronger result. Both to this end and for other purposes, we define some stronger properties than ergodicity but which, when they hold, are nonetheless often easier to prove:

Definition 4.5 *A dynamical system (X, P, T) is called:*

- mixing *if for each measurable $C, D \subset X$ we have*

$$\lim_{n\to\infty} P((T^{-n}C) \cap D) = P(C)P(D); \tag{4.9}$$

- weakly mixing *if for each measurable $C, D \subset X$ we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} |P((T^{-n}C) \cap D) - P(C)P(D)| = 0. \tag{4.10}$$

- 'ergodic' *if for each measurable $C, D \subset X$ we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} P((T^{-n}C) \cap D) = P(C)P(D). \tag{4.11}$$

Eq, (4.9) means that for large n the events $T^{-n}C$ and D become independent (relative to P). These definitions (in connection with chaos) are part of the *ergodic hierarchy*. The implication: mixing $\Rightarrow$ weakly mixing follows from Analysis, since convergence implie Césaro convergence.

Exercise 28 *Prove the implications: weakly mixing $\Rightarrow$ 'ergodic' $\Rightarrow$ ergodic.*

In fact, 'ergodicty' is *equivalent* to ergodicity.[57], but the converse implication requires the weak ergodic theorem (see below). The irrational rotation (4.3) is ergodic but not even weakly mixing: taking small intervals for C, D one sees that $\lim_{n\to\infty} P((T^{-n}C) \cap D)$ may not even exist, since it continues to jump between zero and nonzero values that do not converge to zero. Positively:

Proposition 4.6 *For any $p \in \mathrm{Prob}(A)$, the dynamical system $(A^{\mathbb{N}}, p^{\mathbb{N}}, S)$, cf.* (2.5)*, is mixing, and hence ergodic. This also holds for the two-sided case (Bernoulli shift), where $\mathbb{N}$ is replaced by $\mathbb{Z}$.*

Proof. It can be shown that it is enough to verify (4.9) for all generators of the σ-algebra Σ on which P is defined.[58] So let $C = [\sigma]_L$ and $D = [\tau]_M$, for some $\sigma \in A^L$ and $\tau \in A^M$. Since

$$S^{-n}[\sigma]_L = \{s \in A^{\mathbb{N}} \mid s_n = \sigma_0, s_{n+1} = \sigma_1, \ldots, s_{n+L-1} = \sigma_{L-1}\}; \tag{4.12}$$

$$[\tau]_M = \{s \in A^{\mathbb{N}} \mid s_0 = \tau_0, s_1 = \tau_1, \ldots, s_{M-1} = \tau_{M-1}\}, \tag{4.13}$$

[57] Eq. (4.10) implies (4.11), since $|N^{-1}\sum_{n=1}^{N} f(n) - a| = |N^{-1}\sum_{n=1}^{N}(f(n) - a)| \leq N^{-1}\sum_{n=1}^{N} |f(n) - a|$.

[58] See Viano & Oliveira, 2016, Lemma 7.1.2.

for $n > M$ we simply have

$$(S^{-n}C)\cap D = \{s \in A^{\mathbb{N}} \mid s_0 = \tau_0, \ldots, s_{M-1} = \tau_{M-1}, s_n = \sigma_0, \ldots, s_{n+L-1} = \sigma_{L-1}\}. \tag{4.14}$$

Thus $P = p^{\mathbb{N}}$ gives $P((S^{-n}C)\cap D) = P(C)P(D)$ already for $n > M$. Likewise for $\mathbb{N} \rightsquigarrow \mathbb{Z}$. □

This is a special case of the following elegant result.

Theorem 4.7 *Assuming $P \circ S^{-1} = P$, the (one-sided or two-sided) shift map on $(A^{\mathbb{N}}, P, S)$ or $(A^{\mathbb{Z}}, P, S)$ is ergodic iff P comes from an irreducible stationary Markov chain.*

We just sketch the proof of ergodicity from irreducibility.[59] Using (4.14), for $k > 0$ we find

$$\begin{aligned}
P((S^{-M-k}C)\cap D) &= P(S^{-M-k}C|D)P(D) \\
&= P(s_{M+k} = \sigma_0, \ldots, s_{M+k+L-1} = \sigma_{L-1} | s_0 = \tau_0, \ldots, s_{M-1} = \tau_{M-1})P(D) \\
&= P(s_{M+k} = \sigma_0, \ldots, s_{M+k+L-1} = \sigma_{L-1} | s_{M-1} = \tau_{M-1})P(D) \\
&= P(s_{M+k} = \sigma_0, \ldots, s_{M+k+L-1} = \sigma_{L-1})P(s_{M+k} = \sigma_0 | s_{M-1} = \tau_{M-1})P(D) \\
&= P(s_{M+k} = \sigma_0, \ldots, s_{M+k+L-1} = \sigma_{L-1})P^{k+1}_{\tau_{M-1}\sigma_0}P(D) = \\
&= P^{k+1}_{\tau_{M-1}\sigma_0}P_{\sigma_0\sigma_1}\cdots P_{\sigma_{L-2}\sigma_{L-1}}P(D).
\end{aligned} \tag{4.15}$$

Since M is fixed for given D, cf. (4.13), from (3.35) we then obtain

$$\begin{aligned}
&\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}P((S^{-n}C)\cap D) = \lim_{N\to\infty}\frac{1}{N}\sum_{n=M+1}^{N+M}P((S^{-n}C)\cap D) = \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}P((S^{-M-k}C)\cap D) \\
&= \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}P(D)P^{k+1}_{\tau_{M-1}\sigma_0}P_{\sigma_0\sigma_1}\cdots P_{\sigma_{L-2}\sigma_{L-1}} = P(D)p_{\sigma_0}P_{\sigma_0\sigma_1}\cdots P_{\sigma_{L-2}\sigma_{L-1}} = P(D)P(C),
\end{aligned} \tag{4.16}$$

where, since P is assumed irreducible, we used (3.35). This proves (4.11) and hence ergodicity.[60]

Using some additional assumptions, one can prove the existence of invariant measures (and thence of ergodic measures). Theorem 4.9 below relies on some topology on the set $\mathrm{Prob}(A)$ of all probability measures on A. Let A be a compact metrizable space, or more generally a Polish space. The *weak topology* on $\mathrm{Prob}(A)$ is most easily defined by stating what convergence in this topology means; anticipating the fact that (for the kind of spaces A just mentioned) the weak topology will be metrizable, it is enough to define what convergence of sequences means.[61]

Definition 4.8 *For a sequence (p_n) in $\mathrm{Prob}(A)$ we say that $p_n \to p$ weakly in $\mathrm{Prob}(A)$ iff*

$$\textstyle\int_A dp_n\, f \to \int_A dp\, f \text{ for each } f \in C_b(A).$$

It turns out that if A is compact and metrizable, then so is $\mathrm{Prob}(A)$ in this topology, and if A is Polish, then so is $\mathrm{Prob}(A)$.[62] The *Portmanteau theorem* gives various equivalent criteria for $p_n \to p$ weakly, such as: $p_n(B) \to p(B)$ for any measurable $B \subset A$ for which $p(B^- \backslash \mathring{B}) = 0$.

[59] We follow Shields (1996), Example I.2.8; see Viana & Oliveira (2016), Theorem 7.2.8 for the converse.

[60] There is also a criterion for mixing, which we will not use and hence state without proof (see Viana & Oliveira, 2016, Theorem 7.2.11). It (evidently) uses a strengthening of irreducibility, namely *aperiodicity*: this means that there is $N \in \mathbb{N}$ such that $(P^N)_{ab} > 0$ for all $a, b \in A$; whereas irreducibility meant that for all $a, b \in A$ there is $N \in \mathbb{N}$ such that $(P^N)_{ab} > 0$, cf. Definition (3.5). A stationary Markov chain, then, is mixing iff its stochastic matrix is aperiodic.

[61] The official way would be to either use nets or define the weak topology directly, namely by saying that it is generated by the ε-balls $U_{f,x,\varepsilon} := \{p \in \mathrm{Prob}(A) \mid |\int_A dp\, f - x| < \delta\}$, where $f \in C_b(A)$, $x \in \mathbb{R}$ and $\varepsilon > 0$. Dembo & Zeitouni, Appendix D, and Rassoul-Agha, Appendix B.4, are useful summaries of probability theory on Polish spaces.

[62] A metric on $\mathrm{Prob}(A)$ returning the weak topology is the *Lévy–Prohorov metric* d_P. For any $F \subset A$, in terms of a metric d that makes A Polish, and any $\varepsilon > 0$, define $F^\varepsilon := \{a \in A \mid \exists_{b\in F} : d(a,b) < \varepsilon\} = \bigcup_{a\in F}\mathscr{B}_\varepsilon(a)$, where $\mathscr{B}_\varepsilon(a) = \{b \in A \mid d(a,b) < \varepsilon\}$. Then $d_P(p,q) := \inf\{\varepsilon > 0 \mid \forall_{F\subset A, F\text{closed}} : p(F) \le q(F^\varepsilon) + \varepsilon\}$.

For a finite set A all this simplifies considerably. Let $A^* = \{f : A \to \mathbb{R}\} \cong \mathbb{R}^{|A}$ with the usual (Euclidean) topology, so that $\mathrm{Prob}(A) \subset A^*$ inherits this topology, in which it is closed and bounded, and hence compact. This topology is equivalent with the weak topology.

Theorem 4.9 (Krylov–Bogoliubov) *If X is a compact metric space (with Borel structure),*[63] *then $\mathrm{Prob}^T(X)$ is not empty, and neither is its extreme boundary $\partial_e \mathrm{Prob}^T(X)$ of ergodic measures.*

Proof. The second claim follows from the first via Theorem A.25 (Krein–Milman), since if X is compact and metrizable, then $\mathrm{Prob}(X)$ is a compact convex space (in the weak* = weak = vague topology).[64] The first claim follows from the fact that for any $P \in \mathrm{Prob}(X)$ the sequence $(P_N)_N$,

$$P_N = \frac{1}{N} \sum_{n=0}^{N-1} T^{-n} P, \tag{4.17}$$

has an accumulation point (since $\mathrm{Prob}(X)$ is compact), which is necessarily T-invariant.[65] □

It is an interesting question if there is a *unique* ergodic measure:

Exercise 29 *1. Let $X = [0,1]$ with $T(x) = x^2$. What is $\partial_e \mathrm{Prob}^T(X)$?*

2. Same question for $T(x) = 1 - x$.

3. For any X, what is or what are the ergodic measure(s) for $T(x) = x$?

Under the same assumption (i.e. X compact metric) both $\mathrm{Prob}(X)$ and $\mathrm{Prob}^T(X)$ are *(Choquet) simplices*. These generalize the familiar n-simplices Δ_n, see (A.48), which may defined as $\Delta_n = \mathrm{Prob}(n)$, where n is any set with n elements.

Definition 4.10 *A (Choquet) simplex is a convex compact metrizable space K (in a locally convex vector space) in which any element is, in a suitable unique sense, a limit of a sequence consisting of finite convex sums of extremal boundary points (i.e., points in $\partial_e K$). More precisely: each $x \in K$ is the barycenter of a* unique *probability measure μ supported on $\partial_e K$, and this in turn means that*

$$x = \int_{\partial_e K} d\mu(y)\, y, \tag{4.18}$$

or, equivalently, $\mu(f) = f(x)$ for all affine continuous functions on K.

To see that this captures the intuitive idea of a barycenter, take $K = [0,1]$, so that $\partial_e K = \{0,1\}$, and $x \in [0,1]$. Then

$$\mu = (1-x)\delta_0 + x\delta_1; \qquad \mu(f) = (1-x)f(0) + xf(1). \tag{4.19}$$

For any affine function $f(y) = ay + b$ this indeed gives $\mu(f) = f(x)$. Here μ is unique (for given x), as befits a simplex. But the following compact convex set is not a simplex:

$$B^3 := \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 \leq 1\} \tag{4.20}$$

Exercise 30 *Show that $\partial_e B^3 = \partial B^3 = S^2 := \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = 1\}$, and give an example of some $x \in \mathring{B}^3$ with more than one $\mu \in \mathrm{Prob}(S^2)$ whose barycenter is x.*

[63] Equivalently, X is compact and separable. Among the compact spaces this is the non-pathological case.

[64] See e.g. Denker, Grillenberger, & Sigmund (1976), Proposition (2.8), or Walters (1982), §6.2.

[65] See e.g. Denker, Grillenberger, & Sigmund (1976), Theorem (3.6).

A finite-dimensional simplex is affinely isomorphic to some Δ_n, where $\partial_e\Delta_n$ consists of all unit vectors in $\mathbb{R}^n$, but if X is infinite surprising examples may occur. The strangest of these is the *Poulsen simplex*, in which $\partial_e K$ is dense in K. This is the case already in the simple example $K = \mathrm{Prob}^T(\mathbb{T})$, where $\mathbb{T}$ is the unit circle in $\mathbb{C}$ and T is the doubling map $T(z) = z^2$ (though this is not at all simple to prove). Another example will be given in §10.5. Up to affine isomorphism, there is just one Poulsen simplex (i.e., a Choquet simplex with the said property).[66] The theory of simplices (i.e. Choquet theory) gives the cleanest path to the decomposition theory of invariant probability measures into ergodic ones, which comes down to the theorem that (for X compact and metrizable) the set $\mathrm{Prob}^T(X)$ of T-invariant probability measures on X is a Choquet simplex, so that according to the theory just summarized each $P \in \mathrm{Prob}^T(X)$ is the barycenter of a probability measure supported on the set $\partial_e\mathrm{Prob}^T(X)$ of ergodic probability measures on X.

We now discuss the *ergodic theorem*, which comes in two versions (funnily, ergodicity is not assumed in either of these; but if it is assumed the ergodic theorem is considerably sharpened):[67]

Theorem 4.11 *Let (X,P,T) a dynamical system, let $f \in L^1(X,P)$, and consider the sum*

$$f_N(x) := \frac{1}{N}\sum_{n=0}^{N-1} f(T^n x). \tag{4.21}$$

Then $\lim_{N\to\infty} f_N = f^*$ for some T-invariant function $f^* \in L^1(X,P)$, in both of the following senses:

1. *$f_N(x) \to f^*(x)$ pointwise for P-a.e. $x \in X$ (*Birkhoff's pointwise ergodic theorem*);*
2. *$f_N \to f^*$ in L^1; and if $f \in L^2$, $f_N \to f^*$ in L^2 (*von Neumann's mean ergodic theorem*).*[68]

Exercise 31 *Show that L^1-convergence (part 2) implies*

$$\langle f^*\rangle_P \equiv \int_X dP\, f^* = \int_X dP\, f \equiv \langle f\rangle_P. \tag{4.22}$$

For experts in probability theory we also mention an interesting feature of the limit f^*. Let Σ^T consists of the T-invariant sets in the σ-algebra Σ on which P is defined, cf. (2.2). Then

$$f^* = E_P(f \mid \Sigma^T). \tag{4.23}$$

Recall that if $\Sigma_0 \subset \Sigma$ is a sub-σ-algebra of some σ-algebra Σ on X, and $f : X \to \mathbb{R}$ is Σ-measurable, then the *conditional expectation* $E_P(f \mid \Sigma_0)$ is the unique Σ_0-*measurable* function on X that satisfies

$$\int_B dP\, E_P(f \mid \Sigma_0) = \int_B dP\, f, \tag{4.24}$$

for each $B \in \Sigma_0$ (here uniqueness is up to other possibilities that coincide with some given choice P-a.e.). Thus $E_P(f \mid \Sigma_0)$ is a coarse-grained version of f. For example, let $X = 2^{\mathbb{N}}$ with $P = f^{\mathbb{N}}$, and Σ generated by the cylinder sets, as usual. Now take $N \in \mathbb{N}$ and take Σ_N to consist of all sets $[\sigma]^{\mathbb{N}} = \{s \in 2^{\mathbb{N}} \mid s_{|N} = \sigma\}$, where $\sigma \in 2^N$. Then $E_P(f \mid \Sigma_N)(s)$ cannot depend on the 'tail' coordinates s_M for $M \geq N$. Since f is Σ-measurable, for given $s \in 2^{\mathbb{N}}$ the value $f(s)$ can only depend on $s_{|M}$ for some $M < \infty$ (which depends on s), and for the conditional expectation we find

$$\begin{aligned} E_P(f \mid \Sigma_N)(s) &= f(s) && (N \geq M);\\ E_P(f \mid \Sigma_N)(s) &= 2^{N-M}\sum_{\tau\in 2^{M-N}} f(s_{|N}\tau) && (N < M). \end{aligned} \tag{4.25}$$

[66] Simon (2011), Chapters 10 and 11, is an introduction to Choquet theory. Example 9.7; For the Poulsen simplex see Lindenstrauss, Olsen, & Sternfeld (1978); Gelfert & Kwietniak (2018); and Simon (2011), Example 9.7.

[67] von Neumann's ergodic theorem even holds in L^p for all $1 \leq p < \infty$, cf. Walters (1982), Corollary 1.14.1.

[68] For finite measure spaces L^2-convergence implies L^1 convergence, cf. Exercise 23.

As a special case, take $f = 1_C$ for any $C \in \Sigma$ and specialize the conditional expectation to

$$P(C \mid \Sigma_0) := E_P(1_C \mid \Sigma_0), \tag{4.26}$$

so that for each $B \in \Sigma_0$ we have $\int_B dP\, P(C \mid \Sigma_0) = P(B \cap C)$.

Exercise 32 *Show that if also* $\Sigma_0 = \{X, \emptyset, B, X \backslash B\}$, *with* $0 < P(B) < 1$ *then*

$$P(C|B) = P(B \cap C)/P(B) \;\; (x \in B); \qquad P(C \mid \Sigma_0)(x) = P(C|X \backslash B) \;\; (x \in X \backslash B). \tag{4.27}$$

Before proving Theorem 4.11, we note that it combines with ergodicity to give:

Corollary 4.12 *If* (X, P, T) *is ergodic and* $f \in L^1(X, P)$, *then* $f^*(x) = \langle f \rangle_P$ *for P-a.e.* $x \in X$, *i.e.,*

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) = \int_X dP\, f. \tag{4.28}$$

Exercise 33 *Prove this. Hint: show that* $f^*(Tx) = f^*(x)$ *P-a.e. and use Proposition 4.3.3.*

This corollary was Boltzmann's idea: $\boxed{\textit{time average} = \textit{space average}}$, in that the average of f over a "generic" trajectory in time equals its average over X. Of course, asking this for "every" function f requires such trajectories to exhaust X, at least in some (measure-theoretic) sense.[69]

For each $x \in X$ the point measure δ_x on X is defined by $\delta_x(B) = 1$ if $x \in B$ and $\delta_x(B) = 0$ if $x \notin B$; as functionals of $C_b(X)$ we have $\delta_x(f) = f(x)$. Since weak convergence $P_n \to P$ in $\mathrm{Prob}(X)$ is defined by $P_n(f) \to P(f)$ for each $f \in C_b(X)$, eq. (4.28) may be restated as follows:

Corollary 4.13 *If* (X, P, T) *is ergodic, then, weakly in* $\mathrm{Prob}(X)$ *for P-a.e.* $x \in X$,

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \delta_{T^n x} = P. \tag{4.29}$$

Indeed, each term in the sum on the left is a random variable

$$X \to \mathrm{Prob}(X); \qquad\qquad x \mapsto \delta_{T^n x}, \tag{4.30}$$

and since $\delta_{T^n x} \in \mathrm{Prob}(X)$, for each N also the convex combination

$$P_N(x) := N^{-1} \sum_{n=0}^{N-1} \delta_{T^n x} \tag{4.31}$$

is in $\mathrm{Prob}(X)$. Eq. (4.29) then states that $P_N(x) \to P$, weakly in $\mathrm{Prob}(X)$ for P-a.e. $x \in X$. A nice way of looking at (4.29), read the other way round, is to write it as

$$P(B) = \tau(x, B) := \lim_{N \to \infty} \frac{1}{N} |\{n \in \{0, 1, \ldots, N-1\} \mid T^n x \in B\}|, \tag{4.32}$$

so that the probability of $B \in \Sigma$ is equal to the *sojourn time* of a generic path in B (i.e. the fraction of time a particle on such a path spends in B).

[69] Eq. (4.28) was essentially Boltzmann's "ergodic hypothesis", stated in the context of average values of physical observables f in equilibrium states–notably in the microcanonical ensemble, which involves a hypersurface X of constant energy in phase space, with flat prior P.

In our favourite case $X = A^{\mathbb{N}}$, $T = S$, with $P \circ S^{-1} = P$, the evaluation map at $n = 0$, i.e.,

$$X_0 : A^{\mathbb{N}} \to A, \qquad X_0(s) = s_0, \tag{4.33}$$

canonically induces a map X_0^{-1} between probabilities, viz.

$$X_0^{-1} : \mathrm{Prob}(A^{\mathbb{N}}) \to \mathrm{Prob}(A); \qquad X_0^{-1}P(B) = P(\{s \in A^{\mathbb{N}} \mid s_0 \in B\}) = P(X_0 \in B). \tag{4.34}$$

Since $(\delta_{S^n s})_0 = \delta_{s_n}$, the image of $N^{-1}\sum_{n=0}^{N-1} \delta_{S^n s}$ under (4.34) is the *empirical measure* (or *type*)

$$L_N(s) = \frac{1}{N}\sum_{n=0}^{N-1} \delta_{s_n}, \tag{4.35}$$

where $\delta_{s_n} \in \mathrm{Prob}(A)$ is a point measure on A (since $s_n \in A$). For fixed N, this is a random variable

$$L_N : A^{\mathbb{N}} \to \mathrm{Prob}(A); \qquad s \mapsto L_N(s), \tag{4.36}$$

whose value $L_N(s)$ depends only on $s_{|N} \in A^N$. It is easy to show (from the Portmanteau theorem) that X_0^{-1} is weakly continuous,[70] so that, for any S-invariant $P \in \mathrm{Prob}(A^{\mathbb{N}})$, (4.29) implies that

$$\lim_{N\to\infty} L_N(s) = p; \qquad p(a) := P(X_0 = a), \tag{4.37}$$

for P-almost every $s \in A^{\mathbb{N}}$, provided that $(A^{\mathbb{N}}, P, S)$ is ergodic. The simplest situation where we know this to be the case is the i.i.d. probability $P = p^{\mathbb{N}}$ for some prior $p \in \mathrm{Prob}(A)$, cf. Proposition 4.6. Note that the p in (4.37) is now simply the given p in $P = p^{\mathbb{N}}$. As we shall see shortly, eq. (4.37) implies the strong law of large numbers, which therefore follows from the (pointwise) ergodic theorem. But it holds more generally than for i.i.d. Eq. (4.37) and Theorem 4.7 give

Proposition 4.14 *Let $p \in \mathrm{Prob}(A)$ be the unique stationary probability in an irreducible stationary Markov chain with finite state space A, i.e. $p(a) = P(X_0 = a)$, cf. Theorem 3.6. Then*

$$\lim_{N\to\infty} L_N(s) = p, \tag{4.38}$$

for P-almost every sequence $s \in A^{\mathbb{N}}$ (realizing the chain in the Kolmogorov model).

Corollary 4.15 *Given an injective function $E : A \to \mathbb{R}$, define random variables*

$$S_N : A^{\mathbb{N}} \to \mathbb{R}; \qquad S_N(s) := \frac{1}{N}\sum_{n=1}^{N-1} E(s_n). \tag{4.39}$$

If $P \in \mathrm{Prob}(A^{\mathbb{N}})$, where $P \circ S^{-1} = P$, defines an irreducible stationary Markov chain, then

$$P\left(\lim_{N\to\infty} S_N = \langle E\rangle_p\right) = 1 \qquad \Leftrightarrow \qquad S_N \to \langle E\rangle_p \ (P-\text{almost surely}), \tag{4.40}$$

where $p \in \mathrm{Prob}(A)$ is the unique stationary probability of the Markov chain, cf. Theorem 3.6.

In particular, this is true if if the X_n are i.i.d. by p (and hence $P = p^{\mathbb{N}}$).

[70] See footnote 139, which gives $P_n \to P$ weakly in $\mathrm{Prob}(A^{\mathbb{N}})$ iff $P_n(B) \to P(B)$ for all cylinder sets $[\sigma]_N$, $\sigma \in A^N$. Taking $N = 1$ gives weak continuity of X_0^{-1}.

This follows from Proposition 4.14 by taking the average of E on both sides of (4.38). First,

$$\langle E\rangle_{L_N(s)} = \sum_{a\in A}\frac{1}{N}\sum_{n=0}^{N-1}\delta_{s_n}(a)E(a) = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{a\in A}\delta_{s_n a}E(a) = \frac{1}{N}\sum_{n=1}^{N-1}E(s_n) = S_N(s), \tag{4.41}$$

whereas on the right-hand side this averaging trivially gives $\langle E\rangle_p$. This averaging is weakly continuous by definition of the weak topologies on $\mathrm{Prob}(A^{\mathbb{N}})$ and $\mathrm{Prob}(A)$, and hence, with more ado, the restriction to finite sets A so far may be lifted; any "reasonable" measure space works.

The case of repeated sampling from a probability space (A,p) is worth spelling out: for any measurable function $E : A \to \mathbb{R}$ with finite mean $\langle E\rangle_p$, we have

$$\lim_{N\to\infty}\frac{1}{N}(E(s_1)+\cdots+E(s_N)) = \langle E\rangle_p \equiv \int_A dp\,E, \tag{4.42}$$

for $p^{\mathbb{N}}$-almost all $s \in A^{\mathbb{N}}$. For $f = 1_B$ (with $B \subset A$ measurable), eq. (4.42) gives, for $p^{\mathbb{N}}$-a.e. $s \in A^{\mathbb{N}}$,

$$\tau(s,B) := \lim_{N\to\infty}\frac{1}{N}|\{n\in\{0,1,\ldots,N-1\} \mid s_n \in B\}| = p(B). \tag{4.43}$$

Thus one may verify (but not: define!) the probabilities p from long-term frequencies in a long series of observations. These need not even be independent! It is enough if they form an irreducible stationary Markov chain. Looking at the sampling as a path in time, (4.43) is once again a formula for the *sojourn time* $\tau(s,B)$; i.e. the average time spent in B during a journey s, cf. (4.32).[71]

Although weaker than the strong law, the *weak law of large numbers* is often more useful:[72]

Theorem 4.16 *If random variables (X_n) taking values in $A \subset \mathbb{R}$ are i.i.d. by $p \in \mathrm{Prob}(A)$, then*

$$\lim_{N\to\infty} p^N\left(\left|\frac{1}{N}\sum_{n=0}^{N-1}X_n - \langle X_0\rangle_p\right| < \delta\right) = 1, \tag{4.44}$$

for all $\delta > 0$, provided $\langle |X_0|\rangle_p < \infty$. If in addition $\langle X_0^2\rangle_p < \infty$, then one has the explicit bound

$$p^N\left(\left|\frac{1}{N}\sum_{n=0}^{N-1}X_n - \langle X_0\rangle_p\right| \geq \delta\right) \leq \frac{\mathrm{Var}_p(X_0)}{N\delta^2}. \tag{4.45}$$

Here we could have written $p^{\mathbb{N}}$ instead of p^N (and the i.i.d. assumption could also be relaxed). Eq. (4.44) can be proved either from the strong law,[73] or from Theorem 4.11.2 below,[74] or, if it applies (i.e. for finite variance), from (4.45). The latter follows in turn from Chebyshev's inequality

$$p(|X - \langle X\rangle_p| \geq \delta) \leq \mathrm{Var}_p(X)/\delta^2, \tag{4.46}$$

and the fact that i.i.d. gives

$$\mathrm{Var}_p\left(N^{-1}\sum_{n=0}^{N-1}X_n\right) = N^{-1}\mathrm{Var}(X_0). \tag{4.47}$$

[71] Eq. (4.43) also gives convergence of the *Monte Carlo method* for computing volumes $\mu(B)$ of subsets $B \subset A = [0,1]^d$ (and hence also integrals). In its simplest form: use a random generator to provide $N\cdot d$ elements of $[0,1]$ and hence N elements (d-tuples) of $[0,1]^d$; count the number of points in B, divide this number by N, and let $N\to\infty$. The result is (almost surely) $\mu(B)$.

[72] Here $\mathrm{Var}_p(X) := \langle X^2\rangle_p - \langle X\rangle_p^2$ is the variance of a random variable X with law p, as usual.

[73] See e.g. Klenke (2014), Remark 5.13, based on Fatou's lemma.

[74] If (4.44) fails there is $\delta > 0$ such that for each M there is $N > M$ for which $P(|f_N - f^*|) \geq \delta) \geq \delta$. This blocks convergence $f_N \to f^*$ in L^1 (or in any L^p, $1 \leq p < \infty$). Given Theorem 4.11.2, this is a proof by contradiction of (4.44).

Proof of Theorem 4.11.1. All proofs are long and technical, but the following version at least has a clear structure.[75] One might have the special case of the Strong Law of Large Numbers (SLLN) for a coin toss in mind, cf. Theorem 4.7 and Corollary 4.15. We then have

$$X = 2^{\mathbb{N}}; \quad P = p^{\mathbb{N}}; \quad x = x_0 x_1 \cdots \ (x_n \in \{0,1\}); \quad Tx_n = x_{n+1}; \quad f(x) = x_0; \tag{4.48}$$

$$f(T^n x) = x_n; \quad \frac{1}{N}\sum_{n=0}^{N-1} f(T^n x) = \frac{1}{N}\sum_{n=0}^{N-1} x_n; \quad \int_X dP\, f = \langle x_0 \rangle_P = p = p(1). \tag{4.49}$$

Let $f \in L^1(X,P)$. Without loss of generality we may assume that $f \geq 0$ (i.e. $f(x) \geq 0$ for all x), since the statement of the theorem (involving pointwise limits and integrals) is linear in f; otherwise we write $f = f_+ - f_-$ with $f_\pm \geq 0$ and apply the proof to $f_\pm$ separately. For $x \in X$,

$$f_N(x) := \frac{1}{N}\sum_{n=0}^{N-1} f(T^n x); \qquad \underline{f(x)} := \liminf_{N\to\infty} f_N(x); \qquad \overline{f(x)} := \limsup_{N\to\infty} f_N(x) \tag{4.50}$$

always exist (the lim sup may be ∞, but since $f \in L^1$ we may exclude this value by taking out a measure-zero set from X if necessary; for a coin toss this is unnecessary since $f_N(x) \in [0,1]$). Our aim is to prove that for P-almost every $x \in X$ we have

$$\underline{f(x)} = \overline{f(x)}, \tag{4.51}$$

since in that case the limit itself exists for P-almost every $x \in X$, i.e.

$$f^*(x) := \lim_{N\to\infty} f_N(x) = \underline{f(x)} = \overline{f(x)}. \tag{4.52}$$

Eq. (4.51) will follow from the inequalities (whose derivation, then, is the burden of the proof)

$$\int_X dP\, \underline{f} \geq \int_X dP\, f \geq \int_X dP\, \overline{f}. \tag{4.53}$$

For these imply $\int_X dP\,(\overline{f} - \underline{f}) \leq 0$, which together with the pointwise inequalities $\overline{f}(x) - \underline{f}(x) \geq 0$, which are a trivial consequence of the definitions of lim sup and lim inf, yields (4.51) P-a.e.. Moreover, knowing (4.51) and using (4.53) once again, gives (4.22). Finally, T-invariance of f^* is, once it exists, also obvious from (4.50) and (4.52). Therefore, all we still need is (4.53).

We start with the second inequality in (4.53). As already mentioned, by removing a measure zero set we may assume that $\overline{f(x)} < \infty$. For any $\varepsilon > 0$ and $M \in \mathbb{N}$, define the set

$$X_M := \{x \in X \mid f_M(x) \geq \overline{f(x)}(1-\varepsilon)\}. \tag{4.54}$$

Then $X_M \subset X_{M+1}$. By definition of lim sup, for any x for which $\overline{f(x)} < \infty$ there is $M > 0$ such that

$$f_M(x) \geq \overline{f(x)}(1-\varepsilon), \tag{4.55}$$

so that $\cup_{M\in\mathbb{N}} X_M = X$. Hence for any $0 < \delta < 1$ we can find M such that

$$P(X_M) \geq 1 - \delta. \tag{4.56}$$

[75] We follow (and simplify) a proof along the lines of a proof of Hopf's ergodic theorem by Kamae & Keane (1997) as presented in Dajani & Kalle (2021), §3.1. Hopf's *ratio ergodic theorem* states that if $f, g \in L^1(X,P)$ such that $g(x) \geq 0$ and $\sum_{n=0}^{\infty} g(T^n x) = \infty$, then the ratio $r(x) := \lim_{N\to\infty} \frac{\sum_{n=0}^{N-1} f(T^n x)}{\sum_{n=0}^{N-1} g(T^n x)}$ exists P-a.e., is T-invariant, and satisfies $\int_X dP\, f = \int_X dP\, rg$. For $g = 1_X$ this recovers Birkhoff's ergodic theorem. See also Keane (2005).

To clarify the idea of this part of the proof, we assume that there is an M such that $P(X_M) = 1$, and in addition that if $x \in X_M$, then also $T^k x \in X_M$ for all k. In that case, for any k, since $T^k \circ T^n = T^{k+n}$,

$$\sum_{n=k}^{k+M-1} f(T^n x) = \sum_{n=0}^{M-1} f(T^{k+n}x) = M f_M(T^k x) \geq M\overline{f(x)}(1-\varepsilon). \tag{4.57}$$

Lemma 4.17 *For any $M \in \mathbb{N}$: If (a_n) are non-negative real numbers such that for any $k \geq 0$,*

$$\sum_{n=k}^{k+M-1} a_n \geq c \cdot M, \tag{4.58}$$

for some constant $c \geq 0$, then for all $N > M$ we have

$$\sum_{n=0}^{N-1} a_n \geq c \cdot (N-M). \tag{4.59}$$

Exercise 34 *Prove this lemma.*

Consequently, eqs. (4.50) and (4.57) give

$$f_N(x) = \frac{1}{N}\sum_{n=0}^{N-1} f(T^n x) \geq \frac{N-M}{N}\overline{f(x)}(1-\varepsilon) = \overline{f(x)}(1-\varepsilon) + O(1/N). \tag{4.60}$$

Since the measure P is T-invariant, i.e. $\int_X dP(x)\, f(T^n x) = \int_X dP(x)\, f(x)$, eq. (4.60) gives

$$\int_X dP\, f \geq (1-\varepsilon)\int_X dP\, \overline{f} + O(1/N). \tag{4.61}$$

Letting $N \to 0$ and then $\varepsilon \to 0$ gives the second inequality in (4.53). This is the main idea.

We now refine the argument so that the two simplifying assumptions are unnecessary; but we first present it for the case that f is bounded.[76] In that case we pick a number

$$L > \|f\|_\infty := \sup_{x \in X} |f(x)|. \tag{4.62}$$

and replace f by F, defined by $F(x) = f(x)$ if $x \in X_M$ and $F(x) = L$ if $x \notin X_M$. For given n, in case that $T^n x \in X_M$ we have

$$\sum_{n=k}^{k+M-1} F(T^n x) = \sum_{n=k}^{k+M-1} f(T^n x), \tag{4.63}$$

so we have the bound (4.57). If $T^n x \notin X_M$ some of the terms in the sum equal L, but since $L > \overline{f(x)}$ we also have this bound. Hence

$$\frac{1}{N}\sum_{n=0}^{N-1} F(T^n x) \geq (1-\varepsilon)\cdot\overline{f(x)} + O(1/N); \tag{4.64}$$

$$\Rightarrow \int_X dP\, F \geq (1-\varepsilon)\int_X dP(x)\,\overline{f(x)} + O(1/N), \tag{4.65}$$

[76] This proof also works if f is *essentially* bounded, since the removal of a set of measure zero reduces the proof to the bounded case. The P-almost everywhere nature of the claim in Theorem 4.11 is not jeopardized by this removal.

by Lemma 4.17. Using (4.56), which implies $P(X\backslash X_M) < \delta$, we also have

$$\int_X dPF = \int_{X_M} dPF + \int_{X\backslash X_M} dPF = \int_{X_M} dPf + LP(X\backslash X_M) \leq \int_{X_M} dPf + L\delta. \tag{4.66}$$

Combining (4.65) and (4.66) with the fact that $F = f$ on X_M gives

$$\int_X dPf \geq \int_{X_M} dPf \geq \int_X dPF - \delta L \geq (1-\varepsilon)\int_X dP(x)\,\overline{f(x)} - L\delta + O(1/N). \tag{4.67}$$

We now let $N \to \infty$, $\delta \to 0$, $\varepsilon \to 0$ in this order, and again obtain the second inequality in (4.53).

Finally, in the general case where f is not bounded (or even essentially bounded), at first we take $L > 0$ to be arbitrary (but think of it as large, eventually letting $L \to \infty$) and replace (4.54) by

$$X'_M := \{x \in X \mid f_M(x) \geq \min\{\overline{f(x)}, L\} \cdot (1-\varepsilon)\}. \tag{4.68}$$

We still have $X'_M \subset X'_{M+1}$ and $\cup_M X'_M = X$, so that for any $0 < \delta < 1$ there is M such that (4.56) holds. Using the same F as before, and using the same arguments we now obtain the inequalities

$$\int_X dPF \geq (1-\varepsilon)\int_X dP(x)\,\min\{\overline{f(x)}, L\} + O(1/N); \;\Rightarrow \tag{4.69}$$

$$\int_X dPf \geq (1-\varepsilon)\int_X dP(x)\,\min\{\overline{f(x)}, L\} - L\delta + O(1/N). \tag{4.70}$$

We then let $N \to \infty$, $\delta \to 0$, $\varepsilon \to 0$, and finally $L \to \infty$, which replaces $\min\{\overline{f(x)}, L\}$ by $\overline{f(x)}$.

The proof of the first inequality in (4.53) is similar. It relies on the following:

Lemma 4.18 *Let $f \in L^1(X,P)$. For each $\varepsilon > 0$ there is $\delta > 0$ such that for all measurable $B \subset X$:*

$$P(B) < \delta \qquad \Rightarrow \qquad \int_B dP|f| < \varepsilon. \tag{4.71}$$

Proof (by contradiction). The negation of the claim states that there exists $\varepsilon > 0$ such that for each $\delta > 0$ there is a B with $P(B) < \delta$ and $\int_B dP|f| \geq \varepsilon$. Hence there is a sequence (B_N) such that $P(B_N) \to 0$ whilst $\int_{B_N} dP|f| \geq \varepsilon$ for all N. Consider the sequence $(1_{B_N}|f|)$ in $L^1(X,P)$. This is obviously dominated by $|f| \in L^1(X,P)$, so that by dominated convergence,

$$\lim_{N\to\infty} \int_X dP\, 1_{B_N}|f| = \int_X dP \lim_{N\to\infty} 1_{B_N}|f| = 0, \tag{4.72}$$

since $\lim_{N\to\infty} 1_{B_N} = 0$ P-a.e. But $\int_X dP\, 1_{B_N}|f| = \int_{B_N} dP|f| \geq \varepsilon$, so we have a contradiction. □

We still assume $f \geq 0$. The role of X_M in the previous part of the proof is now played by

$$Y_M := \{x \in X \mid \exists_{m\leq M}(f_m(x) \leq \underline{f(x)}(1+\varepsilon'))\}; \tag{4.73}$$

with $f_M(x)$ instead of $\exists_{m\leq M} f_m(x)$ we would not have $Y_M \subset Y_{M+1}$, which is the case for (4.73). Therefore, if we pick $\varepsilon > 0$ and $\varepsilon' > 0$, and for the former find $\delta > 0$ from Lemma 4.18, we can find M is such that $P(X\backslash Y_M) < \delta$, cf. (4.56). Since f is bounded below by 0, we define $G(x) = f(x)$ if $x \in Y_M$ and $G(x) = 0$ if $x \notin Y_M$, so that $G \leq f$. For any k we therefore obtain from (4.73) that

$$\sum_{n=k}^{k+M-1} G(T^n x) \leq M \cdot \underline{f(x)}(1+\varepsilon'), \tag{4.74}$$

and hence by an obvious adaptation of Lemma 4.17 we find that for all $N > M$ we have

$$\frac{1}{N}\sum_{n=0}^{N-1} G(T^n x) \le \underline{f(x)}(1+\varepsilon') + O(1/N); \;\Rightarrow\; \int_X dP\, G \le (1+\varepsilon')\int_X dP(x)\,\underline{f(x)} + O(1/N);$$

$$\begin{aligned}\Rightarrow\; \int_X dP\, f &= \int_{Y_M} dP\, f + \int_{X\setminus Y_M} dP\, f = \int_{Y_M} dP\, G + \int_{X\setminus Y_M} dP\, f \\ &\le \int_X dP\, G + \int_{X\setminus Y_M} dP\, f \le (1+\varepsilon')\int_X dP(x)\,\underline{f(x)} + \int_{X\setminus Y_M} dP\, f \; + O(1/N) \\ &\le (1+\varepsilon')\int_X dP(x)\,\underline{f(x)} + \varepsilon + O(1/N), \end{aligned} \tag{4.75}$$

where we used Lemma 4.17. Letting $N \to \infty$, then $\varepsilon \to 0$, and then $\varepsilon' \to 0$ turns (4.75) into the first inequality in (4.53), written in the opposite way. □

Proof of Theorem 4.11.2. We give two proofs. The first,[77] ahistorical proof derives part 2 from part 1 and works in any p-norm, $1 \le p < \infty$. If f is (essentially) bounded, then so is each f_N, and since $|f_N(x) - f^*(x)|^p \to 0$ pointwise P-a.e. by part 1, the bounded convergence theorem gives $f \to f^*$ in $L^p(X,P)$. If $f \in L^p$ is not bounded (think of $p = 1$, $X = [0,1]$ with Lebesgue measure, and $f(x) = 1/\sqrt{x}$), for any $\varepsilon > 0$ one can nonetheless find $g \in L^\infty$ (so that also $g \in L^p$, since P is finite) such that $\|f - g\|_p < \varepsilon/4$, and find sufficiently large N such that $\|g_{M+N} - g_N\|_p < \varepsilon/2$ for all M. Then one shows that (f_N) is a Cauchy sequence in L^p by the usual ε-elegance:

$$\|f_{M+N} - f_N\|_p \le \|f_{M+N} - g_{M+N}\|_p + \|g_{M+N} - g_N\|_p + \|g_N - f_N\|_p < \varepsilon/4 + \varepsilon/2 + \varepsilon/4 = \varepsilon. \tag{4.76}$$

Second, von Neumann's mean ergodic theorem (part 2) historically preceded Birkhoff's (part 1), and indeed L^2-convergence is considerably easier (and more elegant) to prove than pointwise convergence. The main technique of the proof is the use of the so-called *Koopman operator*

$$U_T : L^2(X,P) \to L^2(X,P); \qquad\qquad U_T f := f \circ T, \tag{4.77}$$

which by T-invariance of P is an isometry,[78] in that it satisfies

$$\langle U_T f, U_T g\rangle = \langle f, g\rangle, \tag{4.78}$$

where $\langle f,g\rangle := \int_X dP(x)\overline{f(x)}g(x)$ is the inner product in $L^2(X,P)$. A key role will be played by

$$H_T := \ker(U_T - 1_H) = \{f \in L^2 \mid U_T f = f\} \subset L^2(X,P) \equiv H, \tag{4.79}$$

where 1_H is the unit operator on H. Being the kernel of a continuous (i.e. bounded) linear operator, H_T is a closed linear subspace of H. We now compute its orthogonal complement $H_T^\perp$, consisting of all $g \in L^2$ such that $\langle g, f\rangle = 0$ for all $f \in H_T$. The following observation is crucial to this end:

Exercise 35 *Show that $U_T f = f$ iff $U_T^* f = f$.*

For any bounded operator $A : H \to H$ we have $\ker(A^*)^\perp = \mathrm{ran}(A)^-$, where the bar denotes closure. For $A = U_T - 1_H$, by the exercise we have $\ker(A^*) = \ker(A)$ and hence $\ker(A)^\perp = \mathrm{ran}(A)^-$, i.e.,

$$H_T^\perp = \mathrm{ran}(U_T - 1_H)^- = \{U_T g - g, g \in L^2\}^-. \tag{4.80}$$

Trivially,

$$f = p_T f + (1_H - p_T), \tag{4.81}$$

[77] See Walters (1982), Corollary 1.14.1, for details.

[78] this means that $U_T^* U_T = 1_H$. But U_T is unitary only if T is invertible.

where $p_T : H \to H_T$ is the orthogonal projection defined by

$$p_T f = f \ \ (f \in H_T); \qquad\qquad p_T f = 0 \ \ (f \in H_T^\perp), \tag{4.82}$$

Since $(1_H - p_T)f \in H_T^\perp$, if we ignore the closure in (4.80) for the moment there is $g \in H$ such that

$$f = p_T f + U_T g - g. \tag{4.83}$$

By definition of f_N, see (4.21), we have

$$(p_T f)_N = p_T f; \qquad\qquad (U_T g - g)_N = \frac{1}{N}(U_T^N - 1_H)g. \tag{4.84}$$

Hence

$$\|f_N - p_T f\|_2 = \frac{1}{N}\|U_T^N g - g\|_2 \leq \frac{1}{N}\|U_T^N - 1_H\|\|g\|_2 \leq \frac{1}{N}(\|U_T^N\| + \|1_H\|)\|g\|_2 \leq \frac{2}{N}\|g\|_2, \tag{4.85}$$

by the triangle inequality for the operator norm $\|\cdot\|$, and the fact that $\|U\| = 1$ for any (nonzero) isometry U. It follows that $\|f_N - p_T f\|_2 \to 0$, so that $f_N \to p_T f$ in L^2. □

The proof identifies $f^* = \lim_N f_N$ *as* $p_T f$, at least in L^2 (where orthogonal projections are defined).

Exercise 36 *Finish the proof by taking the closure in* (4.80) *into account; this affects* (4.83).

Finally: *Proof of* (4.24). Since f^* is T-invariant and hence Σ^T-measurable, $E_P(f, \Sigma^T)$ is a candidate for f^*. We therefore need to check (4.24), i.e., for each $B \in \Sigma^T$ we should have

$$\int_B dP\, f^* = \int_B dP\, f. \tag{4.86}$$

This follows from the pointwise ergodic theorem applied to $f \cdot 1_B$. By T-invariance of B we have

$$(f \cdot 1_B)^* = f^* \cdot 1_B, \tag{4.87}$$

so that (4.22) gives (4.86) and hence, by (P-a.e.) uniqueness, (4.23). □

5 Asymptotic properties of entropy

After this extensive preparation we finally turn to our main topic! There will be lots of estimates!

- We already defined the *type* or *empirical* measure L_N, cf. (4.35). For fixed $\sigma \in A^N$ we have

$$L_N(\sigma) := \frac{1}{N} \sum_{n=0}^{N-1} \delta_{\sigma(n)}; \qquad\qquad L_N(\sigma) \in \mathrm{Prob}(A). \tag{5.1}$$

  But note that $L_N(\sigma) \in \mathrm{Prob}_N(A) \subset \mathrm{Prob}(A)$, where $\mathrm{Prob}_N(A)$ consists of all "quantized" probability distributions $p : A \to [0,1]$, namely those for which $p(a) \equiv p_a = m_a/N$ for some $m_a \in \{0,1,\ldots,N\}$, with $\sum_{a\in A} m_a = N$. Hence L_N is *a priori* a function $L_N : A^N \to \mathrm{Prob}(A)$, but it may (and will) often be regarded as a function $L_N : A^{\mathbb{N}} \to \mathrm{Prob}(A)$ given by (5.1) in which $\sigma \in A^N$ is simply replaced by $s \in A^{\mathbb{N}}$; in that case $L_N(s)$ clearly depends on $s_{|N}$ only.

- In the opposite direction, the *type class* $T_N(p) \subset A^N$ of $p \in \mathrm{Prob}(A)$ is

$$T_N(p) := \{\sigma \in A^N \mid L_N(\sigma) = p\}. \tag{5.2}$$

  Thus T_N is a function $T_N : \mathrm{Prob}(A) \to \mathscr{P}(A^N)$, with $T_N(p) = \emptyset$ if $p \notin \mathrm{Prob}_N(A)$.

In kinetic gas theory (cf. the Introduction), Boltzmann (1879) looked at N particles each of which could be in some state $a \in A$, so that $\sigma \in T_N(p)$ is a specific way of putting $n_a = Np_a$ particles in state $a \in A$, for each $a \in A$. He famously computed the cardinality of $T_N(p)$: if $p \in \mathrm{Prob}_N(A)$, then

$$|T_N(p)| = \frac{N!}{\prod_{a\in A}(Np_a)!} = \frac{N!}{\prod_{a\in A} n_a!}. \tag{5.3}$$

Here is a group-theoretical way to see this. First, $T_N(p)$ is not empty because $p \in \mathrm{Prob}_N(A)$. Pick som $\sigma \in T_N(p)$. Then any $\sigma' \in T_N(p)$ can be obtained from σ by a permutation π of the N particles, that is, $\sigma' = \sigma \circ \pi$. Define an equivalence relation $\sim$ on the permutation group $\mathfrak{S}_N$ on N symbols by $\pi \sim \pi'$ iff $\sigma \circ \pi = \sigma \circ \pi'$. Then

$$|T_N(p)| = |\mathfrak{S}_N/\sim| = |\mathfrak{S}_N/S_\sigma| = |\mathfrak{S}_N|/|S_\sigma|, \tag{5.4}$$

where $S_\sigma = \{\pi \in \mathfrak{S}_N \mid \sigma \circ \pi = \sigma\}$ is the stabilizer of σ. This gives (5.3), since

$$|\mathfrak{S}_N| = N!; \qquad\qquad |S_\sigma| = \prod_{a\in A} n_a!. \tag{5.5}$$

Exercise 37 *Provide your own purely combinatorial derivation of* (5.3).

This formula leads to an archetypical limit formula for (base-e) Shannon's entropy (1.4):

$$S(p) := -\sum_{a\in A} p(a) \log p(a); \tag{5.6}$$

$$S(p) = \lim_{N\to\infty} \frac{1}{N} \log |T_N(p_N)|, \tag{5.7}$$

for any $p \in \mathrm{Prob}(A)$ and any sequence (p_N) such that $p_N \in \mathrm{Prob}_N(A)$ and $p_N \to p$ (weakly).

Exercise 38 *Prove* (5.7) *from Stirling's formula:*[79]

$$\log(n!) = n \log(n/e) + O(\log n). \tag{5.8}$$

[79] The usual form is $n! \approx n^n e^{-n} \sqrt{2\pi n}$.

We also have a more precise estimate for finite N, which at once implies (5.7), and much more:

$$(N+1)^{-|A|} \cdot e^{NS(p)} \leq |T_N(p)| \leq e^{NS(p)}, \tag{5.9}$$

provided of course that $p \in \mathrm{Prob}_N(A)$. Let us note that one may also write (5.9), as

$$e^{NS(p)-o(N)} \leq |T_N(p)| \leq e^{NS(p)}, \tag{5.10}$$

where the $o(N)$ term is a positive function $f(N)$, depending on $|A|$ but not on p, such that $\lim_{N\to\infty} f(N)/N = 0$. Indeed, from (5.9) we have $f(N) = |A|\log(N+1)$. To get some feeling for this, note that A^N, of which $T_N(p)$ is a subset, has $|A|^N$ elements, so that

$$(N+1)^{-|A|} \cdot e^{N(S(p)-\log|A|)} \leq \frac{|T_N(p)|}{|A|^N} \leq e^{N(S(p)-\log|A|)}. \tag{5.11}$$

The fraction in the middle is $f^N(\sigma \in T_N(p))$, where f is the flat prior on A, i.e., $f(a) = |A|^{-1}$. In view of (1.28) we also recognize the relative entropy (1.27), which now takes the form

$$D(p\|q) := \sum_{a\in A} p(a)\log\left(\frac{p(a)}{q(a)}\right) \quad \text{if } p \ll q; \tag{5.12}$$

$$D(p\|q) := \infty \quad \text{otherwise}, \tag{5.13}$$

where, at least in the case that A is finite considered here, the 'absolute continuity' $p \ll q$ means that $q(a) = 0$ implies $p(a) = 0$, in which case the corresponding term in (5.12) is zero. This condition holds for $q = f$, in which case we have, cf. (1.28),

$$D(p\|f) = -S(p) + \log|A|, \tag{5.14}$$

so that (5.11) may be rewritten as

$$(N+1)^{-|A|} \cdot e^{-ND(p\|f)} \leq f^N(L_N = p) \leq e^{-ND(p\|f)}, \tag{5.15}$$

where $f^N(L_N = p) = f^N(\sigma \in T_N(p))$ is short for $f^N(\{\sigma \in A^N \mid L_N(\sigma) = p\})$.

Proposition 5.1 (Gibbs inequality) *We have $D(p\|q) \geq 0$ with equality $D(p\|q) = 0$ iff $p = q$.*

Exercise 39 *Prove this from Jensen's inequality* (A.4) *with $f(x) = -\log x$, $F(a) = q(a)/p(a)$, and $P = p$, assuming $p \ll q$,*

Hence eq. (5.15) shows that the probability f^N that $L_N(\sigma) = p \neq f$ is exponentially small as $N \to \infty$, whereas $f^N(\sigma \in T_N(f)) \to 1$ as $N \to \infty$. This is our first large deviations result!

But we need not use f, and can generalize (5.15). If $\sigma \in T_N(p)$, then, as a simple exercise,

$$q^N(\sigma) = e^{-N(S(p)+D(p\|q))}, \tag{5.16}$$

for any $p, q \in \mathrm{Prob}(A)$ such that $p \ll q$, where we still assume that $p \in \mathrm{Prob}_N(A)$. Therefore,

$$q^N(L_N = p) = \sum_{\sigma\in T_N(p)} q^N(\sigma) = |T_N(p)|e^{-N(S(p)+D(p\|q))}. \tag{5.17}$$

Using (5.9) in this equality we see that $S(p)$ drops out, so that (5.15) generalizes to

$$(N+1)^{-|A|} \cdot e^{-ND(p\|q)} \leq q^N(L_N = p) \leq e^{-ND(p\|q)}. \tag{5.18}$$

Exercise 40 *Prove eq.* (5.16).

The estimate (5.9) - (5.10) also allows us to compute the asymptotics of the Boltzmann–Planck entropy (1.3). A sensible probabilistic interpretation of this formula (putting $k = 1$) seems to be

$$S_B^{(N)}(p|q) = \log q^N(L_N = p), \tag{5.19}$$

where $p \in \mathrm{Prob}_N(A)$ is a given macrostate and $q \in \mathrm{Prob}(A)$ is a prior (which Boltzmann took to be flat, see below). Eq, (5.18) then immediately gives

$$s_B(p|q) := \lim_{N\to\infty} \frac{S_B^{(N)}(p_N|q)}{N} = -D(p\|q), \tag{5.20}$$

where, as in (5.7), $p \in \mathrm{Prob}(A)$ is arbitrary and (p_N) is any sequence such that $p_N \in \mathrm{Prob}_N(A)$ and $p_N \to p$ (weakly). For the flat prior (1.9), eq. (5.14) yields

$$s_B(p|f) = S(p) - \log|A|. \tag{5.21}$$

Boltzmann (1877) himself interpreted the "W" in (1.3) as $W = |T_N(p)|$. We therefore write

$$\tilde{S}_B^{(N)}(p) := \log|T_N(p)|; \qquad\qquad \tilde{s}_B(p) := \lim_{N\to\infty} \frac{\tilde{S}_B^{(N)}(p_N)}{N}, \tag{5.22}$$

and, straight from (5.7), see that Boltzmann's entropy (asymptotically) coincides with Shannon's:

$$\tilde{s}_B(p) = S(p). \tag{5.23}$$

Proof of (5.9). Upper bound: we have $D(p\|p) = 0$, so that (5.16) becomes, still for $\sigma \in T_N(p)$,

$$p^N(\sigma) = e^{-NS(p)}. \tag{5.24}$$

We use (5.24) in the following computation:

$$p^N(T_N(p)) = \sum_{\sigma\in T_N(p)} p^N(\sigma) = \sum_{\sigma\in T_N(p)} e^{-NS(p)} = |T_N(p)|e^{-NS(p)}, \tag{5.25}$$

and combine this with the property $p^N(\Delta) \le 1$ for any $\Delta \subset A^N$. Hence $|T_N(p)|e^{-NS(p)} \le 1$, which gives the upper bound in (5.9). The lower bound uses the estimate, for any $p, q \in \mathrm{Prob}_N(A)$,

$$p^N(T_N(q)) \le p^N(T_N(p)). \tag{5.26}$$

This may be unsurprising, but it is actually hard to prove (see below). We also use the fact that

$$|\mathrm{Prob}_N(A)| \le (N+1)^{|A|}, \tag{5.27}$$

which follows because in the probabilities $p(a_1) = k_1/N, \ldots, p(a_{|A|}) = k_|A|/N$ comprising $\mathrm{Prob}_N(A)$ each k_i can only take $N+1$ possible values (i.e. $k_i \in \{0, 1, \ldots, N\}$). Also, trivially,

$$A^N = \bigcup_{p\in\mathrm{Prob}_N(A)} T_N(p). \tag{5.28}$$

Using (5.26), (5.27), (5.28), and (5.25), we obtain the lower bound in (5.9):

$$\begin{aligned} 1 = p^N(A^N) = p^N\left(\bigcup_{q\in\mathrm{Prob}_N(A)} T_N(q)\right) &= \sum_{q\in\mathrm{Prob}_N(A)} p^N(T_N(q)) \le \sum_{q\in\mathrm{Prob}_N(A)} p^N(T_N(p)) \\ = |\mathrm{Prob}_N(A)|p^N(T_N(p)) \le (N+1)^{|A|}p^N(T_N(p)) &= (N+1)^{|A|}|T_N(p)|e^{-NS(p)}. \end{aligned} \tag{5.29}$$

To prove (5.26), we compute

$$p^N(T_N(q)) = \sum_{\sigma\in T_N(q)} p^N(\sigma) = \sum_{\sigma\in T_N(q)} \prod_{a\in A} p(a)^{Nq(a)} = |T_N(q)| \prod_{a\in A} p(a)^{Nq(a)} \tag{5.30}$$

$$= \frac{N!}{(Nq_{a_1})!\cdots(Nq_{a_{|A|}})!} \prod_{a\in A} p(a)^{Nq(a)}, \tag{5.31}$$

where we used (5.16) with p and q swapped, and (5.3) with $p \rightsquigarrow q$. Therefore,

$$\frac{p^N(T_N(p))}{p^N(T_N(q))} = \prod_{a\in A} \left(\frac{(Nq_a)!}{(Np_a)!} \cdot p_a^{N(p_a-q_a)} \right). \tag{5.32}$$

The inequality $(m!/n!) \geq n^{m-n}$ then yields (5.26). □

Exercise 41 *Prove this inequality and supply the ensuing final step towards* (5.26)*.*

The use of probabilities $p \in \mathrm{Prob}_N(A)$ was inconvenient in stating (5.7) and (5.20). This may resolved by working with a buffer around $\mathrm{Prob}_N(A)$: motivated by (5.24), for any $\delta > 0$ we define

$$T^E_{N,\delta}(p) := \{\sigma \in A^N \mid e^{-N(S(p)+\delta)} \leq p^N(\sigma) \leq e^{-N(S(p)-\delta)}\}, \tag{5.33}$$

so that at least we have $T_N(p) \subset T^E_{N,\delta}(p)$. The key facts about $T^E_{N,\delta}(p)$ are:

$$\lim_{N\to\infty} p^N(T^E_{N,\delta}(p)) = 1; \tag{5.34}$$

$$(1-\delta)e^{N(S(p)-\delta)} \leq |T^E_{N,\delta}(p)| \leq e^{N(S(p)+\delta)}, \tag{5.35}$$

wherethe upper bound is true for any N, whereas the lower bound holds for sufficiently large N. Eq. (5.34) is a "weak" form of the *Asymptotic Equipartition Property* (AEP), of which (5.41) below will be a" strong" form. Eq. (5.35) follows from (5.34), which we will prove shortly. Consequently, for given $p \in \mathrm{Prob}(A)$ and small $\delta > 0$, for large N the set A^N splits into two parts:

- $T^E_{N,\delta}(p)$, which by (5.35) has about $e^{NS(p)}$ members, i.e. a fraction $e^{-ND(p\|f)}$ of the total, cf. (5.14), which by (5.33) all have approximately the same probability $e^{-NS(p)}$, cf. (5.33).
- Its complement, which even as a whole has negligible probability because of (5.34).

The AEP is spectacular if $p \neq f$: in that case $D(p\|f) > 0$ and hence the strings in the first class form an exponentially small fraction which nonetheless carry almost all of the probability!

Eq. (5.34) follows from Theorem 4.16 by taking $X_n(s) = \log p(s_n)$ (which are i.i.d.). Then

$$\langle X_n \rangle_{p^{\mathbb{N}}} = \sum_\lambda p^N(X_n = \lambda)\cdot\lambda = \sum_{a\in A} p(a)\log p(a) = -S(p); \tag{5.36}$$

$$\left\{ s \in A^N \mid \left| \frac{1}{N}\sum_{n=0}^{N-1} X_n(s) - \langle X_0\rangle_p \right| < \delta \right\} = \left\{ s \in A^N \mid \left| \frac{1}{N}\sum_{n=0}^{N-1} \log p(s_n) + S(p) \right| < \delta \right\}$$

$$= \left\{ s \in A^N \mid \left| \frac{1}{N}\log\prod_{n=0}^{N-1} p(s_n) + S(p) \right| < \delta \right\} = \left\{ s \in A^N \mid \left| \frac{1}{N}\log(p^N(s_{|N})) + S(p) \right| < \delta \right\}$$

$$= \{\sigma \in A^N \mid \sigma \in T^E_{N,\delta}(p)\} = T^E_{N,\delta}(p), \tag{5.37}$$

so that (4.44) is the same as (5.34).

The proof of the upper bound in (5.35) is almost trivial: by definition (5.33) we have

$$1 = p^N(A^N) \geq p^N(T^E_{N,\delta}(p)) = \sum_{\sigma \in T^E_{N,\delta}(p)} p^N(\sigma) \geq |T^E_{N,\delta}(p)| e^{-N(S(p)+\delta)}. \tag{5.38}$$

The proof of the lower bound in (5.35) relies on (5.34), which for $\delta > 0$ gives $M = M(\delta)$ such that

$$p^N(T^E_{N,\delta}(p)) > 1 - \delta, \tag{5.39}$$

for all $N > M$. A similar computation to the one just given then yields the lower bound in (5.35):

$$1 - \delta < p^N(T^E_{N,\delta}(p)) = \sum_{\sigma \in T^E_{N,\delta}(p)} p^N(\sigma) \leq |T^E_{N,\delta}(p)| e^{-N(S(p)-\delta)}. \tag{5.40}$$

There is also a *strong* version of (5.34), called the *Shannon–McMillan–Breiman theorem*:

$$\lim_{N\to\infty} \frac{1}{N} \log p^N(s_{|N}) = -S(p), \tag{5.41}$$

for $p^{\mathbb{N}}$-almost all $s \in A^{\mathbb{N}}$. This removes the δ-buffer $T^E_{N,\delta}(p)$ around $T_N(p)$ and states that the remarkable formula (5.24), originally proved for $\sigma \in T_N(p)$, is asymptotically generic.

Eq. (5.41) is a special case of what in ergodic theory is still called the Shannon–McMillan–Breiman theorem. Recall the setting of (probabilistic) dynamical systems, where we start with a partition (2.28) of some probability space X, which is subsequently refined to (2.29). The latter partition π^N of X then gives rise to the entropy (2.38). As we see from (2.35), this entropy is the *average* of the information function $I_P(\pi^N)$, see (2.37), w.r.t. P. The SMB theorem shows that as $N \to \infty$ one may replace this average by judiciously chosen *pointwise* values of $I_P(\pi^N)$:[80]

Theorem 5.2 *If (X, P, T) is ergodic with finite partition* (2.28)*, then for P-almost every $x \in X$,*

$$h_P(\pi) = \lim_{N\to\infty} \frac{1}{N} I_P(\pi^N)(x), \tag{5.42}$$

where $\pi^N(x)$ is the cell of the partition π^N that contains x, cf. (2.43).

Exercise 42 *Show that* (5.41) *is a special case of* (5.42)*.*

Proof of Theorem 5.2. Consider the function $I_P(\pi^N)$. Recalling (2.51), for $N > 1$ we abbreviate

$$f_N := I_P(\pi | \wedge_{n=1}^{N-1} T^{-n}\pi) = -\sum_{A\in\pi} \sum_{B\in\pi_1^N} 1_{A\cap B} \log\left(\frac{P(A\cap B)}{P(B)}\right), \tag{5.43}$$

cf. (2.39). A nontrivial measure-theoretic argument shows that the sequence (f_N) has a limit

$$f(x) = \lim_{N\to\infty} f_N(x) \tag{5.44}$$

[80] The problem of explicitly characterizing such points (as opposed to just saying 'for P-almost every $x \in X$') occurs throughout ergodic theory, starting of course with the pointwise ergodic theorem (Theorem 4.28), of which Theorem 5.2 is a highly nontrivial corollary. This problem will to some extent be solved in terms of algorithmic randomness theory.

pointwise P-a.e.[81] This is the function we now apply the pointwise (Birkhoff) ergodic theorem to (in the last line below), as well as the dominated convergence theorem. Using (2.40) and (2.54),

$$\begin{aligned} h_P(\pi) &= \lim_{N\to\infty} H_P(\pi|\pi_1^N) = \lim_{N\to\infty} \int_X dP\, I_P(\pi|\pi_1^N) = \lim_{N\to\infty} \int_X dP\, f_N = \int_X dP \lim_{N\to\infty} f_N \\ &= \int_X dP f = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x), \end{aligned} \tag{5.45}$$

for P-almost every $x \in X$. To relate this to Theorem 5.2, we iterate (2.48) and use the property

$$I_P(T^{-1}\pi)(x) = I_P(\pi)(Tx), \tag{5.46}$$

which follows from (2.2), to compute

$$\begin{aligned} I_P(\pi^N) &= I_P(\wedge_{n=0}^{N-1} T^{-n}\pi) = I_P(\wedge_{n=1}^{N-1} T^{-n}\pi \wedge \pi) = I_P(\wedge_{n=1}^{N-1} T^{-n}\pi) + I_P(\pi|\wedge_{n=1}^{N-1} T^{-n}\pi) \\ &= I_P(\wedge_{n=0}^{N-2} T^{-n}\pi)\circ T + f_N = I_P(\wedge_{n=1}^{N-2} T^{-n}\pi \wedge \pi)\circ T + f_N \\ &= I_P(\wedge_{n=1}^{N-2} T^{-n}\pi)\circ T + I_P(\pi|\wedge_{n=1}^{N-2} T^{-n})\circ T + f_N \\ &= I_P(\wedge_{n=0}^{N-3} T^{-n}\pi)\circ T^2 + f_{N-1}\circ T + f_N = \cdots = \sum_{n=0}^{N-1} f_{N-n}\circ T^n, \end{aligned} \tag{5.47}$$

where $f_1 := I_P(\pi)$ by convention (here the dots replace a routine proof by induction). This is not quite the same as the right-hand side of (5.45); the proof is finished by showing that P-a.e.,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} (f(T^n x) - f_{N-n}(T^n(x)) = 0. \tag{5.48}$$

This may be unsurprising in view of (5.44), but the argument is very technical and we omit it.[82] □

6 Entropy and (very basic) coding theory

The AEP has interesting consequences for coding (or data compression). A *binary code* of A^N (for some finite alphabet A) is an injective map $C^{(N)} : A^N \to 2^*$. Given some probability distribution $p^{(N)}$ on A^N, the *average length* (i.e. number of bits) of a codeword is given by

$$\langle \ell(C^{(N)}) \rangle_{p^{(N)}} := \sum_{\sigma\in A^N} p^{(N)}(\sigma)\ell(C^{(N)}(\sigma)). \tag{6.1}$$

We restrict ourselves to *memoryless sources*; this means that A^N is distributed by p^N for some $p \in \mathrm{Prob}(A)$. A simple possibility is to choose a code $C : A \to 2^*$ and extend this to A^N by

$$C^N(\sigma) := C(\sigma_0)\cdots C(\sigma_{N-1}). \tag{6.2}$$

In that case the average length (6.1) is given by

$$\langle \ell(C^{(N)}) \rangle_{p^N} = N \sum_{a\in A} p(a)\ell(C(a)) = N\langle \ell(C) \rangle_p. \tag{6.3}$$

[81] See e.g. Dajani & Kalle, §9.4. The point is to rewrite $I_P(\alpha|\beta)$ as conditional expectation with respect to the σ-algebra $\sigma(\beta)$ generated by the partition β, namely $I_P(\alpha|\beta)(x) = -\sum_{A\in\pi} 1_A(x)\log E_P(1_A|\sigma(\beta))$. This gives the pointwise limit function as $f(x) = I_P(\pi|\sigma(\cup_{n=1}^\infty T^{-n}\pi))$.

[82] See Dajani & Kalle, §9.4.

Another possibility is to encode A^N using some list (perhaps based on listing A). This requires approximately $\log_2(|A|^N) = N\log_2(|A|)$ bits (which is exact if $|A|$ is a power of two). This gives

$$\lim_{N\to\infty}\frac{1}{N}\langle \ell(C^{(N)}_{naive})\rangle_{p^N} = \log_2(|A|) = S_2(f), \tag{6.4}$$

cf. (1.10). But if $p \neq f$ we can do better if we use the AEP. Encode the elements of A^N by:[83]

1. Encoding elements of $T^E_{N,\delta}(p) \subset A^N$ naively via some list. By (5.35) this subset has $\approx e^{NS(p)} = 2^{NS_2(p)}$ elements and hence this requires $NS_2(p)+2$ bits: one extra bit since $S_2(p)$ may not be an integer and we incorporate $\delta > 0$, and another extra bit to add a prefix 0 indicating that we have a string in $T^E_{N,\delta}(p)$.

2. Likewise encoding the complement $A^N \backslash T^E_{N,\delta}(p)$, adding a prefix 1. Taking A^N as an upper bound of the size of $A^N \backslash T^E_{N,\delta}(p)$, this takes $NS_2(f)+2 = N\log_2(|A|)+2$ bits, cf. (1.10).

Call this coding $C^{(N)}_\delta$, for some $\delta > 0$. We may estimate the average codeword length in $C^{(N)}_\delta$ by

$$\begin{aligned}
\langle \ell(C^{(N)}_\delta)\rangle_{p^N} &= \sum_{\sigma\in A^N} p^N(\sigma)\ell(C^{(N)}_\delta(\sigma)) = \sum_{\sigma\in T^E_{N,\delta}(p)} p^N(\sigma)\ell(C^{(N)}_\delta(\sigma)) + \sum_{\sigma\notin T^E_{N,\delta}(p)} p^N(\sigma)\ell(C^{(N)}_\delta(\sigma)) \\
&\leq \sum_{\sigma\in T^E_{N,\delta}(p)} p^N(\sigma)\cdot NS_2(p) + \sum_{\sigma\notin T^E_{N,\delta}(p)} p^N(\sigma)\cdot NS_2(f) + O(1) \\
&\leq p^N(T^E_{N,\delta}(p))\cdot NS_2(p) + p^N(A^N\backslash T^E_{N,\delta}(p))\cdot NS_2(f) + O(1) \\
&\leq N(S_2(p) + \delta S_2(f)) + O(1),
\end{aligned} \tag{6.5}$$

where for the last inequality we used the weak LLN (5.39), or rather its equivalent version

$$p^N(A^N\backslash T^E_{N,\delta}(p)) \leq \delta, \tag{6.6}$$

which is valid for sufficiently large N. In the limit one has a clean result

$$\lim_{\delta\to 0}\lim_{N\to\infty}\frac{1}{N}\langle \ell(C^{(N)}_\delta)\rangle_{p^N} \leq S_2(p). \tag{6.7}$$

On the other hand, the very definition of the coding scheme $C^{(N)}_\delta$ gives the inequalities

$$\begin{aligned}
\langle \ell(C^{(N)}_\delta)\rangle_{p^N} &= \sum_{\sigma\in T^E_{N,\delta}(p)} p^N(\sigma)\ell(C^{(N)}_\delta(\sigma)) + \sum_{\sigma\notin T^E_{N,\delta}(p)} p^N(\sigma)\ell(C^{(N)}_\delta(\sigma)) \geq \sum_{\sigma\in T^E_{N,\delta}(p)} p^N(\sigma)\ell(C^{(N)}_\delta(\sigma)) \\
&\geq \sum_{\sigma\in T^E_{N,\delta}(p)} p^N(\sigma)\cdot NS_2(p) = p^N(T^E_{N,\delta}(p))\cdot NS_2(p) \geq (1-\delta)\cdot NS_2(p),
\end{aligned} \tag{6.8}$$

where we used the weak law of large numbers (5.39). Hence (6.7) may be supplemented with

$$\lim_{\delta\to 0}\lim_{N\to\infty}\frac{1}{N}\langle \ell(C^{(N)}_\delta)\rangle_{p^N} \leq S_2(p), \tag{6.9}$$

so that

$$\lim_{\delta\to 0}\lim_{N\to\infty}\frac{1}{N}\langle \ell(C^{(N)}_\delta)\rangle_{p^N} = S_2(p). \tag{6.10}$$

[83] We follow Cover & Thomas, §3.2.

Compared with (6.4), since $S_2(p) \leq S_2(f)$ with equality iff $p = f$ we see that (unless $p = f$) the improved coding $C_\delta^{(N)}$ considerably compresses the average length of long messages.[84]

This argument illustrates Shannon's *noiseless coding theorem*. We only discuss this theorem for codings of the type (6.2), so that we only need to talk about codes $C : A \to 2^*$. Finally, we only consider *prefix codes* (also called *instantaneous codes*), which are defined as follows.

Definition 6.1 *1. If $\sigma, \sigma' \in 2^*$ (or $\sigma \in 2^{\mathbb{N}}$) we say that σ' is a* prefix *of σ, written $\sigma' \prec \sigma$, if $\sigma = \sigma'\tau$ for some $\tau \in 2^*$ (or $\tau \in 2^{\mathbb{N}}$).*

2. A subset $S \subset 2^*$ is a prefix set *if no $\sigma' \in S$ is a prefix of any $\sigma \in S$.*

3. A map $C : A \to 2^*$ is a prefix code *if it is injective and $\{C(a), a \in A\}$ is a prefix set in 2^*.*

In other words, in a prefix code C, for any $a, b \in A$ there is no $\tau \in 2^*$ such that $C(b) = C(a)\tau$. Prefix codes are the best example of uniquely decodable codes, cf. Cover & Thomas, Chapter 5.

Lemma 6.2 (*Kraft inequality*) *1. Any prefix subset $S \subset 2^*$ satisfies*

$$\sum_{\sigma \in S} 2^{-\ell(\sigma)} \leq 1. \tag{6.11}$$

2. Consequently (since $C(A) \subset 2^*$ is a prefix set), for any prefix code $C : A \to 2^*$ we have

$$\sum_{a \in A} 2^{-\ell(C(a))} \leq 1. \tag{6.12}$$

3. Conversely, for any (at most countable) set A and subset $\{\ell_a\}_{a \in A}$ of $\mathbb{N}$ that satisfies

$$\sum_{a \in A} 2^{-\ell_a} \leq 1, \tag{6.13}$$

there exists a prefix code $C : A \to 2^*$ for which the codeword lengths satisfy

$$\ell(C(a)) = \ell_a. \tag{6.14}$$

Proof. Recall the cylinder set $[\sigma]_{\ell(\sigma)} = \{s \in 2^{\mathbb{N}} \mid s_{|\ell(\sigma)} = \sigma\}$. Since S is a prefix set,

$$[\sigma]_{\ell(\sigma)} \cap [\sigma']_{\ell(\sigma')} = \emptyset \tag{6.15}$$

whenever $\sigma \neq \sigma'$ (check this!). In terms the flat prior f on $2 = \{0, 1\}$, i.e. $f(0) = f(1) = 1/2$ and the ensuing Bernoulli measure $f^{\mathbb{N}}$ on $2^{\mathbb{N}}$, we have $f^{\mathbb{N}}([\sigma]_{\ell(\sigma)}) = 2^{-\ell(\sigma)}$, and hence

$$\sum_{\sigma \in S} 2^{-\ell(\sigma)} = \sum_{\sigma \in S} f^{\mathbb{N}}([\sigma]_{\ell(\sigma)}) = f^{\mathbb{N}}\left(\sum_{\sigma \in S} [\sigma]_{\ell(\sigma)}\right) \leq f^{\mathbb{N}}(2^{\mathbb{N}}) = 1. \tag{6.16}$$

Exercise 43 *Prove the converse, that is, prove part 3 of Lemma 6.2.*[85]

[84] The universal nature of $S_2(p)$ is confirmed if one tries to construct a similar coding based on some alternative subset $B_\delta \subset A^N$ for which $\lim_{N\to\infty} p^N(B_\delta) = 1$, cf. (5.34), so that for large N one can again achieve $p^N(B_\delta) > 1 - \delta$, cf. (5.39), which is the key to (6.8). Since for any probability space (X, P) and $A \subset X$ and $B \subset X$ such that $P(A) > 1 - \varepsilon_1$ and $P(B) = 1 - \varepsilon_2$ one has $P(A \cap B) > 1 - \varepsilon_1 - \varepsilon_2$; to see this, note that $P(A \cap B) = P(A) - P(A \backslash B)$ and $P(A \backslash B) \leq P(X \backslash B) = 1 - P(B)$. We therefore obtain $P(T^E_{N,\delta}(p) \cap B_\delta) > 1 - 2\delta$, so that for small $\delta > 0$ the set B_δ is (probabilistically) similar to $T^E_{N,\delta}(p)$ and we are back to the previous situation. See Cover & Thomas, Theorem 3.3.1.

[85] See e.g. Cover & Thomas, Theorem 5.2.1, or Austin, Lecture 3, Theorem 4.1.

Theorem 6.3 *1. Any prefix code satisfies* $\langle \ell(C) \rangle_p \geq S_2(p)$*, cf.* (6.3).

2. There exists prefix codes C that satisfy

$$S_2(p) \leq \langle \ell(C) \rangle_p \leq S_2(p) + 1. \tag{6.17}$$

In the rare cases where $p(a) = 2^{-k(a)}$ for some integer $k(a) \in \mathbb{N}$, codes exist for which

$$\ell(C(a)) = I_2(a), \tag{6.18}$$

for each $a \in A$; this is the only way to achieve $\langle \ell(C) \rangle_p = S_2(p)$, cf. Cover & Thomas, §5.4. Thus the information $I_2(a)$ in a is roughly the length of the codeword $C(a)$ in an optimal coding C. This should of course also apply if we take (A^N, p^N, C^N) instead of (A, p, C). Indeed, we have

$$S_2(p^N) = -\sum_{\sigma \in A^N} p^N(\sigma) \log_2 p^N(\sigma) = N S_2(p), \tag{6.19}$$

and combining this with (6.3) we see that Theorem 6.3 consistently describes both the first and the second case. Of course, taking $(A^N, P^{(N)}, C^{(N)})$ for (A, p, C) one is not restricted to the choices $P^{(N)} = p^N$ and $C^{(N)} = C^N$: hence Theorem 6.3 also covers general sources and "block codes".

Proof. Part 1 is an exercise. Part 2 follows by construction: for the *Shannon code* C_S, order

$$A = (a_1, \ldots, a_{|A|}); \qquad\qquad p(a_k) \geq p(a_{k+1}), \tag{6.20}$$

and take $C_S(a_k)$ to be the first $\lceil I_2(a_k) \rceil$ bits from the binary expansion of $\sum_{l=1}^{k-1} p(a_l)$. Then trivially

$$\ell(C_S(a)) = \lceil -I_2(a) \rceil, \tag{6.21}$$

i.e. the smallest integer larger than or equal to $I_2(a) = -\log_2(p_a)$, cf. (1.11). This gives the bound

$$\langle \ell(C_S) \rangle_p \leq S_2(p) + 1. \tag{6.22}$$

Exercise 44 *Prove Theorem 6.3.1. Hint: Jensen's inequality for the convex function* $x \mapsto -\log_2(x)$.

The Shannon code is mainly of theoretical interest; in practice other codes C are used with

$$\langle \ell(C) \rangle_p \leq \langle \ell(C_S) \rangle_p. \tag{6.23}$$

The famous *Huffman code* achieves the lowest possible value of $\langle \ell(C) \rangle_p$ among all prefix codes.

Exercise 45 *Look up and describe this code.*

7 Large deviations: Sanov's theorem

The inequalities (5.18) lead to our first result in large deviation theory, called *Sanov's Theorem* (from 1957). This theorem describes fluctuations of the random variables (L_N). We keep A finite. Initially defined by (5.1) as a function $L_N : A^N \to \mathrm{Prob}(A)$, we now regard each L_N as a function

$$L_N : A^{\mathbb{N}} \to \mathrm{Prob}(A); \qquad\qquad L_N(s) = \frac{1}{N} \sum_{n=0}^{N-1} \delta_{s(n)}, \tag{7.1}$$

whose value $L_N(s)$ obviously only depends on $s_{|N}$. Like its original version, L_N is a convex sum of probability distributions (or measures) on A and hence is a probability distribution on A itself. We then have a strong law of large numbers, in which convergence in $\mathrm{Prob}(A)$ is defined weakly:

Proposition 7.1 *For any $q \in \mathrm{Prob}(A)$, as $N \to \infty$ we have $L_N(s) \to q$ for $q^{\mathbb{N}}$-almost every $s \in A^{\mathbb{N}}$.*

Proof. This follows from the usual strong law of large numbers (SLLN) for i.i.d. $\mathbb{R}$-valued random variables.[86] For each $a \in A$ separately, define $\delta_n^{(a)} : A^{\mathbb{N}} \to \{0,1\}$ by $\delta_n^{(a)}(s) := \delta_{s(n)a}$. Then

$$q^{\mathbb{N}}(\delta_n^{(a)} = 1) = q^{\mathbb{N}}(\{s \in A^{\mathbb{N}} \mid \delta_n^{(a)}(s) = 1\}) = q^{\mathbb{N}}(s_n = a) = q(a), \tag{7.2}$$

any by definition of $q^{\mathbb{N}}$ the $\delta_n^{(a)}$ are also independent with respect to $q^{\mathbb{N}}$, so that they are i.i.d. with distribution $P_q(1) = q(a)$. Defining $L_N^{(a)} : A^{\mathbb{N}} \to [0,1]$ by $L_N^{(a)} = \frac{1}{N}\sum_{n=0}^{N-1} \delta_n^{(a)}$, the SSLN gives

$$\lim_{N\to\infty} L_N^{(a)}(s) = \langle \delta_n^{(a)} \rangle_{P_q} = 0 \cdot P_q(\delta_n^{(a)} = 0) + 1 \cdot P_q(\delta_n^{(a)} = 1) = q(a), \tag{7.3}$$

for $q^{\mathbb{N}}$-almost every $s \in A^{\mathbb{N}}$. But clearly $L_N^{(a)}(s) = (L_N(s))(a)$, so that $(L_N(s))(a) \to q(a)$ for each $a \in A$ and hence $L_N(s) \to q$ weakly, for $q^{\mathbb{N}}$-almost every $s \in A^{\mathbb{N}}$. □

Sanov's theorem describes the (exponentially small) probability of "large" or $O(1)$ fluctuations of L_N (i.e. deviation from its mean q) for large N, in contrast to "small" or $O(1/\sqrt{N})$ fluctuations, which are described by the central limit theorem. The main actor in the theorem is the relative entropy (5.12), *seen as a function of p* for some *given* prior $q \in \mathrm{Prob}(A)$. For the moment we maintain our standing assumption that A is a finite set. To emphasize this, we henceforth write

$$D_q(p) := D(p\|q). \tag{7.4}$$

Hence the domain $\mathscr{D}_{D_q} := \{p \in \mathrm{Prob}(A) \mid D_q(p) < \infty\}$ of the function $D_q : \mathrm{Prob}(A) \to [0,\infty]$ consists of all $p \in \mathrm{Prob}(A)$ for which $p \ll q$. Further to Proposition 5.1, we have:

Proposition 7.2 *The function $p \mapsto D(p\|q)$ is strictly convex and continuous on its domain.*[87]

The first part follows from strict convexity of $x \mapsto x\log x$ for $x \geq 0$. Continuity of D_q follows, since the function $x \mapsto x\log x$ is continuous for $x \geq 0$, so that by its definition (5.12), D_q is just a finite sum of continuous functions.[88] □

For any function $I : \mathscr{X} \to [-\infty,\infty]$ and $B \subset \mathscr{X}$, here given by $\mathscr{X} = \mathrm{Prob}(A)$, we write

$$I(B) := \inf_{x\in B} I(x). \tag{7.5}$$

We apply this notation to $\mathscr{X} = \mathrm{Prob}(A)$ and $I(p) = D_q(p)$, where $q \in \mathrm{Prob}(A)$ is a parameter.

Theorem 7.3 *Let A be a finite set and $q \in \mathrm{Prob}(A)$ a prior. Define the empirical measures $L_N : A^{\mathbb{N}} \to \mathrm{Prob}(A)$ by (7.1). Then for any (weakly measurable) subset $B \subset \mathrm{Prob}(A)$ one has*

$$-D_q(\mathring{B}) \leq \liminf_{N\to\infty} \frac{1}{N}\log q^N(L_N \in B) \leq \limsup_{N\to\infty} \frac{1}{N}\log q^N(L_N \in B) \leq -D_q(B^-). \tag{7.6}$$

Here $\mathring{B}$ is the interior of B and B^- is its closure with respect to the weak topology on $\mathrm{Prob}(A)$. In particular, if $B \subset \mathrm{Prob}(A)$ satisfies $D_q(\mathring{B}) = D_q(B^-)$, then (7.6) implies the more palatable limit

$$\lim_{N\to\infty} \frac{1}{N}\log q^N(L_N \in B) = -D_q(B). \tag{7.7}$$

Since for finite A the function $p \mapsto D_q(p)$ is continuous on its domain $\mathscr{D}_{D_q}$ (i.e. the subset of $\mathrm{Prob}(A)$ where it is finite), the stated hypothesis for (7.7) is satisfied if $B \subset (\mathring{B})^- \subset \mathscr{D}_{D_q}$.

[86] The result remains valid for Polish spaces A by taking fixed Borel sets $B \subset \mathrm{Prob}(A)$ instead of single elements of A, or by using Sanov's theorem below.

[87] Moreover, the function $(p,q) \mapsto D(p\|q)$ is jointly convex, but this will not be used.

[88] Continuity on the interior of the domain already followed from Proposition A.6.

Exercise 46 *Show that* (7.6) *is equivalent to*

$$\limsup_{N\to\infty} \frac{1}{N} \log q^N(L_N \in F) \le -D_q(F) \ \ (F \subset \mathrm{Prob}(A) \text{ closed}); \tag{7.8}$$

$$\liminf_{N\to\infty} \frac{1}{N} \log q^N(L_N \in U) \ge -D_q(U) \ \ (U \subset \mathrm{Prob}(A) \text{ open}). \tag{7.9}$$

Eq. (7.6) or eqs. (7.8) - (7.9) is called a *large deviation principle* (LDP). There are good reasons for the careful way this principle is stated:

Exercise 47 *Find examples where* (7.7) *is false if the stated assumptions are not satisfied:*

1. *For general (A,q), give some B for which the limit in* (7.7) *does not exist.*
2. *For $A=\{0,1\}$ and $q(0)=0$, give some $B \subset (\mathring{B})^-$ for which $D_q(\mathring{B})=\infty$ and $D_q(B^-)<\infty$.*

To see what is going on, first assume that $q \in B^-$. Then $D_q(B)=0$ by Proposition 5.1, and hence $q^N(L_N \in B) \to 1$ or $q^{\mathbb{N}}(L_N \in B) \to 1$ from (7.11), as we already knew from the law of large numbers. But if $q \notin B$, then (7.7) states that the "large fluctuation" $q^N(L_N \in B)$ is exponentially improbable, with rate function $D_q(B)$, i.e.

$$q^N(L_N \in B) \approx e^{-ND_q(B)}, \tag{7.10}$$

where the meaning of $\approx$ is *defined* by (7.7); the point is that $D_q(B)>0$ whenever $q \notin B^-$.

Proof of Theorem 7.3. The inequalities that imply the theorem are, for each $B \subset \mathrm{Prob}(A)$,

$$e^{-ND_q(\mathring{B})-o(N)} \le q^N(L_N \in B) \le e^{-ND_q(B^-)+o(N)}. \tag{7.11}$$

The upper bound in (7.11) follows from the upper bound in (5.18), viz.

$$\begin{aligned} q^N(L_N \in B) &= \sum_{p \in B \cap \mathrm{Prob}_N(A)} q^N(L_N = p) \le \sum_{p \in B \cap \mathrm{Prob}_N(A)} e^{-ND_q(p)} \\ &\le |B \cap \mathrm{Prob}_N(A)| e^{-ND_q(B)} \le (N+1)^{|A|} e^{-ND_q(B)} = e^{-ND_q(B)+o(N)}, \end{aligned} \tag{7.12}$$

where we used (5.27). Since $\inf_{x\in B^-} I(x) \le \inf_{x\in B} I(x)$ for any I and B, and hence

$$D_q(B^-) \le D_q(B) \qquad \Rightarrow \qquad \exp(-ND_q(B)) \le \exp(-ND_q(B^-)), \tag{7.13}$$

because $x \mapsto \exp(-x)$ is decreasing, we obtain the upper bound in (7.11).

The lower bound follows by taking $\varepsilon>0$ and finding $p \in \mathring{U}$ such that

$$D_q(p) < D_q(\mathring{U}) + \varepsilon; \tag{7.14}$$

this is possible by the definition of inf, cf. (7.5). Furthermore, for any $p \in \mathrm{Prob}(A)$ we can find a sequence (p_N) in $\mathrm{Prob}_N(A)$ with $\|p-p_N\|_1 = O(1/N)$, and *since $\mathring{U}$ is open* there is a (large) M such that $p_N \in \mathring{U}$ for all $N>M$. Then the lower bound in (5.18) gives

$$q^N(L_N \in U) \ge q^N(L_N = p_N) \ge e^{-ND_q(p_N)-o(N)} \ge e^{-N(D_q(\mathring{U}\|q)+\varepsilon)-o(N)}, \tag{7.15}$$

in which we let $\varepsilon \to 0$. This completes the proof of the first inequality in (7.11), and hence of (7.11). Taking $B=F$ closed then gives (7.8), whilst $B=U$ open yields (7.9).

Finally, if $D_q(\mathring{B}) = D_q(B^-)$, taking $F=B^-$ in (7.8) and $U=\mathring{B}$ in (7.9) gives (7.7). □

Theorem 7.3 in fact implies the SLLN, i.e. $L_N \to q$ pointwise q^N-a.e. To see this, take $\delta > 0$ and let $B_\delta(q)$ be the open δ-ball in $\mathrm{Prob}(A)$ centered at q (with respect to some metric). Let $C_N \subset A^{\mathbb{N}}$ be the event $L_N \notin B_\delta(q)$, i.e. $L_N \in B_\delta(q)^c$. Since $D_q(B_\delta(q)^c) > 0$ by construction, eq. (7.7) guarantees that $\sum_{N=1}^{\infty} P(C_N) < \infty$, so that by the Borel–Cantelli lemma q^N-a.e. $s \in A^{\mathbb{N}}$ lies in only finitely many C_N. Hence for each $\delta > 0$ and q^N-a.e. $s \in A^{\mathbb{N}}$ we have $\lim_{N\to\infty} L_N(s) \in B_\delta(q)$, so that $\lim_{N\to\infty} L_N(s) = q$ almost surely. This argument in quite general, so that the large deviation principle (when it applies) not only *refines* but also *implies* the strong law of large numbers.

We now generalize the previous material to *Polish spaces* A; one advantage of working with such spaces A is that $\mathrm{Prob}(A)$, equipped with the weak topology, is also Polish.

For $p, q \in \mathrm{Prob}(A)$, the *Kullback–Leibler divergence* or *relative entropy* is now defined as

$$D_q(p) := \int_A dq \frac{dp}{dq} \log\left(\frac{dp}{dq}\right) = \int_A dp \log\left(\frac{dp}{dq}\right) \quad \text{if } p \ll q; \tag{7.16}$$

$$D_q(p) := \infty \quad \text{otherwise}, \tag{7.17}$$

where dp/dq is the Radon–Nikodym derivative, which exists because in (7.16) we assume $p \ll q$ (that is, $q(B) = 0$ implies $p(B) = 0$ for any measurable $B \subset A$). Here are the basic facts.[89]

1. Proposition 5.1.1 remains true.

2. In contrast to the finite case, the function $p \mapsto D_q(p)$ may no longer be continuous on its domain (i.e. the subset of $\mathrm{Prob}(A)$ where it is finite), but it is at least lower semicontinuous.

Returning to Sanov's theorem, let (X_n) be i.d.d. random variables taking values in some Polish space A, with distribution $q \in \mathrm{Prob}(A)$, once again seen as a "prior". Let

$$L_N = \frac{1}{N} \sum_{n=0}^{N-1} \delta_{X_n}; \tag{7.18}$$

be the empirical measure; in the Kolmogorov model, $X_n : A^{\mathbb{N}} \to A$ and (7.18) reduces to (7.1). Sanov's theorem then holds also for Polish space: the statement of Theorem 7.3 is the same.[90]

We now compute the Fenchel transform of the function $p \mapsto D_q(p)$ for fixed q. To get into the setting of Definition A.9 we extend D_q to the space $\mathscr{M}(A)$ of all finite singed Borel measures on A, seen as a Banach space in its role of the dual space of $C_b(A)$ with supremum-norm. Neither topology will actually be used; what matters is the duality

$$\langle p, f \rangle = \int_A dp\, f = \langle f \rangle_p. \tag{7.19}$$

With $\mathrm{Prob}(A) \subset \mathscr{M}(A)$ this extension is defined by:

$$D_q(p) := \int_A dp \log\left(\frac{dp}{dq}\right) \quad \text{if } p \in \mathrm{Prob}(A) \text{ and } p \ll q$$

$$D_q(p) := \infty \quad \text{otherwise} \tag{7.20}$$

For any $f \in C_b(A)$, define the "partition function" and the "pressure" by

$$Z_q(f) := \langle e^f \rangle_q = \int_A dq(a)\, e^{f(a)}; \qquad \Pi_q(f) := \log Z_q(f), \tag{7.21}$$

[89] See for example Simon (2011), Chapter 16, or Rassoul-Agha, §5.1.

[90] See e.g. Rassoul-Agha & Seppäläinen, 2015, §5.2, or Dembo & Zeitouni, §6.2.

respectively, so that $Z_q(f) = e^{\Pi_q(f)}$. Assuming A is Polish, we have a clean case of Fenchel duality:

$$D_q(p) = \sup_{f \in C_b(A)} \{\langle f \rangle_p - \Pi_q(f)\} \qquad (D_q = \Pi_q^*); \tag{7.22}$$

$$\Pi_q(f) = \sup_{p \in \mathrm{Prob}(A)} \{\langle f \rangle_p - D_q(p)\} \qquad (\Pi_q = D_q^*); \tag{7.23}$$

in (7.23) we may take the supremum over all $p \in \mathscr{M}(A)$, since in view of (7.20) non-probability measures $p \in \mathscr{M}(A) \backslash \mathrm{Prob}(A)$ do not contribute to the supremum. Let us note that (7.22) follows from (7.23) plus Theorem A.10, which applies because of Proposition 5.1, which as we noted also applies to Polish spaces A. Conversely, (7.23) follows from (7.22) and Theorem A.10, plus Lemma 7.4 below. Thus (7.22) and (7.23) are equivalent and only one needs to be proved.

Lemma 7.4 *The function $f \mapsto \Pi_q(f)$ on $C_b(A)$ is convex and lsc (in norm).*

Exercise 48 *Prove convexity of $\Pi_q(f)$ from Hölder's inequality.*

Lower semicontinuity of Π_q follows from Fatou's lemma, which states that for any continuous (or even lsc) function $F : \mathbb{R} \to [0, \infty]$ and weakly convergent sequence $P_n \to P$ in $\mathrm{Prob}(\mathbb{R})$ we have

$$\liminf_{n \to \infty} \int_{\mathbb{R}} dP_n F \geq \int_{\mathbb{R}} dP F. \tag{7.24}$$

For $f \in C_b(A)$ the probability $q \in \mathrm{Prob}(A)$ induces a unique probability $P \in \mathrm{Prob}(\mathbb{R})$ such that

$$\Pi_q(f) = \log \int_A dq(a)\, e^{f(a)} = \log \int_{\mathbb{R}} dP(x) e^x, \tag{7.25}$$

viz. $P(B) = q(f^{-1}(B))$, where $B \subset \mathbb{R}$ is (Borel) measurable. Likewise, f_n determines P_n. If $f_n \to f$ in $C_b(A)$, then $P_n \to P_n$ weakly in $\mathrm{Prob}(\mathbb{R})$, whence Fatou's lemma with $F(x) = \exp(x)$ gives

$$\liminf_n \Pi_q(f_n) = \liminf_n \log \int_{\mathbb{R}} dP_n(x) e^x \geq \log \int_{\mathbb{R}} dP(x) e^x = \Pi_q(f), \tag{7.26}$$

using also continuity and monotonicity of log. Then (A.9) makes Π_q lower semicontinuous. □

For finite A, eq. (7.22) can be proved by direct computation, as follows. This is very instructive. A function $f \in C_b(A)$ is now simply an $|A|$-tuple $x = (x_a)$, and similarly $p \in \mathscr{M}(A)$ is an $|A|$-tuple $p = (p_a)$; for $p \in \mathrm{Prob}(A)$ we of course need $p(a) \geq 0$ and $\sum_a p(a) = 1$, which also applies to the given prior $q = (q_a)$. Eq. (7.22) then comes down to

$$D_q(p) = \sup_{x \in \mathbb{R}^{|A|}} \{g(x)\}; \qquad g(x) := \sum_a x_a p_a - \log \sum_b q_b e^{x_b}, \tag{7.27}$$

where $D_q(p) = D(p\|q)$ as given by (5.12) - (5.13). To prove that $\sup_x g(x) = D_q(p)$, first assume that $p \in \mathrm{Prob}(A)$ and that $p \ll q$, which now means that $q_a = 0$ implies $p_a = 0$. In this case we try to compute the supremum by extremizing g, hoping to find a maximum. Indeed, we try to solve

$$\frac{\partial g(x)}{\partial x_a} = p_a - \frac{q_a e^{x_a}}{\sum_b q_b e^{x_b}} = 0, \tag{7.28}$$

which gives $p_a = q_a e^{x_a} / \sum_b q_b e^{x_b}$, from which x can be partly solved. If $q_a > 0$ this gives

$$x_a = \log(p_a/q_a) + \log \left(\sum_b q_b e^{x_b} \right), \tag{7.29}$$

whilst $q_a = 0$ simply returns our assumption $p_a = 0$. Either way, a fortunate cancellation gives

$$g(\bar{x}) = \sum_{a \in A} p(a) \log \left(\frac{p(a)}{q(a)} \right) \tag{7.30}$$

at the value $\bar{x}$ of x where (7.28) holds. If $q(a) = 0$, then $p(a) = 0$ and the corresponding term in the sum vanishes. This is also the case of $p(a) = 0$ whilst $q(a) > 0$. To find out if g assumes a maximum at the extremum $\bar{x}$, we compute the Hessian at $\bar{x}$ as

$$\frac{\partial^2 g(\bar{x})}{\partial x_a \partial x_b} = -\delta_{ab} + p_a p_b. \tag{7.31}$$

This matrix is negative-semidefinite,[91] and hence $\bar{x}$ is a maximum of g, so that (7.27) follows.

- If $p \in \mathrm{Prob}(A)$ but $p \ll q$ fails, then there is some a for which $q(a) = 0$ and $p(a) > 0$. In that case (7.28) has no solution and $g(x) \to \infty$ as $x^a \to \infty$, making $\sup_x g(x)$ infinite.
- Similarly, if $p \notin \mathrm{Prob}(A)$, there are two cases:
  1. If at least one $p_a < 0$ then obviously (7.28) has no solution for that value of a (since $q_a \geq 0$), and once again $g(x) \to \infty$, but this time for $x^a \to -\infty$.
  2. If all $p_a \geq 0$ but $\sum_a p_a \neq 1$, then (7.28) has no solution either, since if it had one, then

$$\sum_a p_a = \sum_a \frac{q_a e^{x_a}}{\sum_b q_b e^{x_b}} = 1. \tag{7.32}$$

For pedagogical contrast, instead of (7.22) we now prove (7.23) for general Polish spaces A.[92]

Eqs. (7.16) - (7.17) imply that the supremum in (7.23) may equally well be taken only over those $p \in \mathrm{Prob}(A)$ for which $p \ll q$, so that by the Radon–Nikodym theorem there is $g := dp/dq$ in $L^1(A,q)$, with $g \geq 0$ and $\int dq\, g = \int dp = 1$. Hence eq. (7.23), the one to be proved, becomes

$$\log \int_A dq\, e^f = \sup_{g \in L^1(A,q), g \geq 0, \int dq\, g = 1} \left\{ \int_A dq\, g (f - \log g) \right\}. \tag{7.33}$$

We prove $\leq$ and $\geq$ to give $=$. Taking $g = e^f / \int_A dq\, e^f$ in the curly brackets on he right-hand side gives the left-hand side, which proves $\leq$. Conversely, Jensen's inequality (A.4) for $-\log$ gives

$$\log \int_A dq\, e^f \geq \log \int_A dq\, 1_{g>0}\, g \cdot \frac{e^f}{g} = \log \int_{\{a \in A | g(a) > 0\}} dp\, \frac{e^f}{g} \geq \int_{\{a \in A | g(a) > 0\}} dp\, (f - \log g). \tag{7.34}$$

Since $dp = dq\, g$, this gives $\geq$ in (7.33), which given the proof of $\leq$ above is now proved. As explained above, this gives (7.23), and hence also (7.22). □

To finish this chapter we add some information about the relative entropy $D(p\|q)$, largely via exercises. This does not use large deviation theory, but it sheds light on D in a different way and is useful in both mathematical physics and in information theory (philosophically, *these fields do not overlap*). For simplicity we just discuss the case where p and q are defined on a finite set.[93]

[91] Its negative $M_{ab} = \delta_{ab} - p_a p_b$ is *diagonally dominant*, i.e. for each a we have $|M_{aa}| \geq \sum_{b \neq a} |M_{ab}|$; indeed we have equality for each a, since $|M_{aa}| = M_{aa} = p_a - p_a^2$ and $\sum_{b \neq a} |M_{ab}| = p_a \sum_{b \neq a} p_b = p_a \sum_b p_b - p_a^2 = p_a - p_a^2$. Since it is also symmetric with non-negative diagonal entries (as $p_a \geq p_a^2$ for $0 \leq p_a \leq 1$), M is positive semi-definite.

[92] We follow Dembo & Zeitouni, Lemma 6.2.13. See also Rassoul-Agha & Seppäläinen, Theorems 5.4 and 5.6.

[93] We mainly follow Cover & Thomas (2006), §§2.2–2.5.

If $A = B \times C$, then $p \in \mathrm{Prob}(A)$ gives rise to marginals $r \in \mathrm{Prob}(B)$ and $s \in \mathrm{Prob}(C)$ via:[94]

$$r(b) := \sum_{c \in C} p(b,c); \qquad\qquad s(c) := \sum_{b \in B} p(b,c). \tag{7.35}$$

These give rise to the product probability measure $q = r \times s$ on $B \times C$, and we expect that $D(p\|q)$ contains information about the probabilistic (in)dependence of B and C under p. Indeed, we have:

$$D(p\|r \times s) = S(r) + S(s) - S(p), \tag{7.36}$$

where S is the usual entropy (1.14), so

$$S(p) = -\sum_{b \in B, c \in C} p(b,c) \log p(b,c); \quad S(r) = -\sum_{b \in B} r(b) \log r(b); \quad S(s) = -\sum_{c \in C} s(c) \log r(c). \tag{7.37}$$

If $X : \Omega \to B$ and $Y : \Omega \to C$ are random variables with distributions $r \equiv P_X \in \mathrm{Prob}(B)$ and $s \equiv P_Y \in \mathrm{Prob}(C)$, respectively, and joint distribution $p \equiv P_{X,Y} \in \mathrm{Prob}(B \times C)$, so that $r(b) = P(X = b)$, $s(c) = P(Y = c)$, and $p(b,c) = P(X = b, Y = c)$, where $P \in \mathrm{Prob}(\Omega)$ is the probability that induces all the others, then, writing $S(X) = -\sum_{b \in B} P(X = b) \log P(X = b)$, etc., eq. (7.36) is written as

$$D(P_{X,Y}\|P_X \times P_Y) = S(X) + S(Y) - S(X,Y). \tag{7.38}$$

Note that Propositon 5.1, i.e., $D(P_{X,Y}\|P_X \times P_Y) \geq 0$, then gives subadditivity of the entropy, i.e.,

$$S(X,Y) \leq S(X) + S(Y). \tag{7.39}$$

Introducing the *conditional entropy* $S(Y|X)$, or equivalently $S(s|r)$, via, equivalently,

$$\begin{aligned} S(Y|X) &:= -\sum_{b \in B} P(X = b) \sum_{c \in C} P(Y = c|X = b) \log P(Y = c|X = b) \\ &= -\sum_{b \in B, c \in C} P(X = b, Y = c) \log P(Y = c|X = b); \end{aligned} \tag{7.40}$$

$$S(s|r) := -\sum_{b \in B, c \in C} p(b,c) \log \left(\frac{p(b,c)}{r(b)} \right), \tag{7.41}$$

we find that

$$S(X,Y) = S(X) + S(Y|X); \qquad\qquad S(p) = S(r) + S(s|r), \tag{7.42}$$

which turns (7.36) and (7.38) into

$$D(p\|r \times s) = S(s) - S(s|r) = S(r) - S(r|s); \tag{7.43}$$

$$D(P_{X,Y}\|P_X \times P_Y) = S(Y) - S(Y|X) = S(X) - S(X|Y). \tag{7.44}$$

Exercise 49 *Derive* (7.36) *or* (7.38)*, and* (7.42)*.*

[94] For Borel spaces A: Looking at (probability) measures p on A as functionals on $C_b(A)$ via expectation values, i.e., $p(f) = \langle f \rangle_p := \int_A dp\, f$, we simply have $r(g) = p(g)$ for $g \in C_b(B)$, seen on the right-hand side as a function on A (officially as the pullback $\pi_1^* g$ under the projection $\pi_1 : B \times C \to B$), and likewise $s(h) = p(h)$ for $h \in C_b(C)$.

8 Large deviations: Cramér's theorem

The next major result in large deviation theory is Cramér's theorem (originally from 1938). We first discuss this theorem for finite A, in which case it is a corollary of Sanov's theorem 7.3.[95] Pick an injective "energy" function $E : A \to \mathbb{R}$ and define a stochastic process

$$X_n : A^{\mathbb{N}} \to \mathbb{R}; \qquad X_n(s) = E(s_n), \tag{8.1}$$

with average $S_N := N^{-1}\sum_{n=1}^{N-1} X_n$, as in (1.36). The SLLN states that for any $q \in \mathrm{Prob}(A)$ and $q^{\mathbb{N}}$-almost every $s \in A^{\mathbb{N}}$, we have $S_N(s) \to \langle E\rangle_q$ as $N \to \infty$, where $\langle E\rangle_q := \sum_{a\in A} q(a)E(a)$. Cramér's theorem describes the probability of large deviations from $\langle E\rangle_q$.

Theorem 8.1 *For any injective function $E : A \to \mathbb{R}$, prior $q \in \mathrm{Prob}(A)$, and measurable $C \subset \mathbb{R}$,*

$$-I_q(\mathring{C}) \le \liminf_{N\to\infty} \frac{1}{N}\log q^N(S_N \in C) \le \limsup_{N\to\infty} \frac{1}{N}\log q^N(S_N \in C) \le -I_q(C^-), \tag{8.2}$$

where, writing $\langle E\rangle_p = \sum_{a\in A} p(a)E(a)$ as usual, I_q is defined by the constrained optimization

$$I_q(x) := \inf\{D_q(p) \mid p \in \mathrm{Prob}(A), \langle E\rangle_p = x\}; \tag{8.3}$$

$$I_q(x) := \infty \text{ if no } p \in \mathrm{Prob}(A) \text{ with } \langle E\rangle_p = x \text{ exists.} \tag{8.4}$$

Furthermore, $I_q(C) := \inf_{x\in C} I_q(x)$ as before.[96]

Moreover, the rate function I_q is alternatively given as a Fenchel transform

$$I_q(x) = \sup_{t\in\mathbb{R}}\{xt - \Pi_q(t)\} \qquad (I_q = \Pi_q^*) \tag{8.5}$$

of the "pressure" $\Pi_q(t)$ corresponding to the given "energy" function $E : A \to \mathbb{R}$, defined by

$$\Pi_q(t) := \log \tilde{Z}_q(t); \qquad \tilde{Z}_q(t) := \langle e^{tE}\rangle_q = \sum_{a\in A} q(a)e^{tE(a)}. \tag{8.6}$$

Both I_q and Π_q are convex and lsc (Π_q is even continuous), and hence are also related by

$$\Pi_q(t) = \sup_{x\in\mathbb{R}}\{xt - I_q(x)\} \qquad (\Pi_q = I_q^*). \tag{8.7}$$

Finally, $I_q(x) \in [0,\infty]$ and I_q has a unique minimum $I_q(\mu) = 0$ at $x = \mu := \langle E\rangle_q$.

Similarly to Sanov's theorem one may replace (8.2) by the pair

$$\limsup_{N\to\infty} \frac{1}{N}\log q^N(S_N \in F) \le -I_q(F) \ \ (F \subset \mathbb{R}\,\text{closed}); \tag{8.8}$$

$$\liminf_{N\to\infty} \frac{1}{N}\log q^N(S_N \in U) \ge -I_q(U) \ \ (U \subset \mathbb{R}\,\text{open}). \tag{8.9}$$

To see what is going on, we first study the case $A = \{0,1\}$, $E(a) = a$, and $q = f$. In that case,

$$I_f(x) = (1-x)\log(1-x) + x\log x + \log 2 \qquad (x \in [0,1]); \tag{8.10}$$

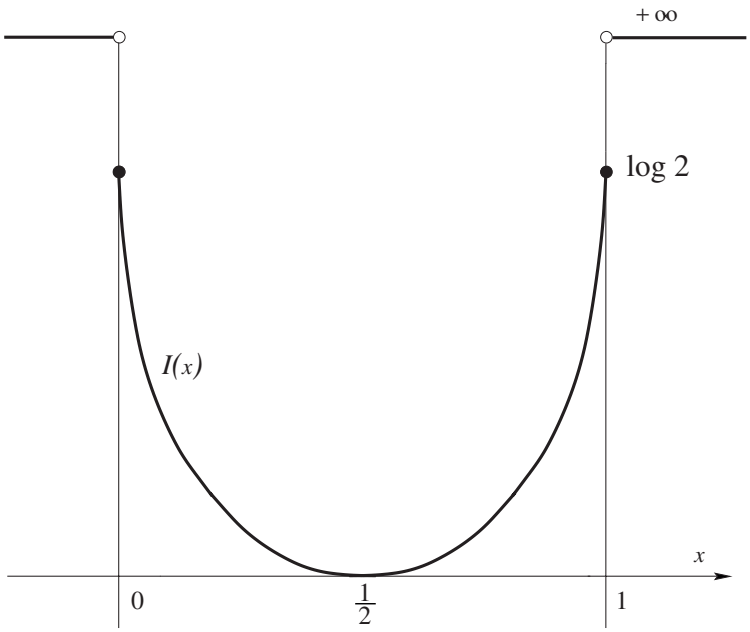
$$I_f(x) = \infty \qquad (x \notin [0,1]). \tag{8.11}$$

This function is plotted in the graph below.[97]

[95] We roughly follow Dembo & Zeitouni, §2.1.2.

[96] Eq. (8.4) follows from (8.3) if we agree that $\inf\emptyset = \infty$). Let $S \subset [-\infty,\infty]$. Then $z = \inf S$ iff: (1) $\forall_{s\in S}(z \le s)$; (2) $\forall_w((\forall_{s\in S}(w \le s)) \to w \le z)$. Now $\forall_{s\in S}F(s)$ is an abbreviation of $\forall_s(s \in S \to F(s))$. If $S = \emptyset$, then $s \in S$ is false and hence any implication of it is true, so (1) is always true. Likewise, in (2) the antecedent $\forall_{s\in S}(w \le s)$ is always true, and so to make (2) true the conclusion $w \le z$ must be true for all w. This forces $z = \infty$, i.e. $\inf\emptyset = \infty$. Similarly, $\sup\emptyset = -\infty$.

[97] The picture is copied from Olivieri & Vares, (2004), p. 4; the I in the picture is our I_f.



1. It is easy to check that (8.10) - (8.11) follows from (8.3) - (8.4). We know (or compute) that

$$D_f(p) = -S(p) + \log 2, \tag{8.12}$$

cf. (5.14), where $S(p)$ is given by (1.26) as usual. Since $p \in \text{Prob}(2)$ is given by $p(1) \equiv p \in [0,1]$, so that $p(0) = 1-p$, we obtain

$$D_f(p) = (1-p)\log(1-p) + p\log p + \log 2. \tag{8.13}$$

We also have $\langle E \rangle_p = p$ in (8.3), so that provided $x \in [0,1]$, the set of all $p \in \text{Prob}(s)$ for which $\langle E \rangle_p = x$ has a single element $p = x$. Hence (8.3) becomes (8.10). For $x \notin [0,1]$, the set just mentioned is empty . Recalling that $\inf \emptyset = \infty$, we infer that (8.4) gives (8.11).

2. We also show that (8.5) yields (8.10) - (8.11), which is more difficult. We compute

$$\Pi_f(t) = \log\left(p(0)e^0 + p(1)e^t\right) = \log(\tfrac{1}{2} + \tfrac{1}{2}e^t) = -\log 2 + \log\left(1 + e^t\right); \tag{8.14}$$

$$I_f(x) = \sup_{t\in\mathbb{R}}\{xt - \Pi_f(t)\} = \log 2 + \sup_{t\in\mathbb{R}}\left\{xt - \log(1+e^t)\right\}. \tag{8.15}$$

From this, proving (8.10) - (8.11) is largely a matter of straightforward calculus:

Exercise 50 *Consider the function $f_x(t) := xt - \log(1+e^t)$. Prove the following cases:*

- *For $x \in (0,1)$ the function $t \mapsto f_x(t)$ has a unique maximum at $t = t(x)$, at which value:*

$$f_x(t(x)) = (1-x)\log(1-x) + x\log x. \tag{8.16}$$

- *For $x = 0$ our function assumes no* maximum *but has a finite* supremum*, given by*

$$\sup_{t\in\mathbb{R}}\{-\log\left(1+e^t\right)\} = -\inf_{t\in\mathbb{R}}\{\log\left(1+e^t\right)\} = -\lim_{t\to-\infty}\log\left(1+e^t\right) = 0. \tag{8.17}$$

- *Similarly for $x = 1$: no maximum but a finite supremum, this time given by*

$$\sup_{t\in\mathbb{R}}\{t - \log\left(1+e^t\right)\} = \lim_{t\to\infty}(t - \left(1+e^t\right)) = 0. \tag{8.18}$$

- *For $x \notin [0,1]$, the function $t \mapsto f_x(t)$ has neither extrema nor asymptotes, and*

$$\sup_{t\in\mathbb{R}}\{f_x(t)\} = \infty. \tag{8.19}$$

3. The general properties of I_q claimed in the theorem are easily checked for I_f: for example, I_f is clearly convex on all of $\mathbb{R}$ and continuous on $(0,1)$, which is the interior of its domain $\mathscr{D}_{I_f}$ (defined as the set of points where I_f takes finite values). At the boundary points $0,1$ of $\mathscr{D}_{I_f}$ our function I_f is discontinuous but still lower semicontinuous (lsc), since it jumps to a higher value ∞. As argued in Appendix A, for convex functions the natural continuity condition is lower semicontinuity and so the natural combination is "convex + lsc", as for rate functions like I_q, or "minus" these properties namely "concave + upper semicontinuous (usc)", as is the case for entropies like $S = -D_f + \log 2$, cf. (8.12).

4. In this case, Theorem 8.1 is a special case of Theorem 7.3. To see this, note that

$$S_N(s) = (L_N(s))(1), \tag{8.20}$$

where we recall that like $p \in \mathrm{Prob}(2)$ above, we also have $L_N(s) \in \mathrm{Prob}(2)$, which like p is entirely defined by its value at $a = 1$. Hence $S_N \in C$ (for some $C \subset \mathbb{R}$) is the same as $L_N(1) \in C$, which by Theorem 7.3 is controlled by $D_f(C) = \inf_{p\in C}\{D_f(p)\}$. But clearly,

$$\begin{aligned}\inf_{p\in C}\{D_f(p)\} &= \inf\{D_f(p) \mid p = x, x \in C\} = \inf_{x\in C}\inf\{D_f(p) \mid p = x\} \\ &= \inf_{x\in C}\inf\{D_f(p) \mid \langle E\rangle_p = x\} = \inf_{x\in C} I_f(x) = I_f(C),\end{aligned} \tag{8.21}$$

since we already noted that $\langle E\rangle_p = p$ in our case (the general case is actually similar).

Exercise 51 *Repeat this entire analysis for an arbitrary prior $q \in \mathrm{Prob}(2)$.*

Before proving Theorem 8.1, we study (8.3) a bit more.[98] We assume $q(a) > 0$ for each $a \in A$. Let

$$E_- := \min_a E(a); \qquad\qquad E_+ := \max_a E(a). \tag{8.22}$$

For $\beta \in \mathbb{R}$, define $q_\beta \in \mathrm{Prob}(A)$ by

$$q_\beta(a) := \frac{q(a)e^{-\beta E(a)}}{Z_\beta}; \qquad\qquad Z_\beta := \sum_{b\in A} q(b)e^{-\beta E(b)}. \tag{8.23}$$

We extend these probabilities to $\beta = \pm\infty$ as follows. Let $a_\pm \in A$ be the elements for which $E_\pm = E(a_\pm)$; these are unique since E was assumed to be injective. We then define $q_{\pm\infty} \in \mathrm{Prob}(A)$ by

$$q_{\pm\infty}(a_\mp) = 1. \tag{8.24}$$

One may also obtain these by taking the limit $\beta \to \pm\infty$ in (8.23).

[98] The following theorem and proof are taken from Austin, Lecture 11. See also Borwein & Zhu, §4.7.

Proposition 8.2 *For each $x \in [E_-, E_+]$ there exists a unique $\beta(x) \in [-\infty, \infty]$ such that the infimum in* (8.3)*, is a unique minimum at $p = q_{\beta(x)}$. Moreover, the value of I_q is given by*

$$I_q(x) = D_q(q_{\beta(x)}) = -\beta x - \log Z_\beta \qquad (E_- < x < E_+); \tag{8.25}$$

$$I_q(E_\pm) = D_q(q_{\mp\infty}) = -\log q(a_{\mp\infty}) \qquad (x = E_\pm); \tag{8.26}$$

$$I_q(x) = \infty \qquad (x \notin [E_-, E_+]). \tag{8.27}$$

Note also that $q^N(S_N \in C) = 0$ whenever $C \cap [E_-, E_+] = \emptyset$, even for finite N. Eqs. (8.25) and (8.2) merely imply this for $N \to \infty$, but it makes (8.25) less surprising than it may appear. For $0 < \beta < \infty$ we write $\beta = 1/T$, $D_q(q_\beta) = -S$, $x = U$, and $\log Z_\beta = -\beta F$. Eq. (8.25) then gives the free energy

$$F = U - TS. \tag{8.28}$$

If in view of Proposition 8.2 we write the second member of the Fenchel duality (7.22) - (7.23) as

$$-\log Z_q(-\beta E) = \inf_{p \in \mathrm{Prob}(A)} \{\beta \langle E \rangle_p + D_q(p)\}, \tag{8.29}$$

where Z_q is defined by (7.21), and notice that $Z_q(-\beta E)$ equals Z_β in (8.23), we conclude that not only the constrained infimum but also the unconstrained one (8.29) is attained at $p = q_\beta$.

Proof of Proposition 8.2. The case $x = E_\pm$ is left to the reader. Assume $E_- < x < E_+$. Since $\langle E \rangle_{q_\beta} \to E_\pm$ as $\beta \to \mp\infty$ by Laplace or directly, the desired value of β for which

$$\langle E \rangle_{q_\beta} = x \tag{8.30}$$

exists by continuity of $\beta \mapsto \langle E \rangle_{q_\beta}$ and the intermediate value theorem. Eq. (8.25) then follows from a simple computation based on (8.23) and (5.12). To prove uniqueness of the minimizer q_β, first note from (5.13) that a possible other minimizer $p \in \mathrm{Prob}(A)$ must satisfy $p \ll q$ and hence $p \ll q_\beta$, so that the Radon–Nikodym derivative $f(a) = p(a)/q_\beta(a)$ (which for finite A is just a quotient) exists. Then, by (5.12), we obtain

$$D_q(p) = D_{q_\beta}(p) + D_q(q_\beta), \tag{8.31}$$

where we used the constraint $\langle E \rangle_p = x$ under which p is supposed to minimize $D_q(p)$, and (8.25). By Proposition 5.1 we conclude that $D_q(p) \geq D_q(q_\beta)$ with equality iff $p = q_\beta$. □

Exercise 52 *Prove* (8.31) *for $E_- < x < E_+$. Also, prove the theorem for $x \notin (E_-, E_+)$.*

We also note that I_q is continuous on (E_-, E_+). If $C \subset (\mathring{C})^- \subset [E_-, E_+]$, then (8.2) implies

$$\lim_{N\to\infty} \frac{1}{N} \log q^N(S_N \in C) = -I_q(C). \tag{8.32}$$

The existence and uniqueness part of the Proposition 8.2 is a special case of the following:[99]

Proposition 8.3 *Let A be a Polish space. If $C \subset \mathrm{Prob}(A)$ is closed and convex and satisfies*

$$D_q(C) = D_q(\mathring{C}) < \infty, \tag{8.33}$$

there is a unique $p \in C$ that minimizes $D_q(p)$ within C. For the conditional probability measures $P_N \in \mathrm{Prob}(\mathrm{Prob}(A))$ defined by $P_N(B) := q^N(L_N \in B \mid L_N \in C)$ (i.e. $P_N = q^N(\cdot \mid C)$) we have

$$\lim_{N\to\infty} P_N = \delta_p \tag{8.34}$$

exponentially fast, in that for any weak nbhd $U \in \mathscr{O}_p(\mathrm{Prob}(\mathrm{Prob}(A)))$ there is $b > 0$ such that

$$p_N(U^c) \leq e^{-bN}. \tag{8.35}$$

[99] We omit the proof; see Dembo & Zeitouni, Theorem 3.3.3, or Rassoul-Agha, §5.3.

Proof of Theorem 8.1. We re-express the average S_N in terms of the empirical measure (7.1):

$$\begin{aligned}\langle E\rangle_{L_N(\sigma)} &= \sum_{a\in A} L_N(\sigma)(a)E(a) = \frac{1}{N}\sum_{a\in A}\sum_{n=0}^{N-1}\delta_{\sigma_n a}E(a) = \frac{1}{N}\sum_{n=0}^{N-1}E(\sigma_n) = \frac{1}{N}\sum_{n=0}^{N-1}X_n(\sigma)\\ &= S_N(\sigma),\end{aligned} \tag{8.36}$$

i.e. $S_N(\sigma)$ is the expectation of E in $L_N(\sigma)\in\mathrm{Prob}(A)$. Hence for every subset $C\subset\mathbb{R}$ we have

$$S_N\in C \qquad \text{iff} \qquad L_N\in B := \{p\in\mathrm{Prob}(A)\mid \langle E\rangle_p\in C\}. \tag{8.37}$$

Similarly to (8.21), we observe that, by basic set theory and the definition of an infimum,

$$\{p\in\mathrm{Prob}(A)\mid\langle E\rangle_p = C\} = \{p\in\mathrm{Prob}(A)\mid\langle E\rangle_p = x, x\in C\}; \tag{8.38}$$

$$\inf\{D_q(p)\mid p\in\mathrm{Prob}(A),\langle E\rangle_p\in C\} = \inf_{x\in C}\{D_q(p)\mid p\in\mathrm{Prob}(A),\langle E\rangle_p = x\}. \tag{8.39}$$

The idea of the proof is that we know from Sanov's theorem how the asymptotics of $q^N(L_N\in B)$ is controlled by the rate function D_q via (7.6) with (7.5). From (8.37) and (8.39) we have:

$$\begin{aligned}D_q(B) &= \inf_{p\in B}\{D_q(p)\} = \inf\{D_q(p)\mid p\in\mathrm{Prob}(A),\langle E\rangle_p\in C\}\\ &= \inf_{x\in C}\{D_q(p)\mid p\in\mathrm{Prob}(A),\langle E\rangle_p = x\} = \inf_{x\in C} I_q(x) = I_q(C).\end{aligned} \tag{8.40}$$

Hence (8.2) follows from (7.6), except for the complication that the equations do not just contain B and C but also their interiors and closures. This calls for some additional topological reasoning. For example, since $p\mapsto\langle E\rangle_p$ is (weakly) continuous, if $B\subset\mathrm{Prob}(A)$ corresponds to $C\subset\mathbb{R}$ via (8.37), then C is open if B is open and hence (8.9) follows from (7.9). The derivation of (8.8) from (7.8) needs some more care and follows from the detailed properties of I_q as stated.

The equivalence between (8.3) and (8.5) is a central feature of large deviation theory and statistical mechanics. It is a special case of *Fenchel–Rockafellar duality*, but we first sketch a direct proof. This is based on the Fenchel transform (7.22) and the Lagrange multiplier method for computing a constrained extremum like (8.3). As in (7.27), for finite A eq. (7.22) simply reads

$$D_q(p) = \sup_{y\in\mathbb{R}^{|A|}}\left\{\sum_a y_a p_a - \log\sum_b q_b e^{y_b}\right\}. \tag{8.41}$$

Furthermore, because of (7.20) we may take the infimum in (8.3) over all $p: A\to\mathbb{R}$ instead of all $p\in\mathrm{Prob}(A)$, i.e. we may forget the conditions $p(a)\geq 0$ and $\sum_a p(a) = 1$. Calling the Lagrange multiplier t, the constrained extremum in (8.3) is therefore found (if it exists) by minimizing

$$(p,t)\mapsto D_q(p) - t\left(\sum_a E_a p_a - x\right) = \sup_{y\in\mathbb{R}^{|A|}}\left\{\sum_a (y_a - tE_a)\cdot p_a - \log\sum_b q_b e^{y_b}\right\} + tx. \tag{8.42}$$

Minimizing with respect to p by naively putting $\partial/\partial p_a$ of the above expression to zero gives $y_a = tE_a$, which clears the first term after the sup. Substituted into the second term, it also gives

$$\sum_b q_b e^{y_b} = \langle e^{tE}\rangle_q. \tag{8.43}$$

Minimizing with respect to t returns the constraint $\sum_a E_a p_a = x$, but if we write this second minimization as $\inf_t$, we see at once that I_q as defined in (8.3) equals I_q as stated in (8.5).

There was some handwaving in the above argument. A real proof of the equality of (8.3) and (8.5) is based on Fenchel–Rockafellar duality, cf. Theorem A.18 and Corollary A.19:[100]

[100] See Borwein & Zhu, Theorem 4.7.1.

Proposition 8.4 *For a Banach space X, let $f : X \to (-\infty,\infty]$ be lsc and convex, and $T : X \to \mathbb{R}^d$ linear and continuous. If $x \in \mathbb{R}^d$ lies in the interior (or the core) of $T(\mathscr{D}_f)$, then*

$$\inf_{p\in X}\{f(p) \mid Tp = x\} = \sup_{t\in\mathbb{R}^d}\{\langle t,x\rangle - f^*(T^*t)\}. \tag{8.44}$$

and the supremum is even a maximum (i.e. it is attained). Here f^* is the Fenchel transform of f.

This applies to (8.3) since as noted before we may replace $p \in \mathrm{Prob}(A)$ by $p : A \to \mathbb{R}$. Hence in Proposition 8.4 we may take $X = \mathbb{R}^{|A|}$, $f = D_q$, $d = 1$, and $T : \mathbb{R}^{|A|} \to \mathbb{R}$ defined by

$$Tp = \langle E\rangle_p = \sum_a E_a p_a. \tag{8.45}$$

Then $f^* : X^* = \mathbb{R}^{|A|} \to (-\infty,\infty]$ is given by $D_q^* = \Pi_q$, see (7.23), and $T^* : \mathbb{R} \to \mathbb{R}^{|A|}$ is given by

$$T^*t = tE, \tag{8.46}$$

because $\langle T^*t, p\rangle := \langle t, Tp\rangle$ gives $\sum_a p_a(T^*t)_a = t\sum_a E_a p_a$, and hence $(T^*t)_a = tE_a$. Note that

$$\Pi_q(tE) = \Pi_q(t), \tag{8.47}$$

with abuse of notation, where the left-hand side is defined in (7.21) whilst the right-hand side comes from (8.6). Thus the equality in (8.44) is almost the equality between (8.3) and (8.5); except that (8.44) is only stated for $x \in \mathrm{int}(T(\mathscr{D}_f))$. Assuming once again that $q_a > 0$ for all $a \in A$ for simplicity, we have $\mathscr{D}_f = \mathrm{Prob}(A)$ and hence

$$T(\mathscr{D}_f) = [E_-, E_+]; \qquad \mathrm{int}(T(\mathscr{D}_f)) = (E_-, E_+). \tag{8.48}$$

- Eq. (8.44) therefore establishes the equality between (8.3) and (8.5) for all $x \in (E_-, E_+)$.

Extending this analysis to $x \notin (E_-, E_+)$ is similar to the case $A = 2$, $E(a) = a$, $q = f$:

- For the boundary points $x = E_\pm$ we first recall that $E_\pm = E(a_\pm)$. If $x = E_-$, the constraint $\langle E\rangle_p = E_-$ can only be satisfied by $p(a_-) = 1$, so that $D_q(p) = -\log q(a_-)$. But this is also the supremum of (8.49) for $x = E_-$, which is reached as $t \to -\infty$. The constraint $\langle E\rangle_p = E_+$ can only be satisfied by $p(a_+) = 1$, so that $D_q(p) = -\log q(a_+)$; which is the supremum of (8.49) for $x = E_+$, but this time as $t \to \infty$.

- For $x > E_+$ or $x < E_-$ we have, using (8.3), $I_q(x) = \inf\emptyset = \infty$. But using (8.5) we also find $I_q(x) = \infty$ by an argument similar to the one between (7.31) and (7.33):

  Exercise 53 *Show that for $x > E_+$ or $x < E_-$ the function*

  $$t \mapsto tx - \Pi_q(t) = tx - \log\sum_a q_a e^{tE_a} \tag{8.49}$$

  has no maximum and *that its supremum is infinite.*

Finally, the general properties of I_q and Π_q easily follow: we have $I_q(x) \geq 0$ by (8.3) and Proposition 5.1, convexity and lsc of I_q follow from (8.5). Convexity and lsc of Π_q will be proved more generally in Theorem 8.5 below, upon which (8.7) follows from Theorem A.10. □

We now turn to a more general (though still one-dimensional) form of Cramér's theorem:

Theorem 8.5 *Let (X_n) be i.i.d. $\mathbb{R}$-valued random variables distributed by $q \in \mathrm{Prob}(\mathbb{R})$ such that*

$$\tilde{Z}_q(t) := \langle e^{tX_1} \rangle_q < \infty \tag{8.50}$$

for each $t \in \mathbb{R}$,[101] *so that also $\mu := \langle X_1 \rangle_q < \infty$. Define the "pressure"*

$$\Pi_q(t) := \log \tilde{Z}_q(t) \tag{8.51}$$

and its full and partial Fenchel transforms $I_q^{(\pm)} : \mathbb{R} \to [0,\infty]$ *(cf. Appendix A) as*

$$I_q(x) = \sup_{t\in\mathbb{R}} \{xt - \Pi_q(t)\}; \qquad I_q^+(x) = \sup_{t\geq 0} \{xt - \Pi_q(t)\}; \qquad I_q^-(x) = \sup_{t\leq 0} \{xt - \Pi_q(t)\}. \tag{8.52}$$

1. *Each $I_q^{(\pm)}$ is convex and lsc, and I_q has a unique minimum $I_q(\mu) = 0$ at $x = \mu$.*
2. *The function Π_q is also convex and lsc (even continuous), and is related to I_q by*

$$\Pi_q(t) = \sup_{x\in\mathbb{R}} \{xt - I_q(x)\}. \tag{8.53}$$

3. *The averages* (1.36) *satisfy an* LDP *that can be expressed in the following equivalent ways:*

$$\lim_{N\to\infty} \frac{1}{N} \log q^N(S_N \geq x) = -I_q^+(x) \qquad (x \in \mathbb{R}); \tag{8.54}$$

$$\lim_{N\to\infty} \frac{1}{N} \log q^N(S_N \leq x) = -I_q^-(x) \qquad (x \in \mathbb{R}); \tag{8.55}$$

$$\lim_{N\to\infty} \frac{1}{N} \log q^N(S_N \geq x) = -I_q(x) \qquad (x \geq \mu); \tag{8.56}$$

$$\lim_{N\to\infty} \frac{1}{N} \log q^N(S_N \leq x) = -I_q(x) \qquad (x \leq \mu). \tag{8.57}$$

The point is that the functions $I_q^{(\pm)}$ are computable from $\Pi_q(t)$. The more general LD property

$$\lim_{N\to\infty} \frac{1}{N} \log q^N(S_N \in B) = -I_q(B), \tag{8.58}$$

for subsets $B \subset \mathbb{R}$ for which $I_q(\mathring{B}) = I_q(B^-)$, cf. (7.7), follows from this, but is not usually stated as part of Cramér's theorem. For example, this condition holds if $B \subset \mathring{D}_{I_q}$ is an open interval, since in that case I_q is continuous on B by Proposition A.6. Eq. (8.58) then follows from the claimed properties of I_q, plus the inference from (8.56) that I_q is increasing for $x \geq \mu$, and similarly, by (8.57) is decreasing for $x \leq \mu$. Theorem 8.5 also implies statements à la (7.8), (7.9), and (7.7).

Proof.[102] We will prove (8.54). The others then follow, as we now show. Jensen for the convex function $x \mapsto e^x$ gives $\tilde{Z}_q(t) \geq e^{t\mu}$, so that for $t \leq 0$ and $x \geq \mu$ one has $\tilde{Z}_q(t) \geq e^{tx}$ and hence $xt - \Pi_q(t) \leq 0$. For $t = 0$ one has $x \cdot 0 - \tilde{Z}_q(0) = 0$. Hence the supremum $\sup_{t\in\mathbb{R}}$ in (8.52) must be reached for $t \geq 0$. This shows that $I_q(x) = I_q^+(x)$ for $x \geq \mu$, which yields (8.56) given (8.54).

Next, changing X_n to $-X_n$ gives

$$-\lim_{N\to\infty} \frac{1}{N} \log q^N(-S_N \geq -x) = \sup_{t\in\mathbb{R}} \{-xt - \Pi_q(-t)\} = \sup_{t\in\mathbb{R}} \{xt - \Pi_q(t)\} = I_q(x), \tag{8.59}$$

[101] This assumption can be relaxed, see Cerf & Petit (2010) and Rassoul-Agha, §2.4, but it simplifies the proof.

[102] There is no simple proof of Cramér's theorem. The one below combines ideas from Cerf & Petit (2010) and Rassoul-Agha, §2.4, and despite its length is felt to be pedagogical since it is largely built from "canonical" arguments that often recur in large deviation theory. See also Den Hollander (2000), §I.3, and Olivieri & Vares (2004).

valid for $-x \geq \langle -X_1 \rangle_q = -\mu$, i.e. $x \leq \mu$. Hence (8.57) follows from (8.56), *with the same rate function*. The derivation of (8.55) from (8.54) is quite similar.

We continue with a proof of the basic properties of I_q that are claimed.

1. Positivity of $I_q^{(\pm)}$ follows from the fact that $\Pi_q(0) = \log 1 = 0$, so that $t \mapsto xt - \Pi_q(t)$ assumes the value zero in any case, so that its supremum over t must be $I_q(x) \geq 0$.

2. Convexity and lsc of the $I_q^{(\pm)}$ follow because these functions are *defined* as (partial) Fenchel transforms, see Definition A.9 and following text (briefly: they are convex and lsc because they are suprema of affine–hence convex–and continuous–hence lsc–functions).

3. Jensen for the convex function $x \mapsto e^x$ gives $\langle e^{tX_1} \rangle_q \geq e^{t\langle X_1 \rangle_q} = e^{t\mu}$ and hence $\Pi_q(t) \geq t\mu$ for all t, so that $\mu t - \Pi_q(t) \leq 0$ and hence $I_q(\mu) = \sup_t \{\mu t - \Pi_q(t)\} \leq 0$. Hence $I_q(\mu) = 0$ by point 1. This local minimum is a global minimum since I is convex, see Proposition A.7. It is also unique: the minimum value $I_q(\mu) = 0$ is assumed for $t = 0$, and if (first) $x > \mu$, then since $\Pi_q'(0) = \mu$ and $t \mapsto \Pi_q(t)$ is continuous (even C^1), for sufficiently small t we have $\Pi_q'(t) < x$. Hence if we define $g_x(t) := xt - \Pi_q(t)$, we have $g_x'(t) = x - \Pi_q'(t) > 0$ and so $g_x(t) > g_x(0) = 0$. Consequently, $I_q(x) = \sup_t g_x(t) > 0$. Similarly, if $x < \mu$, then $\Pi_q'(t) > x$ so that $g_x'(t) < 0$ for small t, so that again $g_x(t) > 0$ and hence $I_q(x) > 0$. All in all, we have

$$I_q(\mu) = 0; \qquad I_q(x) > 0 \quad (x \neq \mu). \tag{8.60}$$

4. The function $t \mapsto \Pi_q(t) := \log \tilde{Z}_q(t)$ is convex and continuous (and hence lsc), and so by Fenchel duality (Theorem A.10), the same is true for I_q *as defined by* sing (8.52).

   Continuity of $\tilde{Z}_q$ and hence of $\log \tilde{Z}_q$ follows from Lebesgue monotone convergence: write

$$\tilde{Z}_q(t) = \int_{\mathbb{R}} dq(x)\, e^{tx}. \tag{8.61}$$

Convexity of $\log \tilde{Z}_q$ follows from Hölder's inequality, as in Exercise 48, but do it again:

Exercise 54 *Prove convexity of* $\log \tilde{Z}_q$.

It remains to prove (8.54), which we do through the following steps:

1. We show that the following limit exists for all $x \in \mathbb{R}$ and defines a convex lsc function:

$$I_q^+ : \mathbb{R} \to [0, \infty]; \qquad I_q^+(x) := -\lim_{N \to \infty} \frac{1}{N} \log q^N(S_N \geq x). \tag{8.62}$$

2. We will separately prove the inequalities

$$\Pi_q(t) \geq \sup_{x \in \mathbb{R}} \{tx - I_q^+(x)\}; \tag{8.63}$$

$$\Pi_q(t) \leq \sup_{x \in \mathbb{R}} \{tx - I_q^+(x)\}, \tag{8.64}$$

   for $t \geq 0$. This gives

$$\Pi_q(t) = \sup_{x \in \mathbb{R}} \{tx - I_q^+(x)\} \quad (t \geq 0). \tag{8.65}$$

3. Corollary A.17, which is necessary because Theorem A.10 does not quite apply, then gives

$$I_q^+(x) = \sup_{t \geq 0} \{tx - \Pi_q(t)\}. \tag{8.66}$$

Combining (8.66) with (8.62), eq. (8.54) follows. In order to derive (8.62), we need a lemma:

Lemma 8.6 *The sequence $a_N = -\log q^N(S_N \geq x)$ is subadditive, i.e. $a_{M+N} \leq a_M + a_N$.*

Proof This eventually follows from the obvious implication, cf. (1.36),

$$\frac{1}{N}\sum_{n=0}^{N-1} X_n \geq x,\ \frac{1}{M}\sum_{n=N}^{N+M-1} X_n \geq x \qquad \Rightarrow \qquad \frac{1}{N+M}\sum_{n=0}^{N+M-1} X_n = S_{N+M} \geq x. \tag{8.67}$$

Since the two events on the left are independent because the X_n are i.i.d., and for the same reason

$$q^{\mathbb{N}}\left(\frac{1}{M}\sum_{n=N}^{N+M-1} X_n \geq x\right) = q^{\mathbb{N}}\left(\frac{1}{M}\sum_{n=0}^{M-1} X_n \geq x\right) = q^{\mathbb{N}}(S_M \geq x), \tag{8.68}$$

eq. (8.67) implies

$$q^{\mathbb{N}}(S_N \geq x)\cdot q^{\mathbb{N}}(S_M \geq x) \leq q^{\mathbb{N}}(S_{N+M} \geq x). \tag{8.69}$$

Since $q^{\mathbb{N}}(S_N \geq x) = q^N(S_N \geq x)$ ect., this gives $a_{M+N} \leq a_M + a_N$, because

$$\begin{aligned}\frac{-\log q^{N+M}(S_{N+M} \geq x)}{N+M} &\leq \frac{-\log q^N(S_N \geq x) - \log q^M(S_M \geq x)}{N+M} \\ &\leq -\frac{1}{N}\log q^N(S_N \geq x) - \frac{1}{M}\log q^M(S_M \geq x).\end{aligned} \tag{8.70}$$

Lemma 2.2 then shows that the limit in (8.62) exists for all $x \in \mathbb{R}$ (possibly ∞) and equals

$$I_q^+(x) = \inf_N \frac{1}{N}\log\left(\frac{1}{q^N(S_N \geq x)}\right). \tag{8.71}$$

In particular, $I_q^+(x) = \infty$ iff $q^N(S_N \geq x) = 0$ for all N; for $N = 1$ this gives (for later use):

$$\mathscr{D}_{I_q^+} = \{x \in \mathbb{R} \mid q(X_1 \geq x) > 0\}. \tag{8.72}$$

Another consequence of (8.71) we will have occasion to use is that for any $N \geq 1$,

$$q^N(S_N \geq x) \leq e^{-NI_q^+(x)}. \tag{8.73}$$

Convexity of $I_q : \mathbb{R} \to [0,\infty]$ follows from the special case

$$I_q^+(\tfrac{1}{2}(x+y)) \leq \tfrac{1}{2}(I_q^+(x) + I_q^+(y)), \tag{8.74}$$

which is proved by the same trick that proved subadditivity; this time, the relevant implication is

$$\frac{1}{N}\sum_{n=0}^{N-1} X_n \geq x,\ \frac{1}{N}\sum_{n=N}^{2N-1} X_n \geq y \qquad \Rightarrow \qquad \frac{1}{2N}\sum_{n=0}^{2N-1} X_n = S_{2N} \geq \tfrac{1}{2}(x+y). \tag{8.75}$$

Since I_q^+ is increasing by (8.62), Lemma A.5 gives convexity of I_q^+.

We now show that I_q^+ is lsc. Since the function $x \mapsto P([x,\infty))$ is usc for any probability measure P on $\mathbb{R}$, for finite N the function $x \mapsto 1/q^N(S_N \geq x)$ is lsc. We would be ready if the infimum in (8.71) were a supremum, cf. Proposition A.8, but life it isn't so simple and we need a direct argument. Eq. (8.71) implies that if $q(X_1 \geq x) > 0$ for all $x \in \mathbb{R}$, then $\mathscr{D}_{I_q^+} = \mathbb{R}$ and we are ready by Proposition A.6. If not, define $x_0 := \inf\{x \in \mathbb{R} \mid q(X_1 \geq x) > 0)\}$. Then:

- either $\mathscr{D}_{I_q^+} = (-\infty, x_0)$ and we are again ready by Proposition A.6;

- or $\mathscr{D}_{I_q^+} = (-\infty, x_0]$ and hence $\mathring{\mathscr{D}}_{I_q^+} = (-\infty, x_0)$. On the latter, I_q^+ is finite, because (8.71) for $N = 1$ implies that $I_q^+(x_0) \leq \log(1/q(X_1 \geq x_0))$, and by assumption $q(X_1 \geq x_0) > 0$. The remark after (8.71) gives $I_q^+(x) = \infty$ for all $x > x_0$, and hence lsc of I_q^+ follows once more.

This completes step 1. We now start step 2 by proving (8.63). This follows from a basic inequality in probability theory, which is called either *Chebyshev's inequality* or *Markov's inequality*:

Lemma 8.7 (Chebyshev–Markov) *For any real-valued random variable X distributed by P, any real-valued non-decreasing function f on (at least) the range of X, and any $s \in \mathbb{R}$ with $f(s) > 0$,*

$$P(X \geq s) \leq \frac{\langle f(X) \rangle_P}{f(s)}. \tag{8.76}$$

Exercise 55 *Prove this.*

The best-known application is the weak law of large numbers, which follows by taking $|X|$ instead of X and $f(x) = x^2$ just defined on $\mathbb{R}^+$, where it is indeed non-decreasing, so that if $\langle X \rangle_P < \infty$,

$$P(|X - \langle X \rangle_P| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}, \tag{8.77}$$

where $\mathrm{Var}(X) := \langle |X - \langle X \rangle_P|^2 \rangle_P = \langle X^2 \rangle - \langle X \rangle_P^2$. For i.i.d. (X_n) we then take S_N for X, and since

$$\mathrm{Var}(S_N) = \mathrm{Var}(X_1)/N, \tag{8.78}$$

for all $\varepsilon > 0$ we obtain $P(|S_N - \mu| \geq \varepsilon) \to 0$ and hence $P(|S_N - \mu| < \varepsilon) \to 1$ as $N \to \infty$, cf. (1.37).

For our proof of (8.63) we use Lemma 8.7 with $f(x) = e^{tx}$ for $t \geq 0$, so that for $s \in \mathbb{R}$,

$$P(X \geq s) \leq e^{-st} \langle e^{tX} \rangle_P, \tag{8.79}$$

sometimes called the *exponential Chebyshev inequality*. Since the (X_n) are i.i.d., we have

$$\langle e^{tNS_N} \rangle_{q^N} = \langle e^{t \sum_{n=0}^{N-1} X_n} \rangle_{q^N} = \langle e^{tX_1} \rangle_q^N. \tag{8.80}$$

Therefore, eqs. (8.50), (8.51), and (8.79) give

$$\Pi_q(t) = \log \langle e^{tX_1} \rangle_q = \frac{1}{N} \log \langle e^{tNS_N} \rangle_{q^N} \geq \frac{1}{N} \log \left(e^{tNx} q^N(S_N \geq x) \right) = tx + \frac{1}{N} \log \left(q^N(S_N) \right), \tag{8.81}$$

where we took $X = NS_N$ and $s = Nx$ in (8.79). From (8.71) we then obtain (8.63).

The converse inequality (8.64) will first be proved for $t = 0$, where it is an equality:

$$\sup_{x \in \mathbb{R}} \{-I_q^+(x)\} = \sup_{x \in \mathbb{R}} \sup_{N \geq 1} \{N^{-1} \log q^N(S_N \geq x)\} = 0 = \Pi_q(0), \tag{8.82}$$

since $q^N(S_N \geq x)$ reaches its supremum 1 as $x \to -\infty$ for any N and log is increasing.

To prove (8.64) for $t > 0$ we rely on another recurrent theme, namely *Laplace's method*:[103]

Lemma 8.8 *Let $I \subset \mathbb{R}$ be a bounded interval. If $S : I \to \mathbb{R}$ is lower semicontinuous, then*

$$\lim_{N \to \infty} \frac{1}{N} \log \int_I dx\, e^{NS(x)} = \sup_{x \in I} S(x). \tag{8.83}$$

[103] See Dorlas, Theorem 19.1, for a nice proof, or Theorem 9.5 and its corollary (9.14) below

For simplicity we assume X_1 is bounded,[104] say $-K \le X_1 \le K$ for some $K > 0$, which gives the same bound for S_N (uniformly in N). First, for $t > 0$ we estimate

$$-tNe^{-tNK} + \langle e^{tNS_N} \rangle \le tN \int_{-K}^{K} dx\, e^{tNx} e^{-NI_q^+(x)}, \tag{8.84}$$

where the averaging brackets were brought inside using Fubini's theorem, we noted that

$$\langle 1_{S_N \ge x} \rangle_{q^N} = q^N(S_N \ge x), \tag{8.85}$$

and then used (8.73). Combining this with the first step in (8.81) gives, with $S(x) := tx - I_q^+(x)$,

$$\Pi_q(t) = \frac{1}{N} \log \langle e^{tNS_N} \rangle \le \frac{1}{N} \log(tN) + \frac{1}{N} \log \left(\int_{-K}^{K} dx\, e^{NS(x)} + e^{-tNK} \right). \tag{8.86}$$

Letting $N \to \infty$ (recalling that $t > 0$) and then $K \to \infty$ and using Lemma 8.8 gives (8.64). □

Exercise 56 *Prove* (8.84).

9 Large deviations: General theory

The theorems of Sanov and Cramér were the first *large deviation principles* (LDP). These involve:

- *A sequence (X_N) of random variables $X_N \in \mathscr{X}$*, where $\mathscr{X}$ is some some regular topological space, and X_N is distributed by $P_N \in \mathrm{Prob}(\mathscr{X})$. This is short for $X_N : \Omega \to \mathscr{X}$ (or more generally: $X_N : \Omega_N \to \mathscr{X}$), where Ω (or Ω_N) carries a probability measure $\mathbb{P}$ (or $\mathbb{P}_N$) so that for (measurable) $B \subset \mathscr{X}$ we then have $P_N(B) = \mathbb{P}(X_N \in B)$ (or $P_N(B) = \mathbb{P}_N(X_N \in B)$.
  - For Sanov, X_N is the empirical measure $L_N : A^{\mathbb{N}} \to \mathrm{Prob}(A)$, see (7.1) and (7.18), and P_N comes from the Bernoulli measure $q^{\mathbb{N}}$ on $A^{\mathbb{N}}$, given some prior $q \in \mathrm{Prob}(A)$.
  - For Cramér, X_N is $S_N : A^{\mathbb{N}} \to \mathbb{R}$, see (1.36) and (8.1), with distribution P_N similarly determined by the Bernoulli measure $q^{\mathbb{N}}$ on $A^{\mathbb{N}}$.
- *A lsc (but not necessarily convex) rate function $I : \mathscr{X} \to [0,\infty]$.*[105] Lower semicontinuity implies that for any $s \in [0,\infty)$ the set $\{x \in \mathscr{X} \mid I(x) \le s\}$ is closed; if it is also compact for all s, then I is called *tight* or *good*. This is normally the case.[106]
  - For Sanov this is the relative entropy (5.12) - (5.13), and generally (7.20). For finite A it is easy to see that $I = D_q$ is tight; in general this requires functional analysis.[107]
  - For Cramér this is the function defined by (8.3), which is also tight.[108]

[104] See Cerf & Petit, page 928, for the unbounded case.

[105] Lower semicontinuity can be imposed without loss of generality: if $I : \mathscr{X} \to [0,\infty]$ satisfies (9.3), then the maximal lsc function $\tilde{I}$ majorized by I, obtained or defined by $\tilde{I}(x) = \liminf_{y \to x} I(y)$, also satisfies (9.3), see Rassoul-Agha, §2.2. Convexity does not always hold, but in our examples it usually does.

[106] $\mathscr{X}$ is usually Polish, in which case compact sets are closed, and hence a tight rate function is automatically lsc.

[107] See the comments preceding Theorem 7.3 for the topology on $\mathrm{Prob}(A)$. For finite A, the space $\mathrm{Prob}(A) \subset A^*$ is compact. For given prior $q \in \mathrm{Prob}(A)$ the subset $\mathscr{D}_{D_q} = \{p \in \mathrm{Prob}(A) \mid p \ll q\}$ is closed in $\mathrm{Prob}(A)$ and hence is also compact: for if $p_n \ll q$, then $q(a) = 0$ enforces $p_n(a) = 0$; if $p_n \to p$ weakly this implies $p(a) = 0$. Furthermore, by Proposition 5.1.2, the function D_q is continuous on its domain $\mathscr{D}_{D_q}$, so that the set $\{p \in \mathrm{Prob}(A) \mid D_q(p) \le s\}$ is closed and hence compact. This is also true in general (i.e. if A is Polish), but with a very technical proof; see Dembo & Zetouni, Lemma 6.2.12. The key result from functional analysis used in the proof is the Eberlein–Smulian theorem.

[108] For finite A this follows from Theorem 9.4 below.

- *Estimates on the asymptotic properties of P_N*, which may be given either in the form:[109]

$$\limsup_{N\to\infty}\frac{1}{N}\log P_N(F)\le -I(F) \qquad (F\subset\mathscr{X}\text{ closed}); \tag{9.1}$$

$$\liminf_{N\to\infty}\frac{1}{N}\log P_N(U)\ge -I(U) \qquad (U\subset\mathscr{X}\text{ open}), \tag{9.2}$$

where for $B\subset\mathscr{X}$ we define $I(B):=\inf_{x\in B}I(x)$; or in the equivalent form, for any $B\subset\mathscr{X}$:

$$-I(\mathring{B})\le\liminf_{N\to\infty}\frac{1}{N}\log P_N(B)\le\limsup_{N\to\infty}\frac{1}{N}\log P_N(B)\le -I(B^-), \tag{9.3}$$

where $\mathring{B}$ and B^- are the interior and closure of B, respectively. Here $P_N(X_N\in B)\equiv P_N(B)$.

Exercise 57 *Show that* (9.1) - (9.2) *and* (9.3) *are equivalent.*

Definition 9.1 *If these bullets apply, we say that $(\mathscr{X},X_N,P_N,I)$, or $(\mathscr{X},P_N,I)$, satisfies an* LDP.

Further properties of the rate function I like convexity of tightness will be mentioned explicitly. If

$$I(\mathring{B})=I(B^-), \tag{9.4}$$

which is a condition on both B and I, eq. (9.3) obviously implies the direct estimate

$$\lim_{N\to\infty}\frac{1}{N}\log P_N(B)=-I(B). \tag{9.5}$$

Proposition 9.2 *The rate function in an* LDP *is unique, and is given by*

$$I(x)=\sup_{U\in\mathscr{O}_x(\mathscr{X})}\{-\liminf_{N\to\infty}\frac{1}{N}\log P_N(U)\}, \tag{9.6}$$

where $\mathscr{O}_x(\mathscr{X})$ is the set of open neighbourhoods of x.

Proof. Since $x\in U$, eq. (9.2) immediately gives

$$I(x)\ge\inf_{y\in U}I(y)=I(U)\ge-\liminf_{N\to\infty}\frac{1}{N}\log P_N(U). \tag{9.7}$$

Taking the supremum over all $U\in\mathscr{O}_x(\mathscr{X})$ gives $\ge$ after $I(x)$ in (9.6). Conversely, take $t<I(x)$ and use Lemma A.4 to find $V_0\in\mathscr{O}_x(\mathscr{X})$ such that $t<I(y)$ for all $y\in V_0$. Since $\mathscr{X}$ is regular one can separate x and the closed set $F=\{x\in\mathscr{X}\mid I(x)\le t\}$ by open sets, say $V_1\ni x$ and $V_2\supset F$, respectively, i.e. $V_1\cap V_2=\emptyset$. Let $W=V_0\cap V_1$. Now $V_2^c=\mathscr{X}\setminus V_2$ is closed and $W\subset V_2^c$, so also $W^-\subset V_2^c$. By construction $I(x)>t$ on V_2^c, and so $I(x)>t$ on W^-. Hence (9.1) gives

$$\begin{aligned}\sup_{U\in\mathscr{O}_x(\mathscr{X})}\{-\liminf_{N\to\infty}\frac{1}{N}\log P_N(U)\}&\ge-\liminf_{N\to\infty}\frac{1}{N}\log P_N(W)\\&\ge-\liminf_{N\to\infty}\frac{1}{N}\log P_N(W^-)\ge-\limsup_{N\to\infty}\frac{1}{N}\log P_N(W^-)\\&\ge I(W^-)=\inf_{x\in W^-}I(x)\ge t.\end{aligned} \tag{9.8}$$

Thus $y\ge t$ is true for all $t<s:=I(x)$. Hence $y\ge s$, which proves $\le$ after $I(x)$ in (9.6). □

[109] See Theorems 7.3 and 8.1 for Sanov and Cramér, respectively.

Proposition 9.3 *If $(\mathscr{X}, X_N, P_N, I)$ satisfies an* LDP *with tight rate function I:*

1. *The zero set $I^{-1}(\{0\}) \subset \mathscr{X}$ is compact and nonempty.*
2. *If $B^- \cap I^{-1}(\{0\}) = \emptyset$ for some measurable set $B \subset \mathscr{X}$, then*

$$\lim_{N\to\infty} P_N(B) = 0. \tag{9.9}$$

Exercise 58 *Prove part 1.*

Proof of part 2. If $B^- \cap I^{-1}(\{0\}) = \emptyset$, then $I(B^-) > 0$: if $\inf_{x\in B^-} I(x) = 0$, then this infimum would equal $\inf_{x\in B^-\cap K_s} I(x)$ for any $s > 0$, and since $B^- \cap K_s$ is compact and I is lsc the infimum is attained at some $x \in B^-$, contradicting the assumption on B. Then (9.3) enforces (9.9). □

To go beyond the two cases we have discussed, which mathematically are based on i.i.d. variables and physically describes non-interacting particles, we now discuss some general techniques and results. One of these, which implicitly was already used in deriving Cramér's theorem from Sanov's, is the powerful *contraction principle*, which creates a new LDP from a given one.[110]

Theorem 9.4 *Let $f : \mathscr{X} \to \mathscr{Y}$ be a continuous map between Hausdorff spaces, and suppose $(\mathscr{X}, X_N, P_N, I)$ satisfies an* LDP *with tight rate function I. Then $(\mathscr{Y}, f\circ X_N, P_N \circ f^{-1}, J)$, with*

$$J(y) := \inf_{x\in\mathscr{X}|f(x)=y} I(x) \;\; (y \in \mathrm{range}(f)); \qquad J(y) = \infty \;\; (y \notin \mathrm{range}(f)), \tag{9.10}$$

satisfies an LDP*, again with a tight rate function.*

Proof. Eqs. (9.1) - (9.2) transfer from (X_N) to $(Y_N = f\circ X_N)$ almost by definition, since continuity of f guarantees that for $B' \subset \mathscr{Y}$ open/closed also $B = f^{-1}(B') \subset \mathscr{X}$ is open/closed. For example,

$$\begin{aligned}\limsup_{N\to\infty} \frac{1}{N}\log(P_N \circ f^{-1}(F')) \le I(f^{-1}(F')) &= \inf_{x\in X|x\in f^{-1}(F')} I(x) = \inf_{x\in X|f(x)\in F'} I(x) \\ &= \inf_{x\in X|f(x)=y, y\in F'} I(x) = \inf_{y\in F'} J(y) = J(F'),\end{aligned} \tag{9.11}$$

where $F' \subset \mathscr{Y}$ is closed, and similarly for $U' \subset \mathscr{Y}$ open. Now, for $s \in \mathbb{R}$, we claim that

$$\{y \in \mathscr{Y} \mid J(y) \le s\} = f(\{x \in \mathscr{X} \mid I(x) \le s\}). \tag{9.12}$$

Assuming this for the moment, the argument of f is a compact set because I is tight. Since the continuous image of a compact set is compact, eq. (9.12) makes the set on the left-hand side compact, so that J is tight and hence also lsc. To prove (9.12) we write it as $A = B$. Then $y \in B$ iff there is $x \in \mathscr{X}$ such that $f(x) = y$ and $I(x) \le s$, whereas $y \in A$ iff $\inf_{x\in\mathscr{X}|f(x)=y} I(x) \le s$. Hence $B \subset A$. For the converse inclusion, first note that only those y contribute to A for which $J(y) < \infty$, so that $f^{-1}(\{y\})$ is non-empty, and closed by continuity of f and the fact that singletons in Hausdorff spaces are closed. Suppose $y \in A$ and $y \notin B$. Then $I(x) > s$ for all $x \in f^{-1}(\{y\})$. But consider

$$K_n := \{x \in f^{-1}(\{y\}) \mid I(x) \le s + 1/n\}, \tag{9.13}$$

which is compact because I is tight and $f^{-1}(\{y\})$ is closed. The assumptions $y \in A$ and $y \notin B$ imply that infinitely many K_n are nonempty. Since the K_n are nested, the intersection of the nonempty

[110] We combine Dembo & Zeitouni, §4.2.1, and Rassoul-Agha, §3.1. The latter also gives a more general result: even if I is not tight $(\mathscr{Y}, f\circ X_N, P_N\circ f^{-1}, \tilde{J})$ satisfies an LDP, where $\tilde{J}$ is the maximal lsc function majorized by J.

K_n is not empty, so there is an $x \in f^{-1}(\{y\})$ with $I(x) \le s$, contradicting $I(x) > s$. Hence $y \in A$ implies $y \in B$, i.e. $A \subset B$, and with the earlier $B \subset A$ we conclude that $A = B$, which is (9.12). □

Let us rehearse how this applies to the theorems of Sanov and Cramér: in the former (Theorem 7.3) we have $\mathscr{X} = \mathrm{Prob}(A)$ and $X_N = L_N$, and in the latter (Theorem 8.1) we have $\mathscr{Y} = \mathbb{R}$ and $X_N = S_N$. For given $E : A \to \mathbb{R}$, the function $f : \mathrm{Prob}(A) \to \mathbb{R}$ is given by $f(p) = \langle E \rangle_p$.

Another way to generate a new LDP from an old one is *Varadhan's theorem*:

Theorem 9.5 *Let $(\mathscr{X}, X_N, P_N, I)$ satisfy an* LDP *with $\mathscr{X}$ Polish and I tight.*

1. *For $F : \mathscr{X} \to \mathbb{R}$ continuous and bounded from above,*[111] *we have "Laplacian" asymptotics*

$$\lim_{N\to\infty} \frac{1}{N} \log \int_{\mathscr{X}} dP_N(x) e^{NF(x)} = \sup_{x\in\mathscr{X}} \{F(x) - I(x)\}. \tag{9.14}$$

2. *Given the (P_N), the following perturbed probability measures on $\mathscr{X}$,*

$$dQ_N(x) = \frac{1}{Z_N} dP_N(x)\, e^{NF(x)} \qquad\qquad Z_N = \int_{\mathscr{X}} dP_N(x) e^{NF(x)} \tag{9.15}$$

satisfy an LDP *on $\mathscr{X}$ with rate function*

$$I_F(x) := I(x) - F(x) - \inf_{y\in\mathscr{X}} \{I(y) - F(y)\} = I(x) - F(x) + \sup_{x\in\mathscr{X}} \{F(x) - I(x)\}. \tag{9.16}$$

if the distribution P_N of $X_N : \Omega_N \to \mathscr{X}$ originates in $\mathbb{P}_N \in \mathrm{Prob}(\Omega_N)$, then by definition

$$\langle e^{NF} \rangle_{P_N} = \int_{\mathscr{X}} dP_N(x)\, e^{NF(x)} = \langle e^{NF(X_N)} \rangle_{\mathbb{P}_N} = \int_{\Omega_N} d\mathbb{P}_N(\omega)\, e^{NF(X_N(\omega))}. \tag{9.17}$$

Note that steepest descent method for fixed probability measures is a special case: if P_N is independent of N we have $I(x) = 0$ and hence (9.14) reproduces the usual Laplacian asymptotics

$$\lim_{N\to\infty} \frac{1}{N} \log \int_{\mathscr{X}} dP(x) e^{NF(x)} = \sup_{x\in\mathscr{X}} \{F(x)\}. \tag{9.18}$$

Proof. Wlog we assume $F \le 0$ (if $F \le K$, replace F by $F - K$). For $i = 1, \ldots, n^2$, define

$$G_i := \left\{ x \in \mathscr{X} \mid -\frac{i}{n} \le F(x) \le -\frac{i-1}{n} \right\}. \tag{9.19}$$

Each G_i is closed, F varies at most by $1/n$ within each G_i, and $F \le -n$ on $G = \mathscr{X} \setminus \cup_i G_i$. Thus

$$\begin{aligned} \int_{\mathscr{X}} dP_N(x) e^{NF(x)} &\le \sum_i P_N(G_i) e^{NF^*(G_i)} + P_N(G) e^{-nN} \le n \max_i \{P_N(G_i) e^{NF^*(G_i)}\} + e^{-nN} \\ &\le (n+1) \max\{e^{-nN}, \max_i \{P_N(G_i) e^{NF^*(G_i)}\}\}, \end{aligned} \tag{9.20}$$

where $F^*(G_i) := \sup_{x\in G_i} F(x)$, analogous to $I(B) \equiv I_*(B) = \inf_{x\in B} I(x)$. Eq. (9.1) then gives

$$\limsup_{N\to\infty} \frac{1}{N} \log \int_{\mathscr{X}} dP_N(x) e^{NF(x)} \le \max\{-n, \sup_{x\in\mathscr{X}} \{F(x) - I(x) + 1/n\}\}. \tag{9.21}$$

[111] Various weaker but more contrived conditions suffice. For example, $\limsup_{N\to\infty} \frac{1}{N} \log \int_{\mathscr{X}} dP_N(x) e^{\gamma NF(x)} < \infty$ for some $\gamma > 1$, or $\lim_{M\to\infty} \limsup_{N\to\infty} \frac{1}{N} \log \int_{F(x)\ge M} dP_N(x) e^{NF(x)} - \infty$. (Rassoul-Agha, §3.2; Dembo & Zeitouni, §4.3).

Exercise 59 *Prove this.*

Letting $n \to \infty$ gives

$$\limsup_{N\to\infty} \frac{1}{N} \log \int_{\mathscr{X}} dP_N(x) e^{NF(x)} \leq \sup_{x\in\mathscr{X}} \{F(x) - I(x)\}. \tag{9.22}$$

To prove a lower bound, take $x_0 \in \mathscr{X}$ and $\delta > 0$ arbitrary, with ensuing open set

$$U_\delta(x_0) := \{x \in \mathscr{X} \mid F(x) > F(x_0) - \delta\}, \tag{9.23}$$

which is nonempty since x_0 lies in it. This time using (9.2), we estimate

$$\liminf_{N\to\infty} \frac{1}{N} \log \int_{\mathscr{X}} dP_N(x) e^{NF(x)} \geq F(x_0) - I(x_0) - \delta. \tag{9.24}$$

Since this true for any $x_0 \in \mathscr{X}$ and $\delta > 0$, we obtain

$$\liminf_{N\to\infty} \frac{1}{N} \log \int_{\mathscr{X}} dP_N(x) e^{NF(x)} \geq \sup_{x\in\mathscr{X}} \{F(x) - I(x)\}, \tag{9.25}$$

which together with (9.22) implies (9.14). A similar proof yields (9.16): the term $I(x) - F(x)$ comes from P_N and $F(x)$, whilst the supremum comes from Z_N, exactly as in (9.14). □

Exercise 60 *Prove* (9.24).

In a stronger version of Varadhan's theorem,[112] the assumption that F is bounded is replaced by the mere existence of some constant $\alpha > 0$ such that

$$\sup_N \left(\int_{\mathscr{X}} dP_N(x) e^{\alpha N F(x)} \right)^{1/N} < \infty. \tag{9.26}$$

This version has a corollary to the effect that under additional assumptions the rate function is a Fenchel transform. Assume that $\mathscr{Y}$ is a (real) topological vector space in separating duality with $\mathscr{X}$, and in (9.14) take $F(x) = \langle y, x\rangle$ for $y \in \mathscr{Y}$, where $\langle \cdot, \cdot \rangle$ is the pairing between $\mathscr{X}$ and $\mathscr{Y}$.

Corollary 9.6 *Let $(\mathscr{X}, X_N, P_N, I)$ satisfy an* LDP *with $\mathscr{X}$ Polish and I convex,*[113] *and assume*

$$\sup_N \left(\int_{\mathscr{X}} dP_N(x) e^{N\langle y,x\rangle} \right)^{1/N} < \infty \tag{9.27}$$

for all $y \in \mathscr{Y}$. Then the pressure $\Pi : \mathscr{Y} \to (-\infty, \infty]$ *defined by*

$$\Pi(y) := \lim_{N\to\infty} \log \left(\left(\int_{\mathscr{X}} dP_N(x) e^{N\langle y,x\rangle} \right)^{1/N} \right) = \lim_{N\to\infty} \frac{1}{N} \log \int_{\mathscr{X}} dP_N(x) e^{N\langle y,x\rangle}, \tag{9.28}$$

exists and is convex and lsc, and one has a Fenchel duality $I = \Pi^*$ and $\Pi = I^*$. This remains true if the LDP *holds on a closed convex subset $\mathscr{C} \subset \mathscr{X}$ and I is extended to $\mathscr{X}$ by $I(x) = \infty$ on $\mathscr{X} \setminus \mathscr{C}$.*

Exercise 61 *Compute the pressure $\Pi(y)$ in out two familiar special cases:*

[112] See e.g. Rassoul-Agha, §3.2.

[113] Recall that our rate functions are lsc by definition.

1. *For Sanov's theorem, taking $\mathscr{Y} = C(A)$ for compact (metrizable) A (or finite A if you like) and $\mathscr{X} = \mathscr{M}(A)$, the space of signed measures on A with subspace $\mathscr{C} = \mathrm{Prob}(A)$.*

2. *For Cramèr's theorem, taking $\mathscr{X} = \mathscr{Y} = \mathbb{R}$.*

In both cases, subsequently rederive the rate functions from the above Corollary.

Bryc's theorem is a converse to Varadhan's theorem. A sequence of probability measures (P_N) on $\mathscr{X}$ is called *exponentially tight* if for each $0 < s < \infty$ there a compact set $C_s \subset \mathscr{X}$ such that

$$\limsup_{N\to\infty} \frac{1}{N} \log P_N(\mathscr{X} \backslash C_s) < -s. \tag{9.29}$$

This suggests that for large N the probability P_N is concentrated on a compact set. It can be shown that if $\mathscr{X}$ is Polish and we have an LDP with tight rate function, then this function is automatically tight (Agha-Rassoud, Theorem 2.21). But Bryc's theorem works in the opposite direction.

Theorem 9.7 *Let (P_N) be an exponentially tight sequence of probability measures on a Polish space $\mathscr{X}$. Suppose that for each $F \in C_b(\mathscr{X})$ the following limit exist:*

$$P(F) := \lim_{N\to\infty} \frac{1}{N} \log \int_{\mathscr{X}} dP_N(x) e^{NF(x)} < \infty. \tag{9.30}$$

Then $(\mathscr{X}, P_N, I)$ satisfies an LDP *with exponentially tight (and hence tight) rate function*

$$I(x) := \sup_{F\in C_b(\mathscr{X})} \{F(x) - P(F)\}. \tag{9.31}$$

We will not use this result and therefore omit the proof; see Rassoul-Agha, §3.3, or Dembo & Zeitouni, §4.4. If F is a linear function (in which case it typically cannot be bounded), eq. (9.31) looks like a Fenchel transform. This situation is covered by the important *Gärtner–Ellis theorem*, which similarly gives existence of an LDP on the basis of the existence of a partition function.[114]

Theorem 9.8 *Let (P_N) be a sequence of probability measures on $\mathscr{X} = \mathbb{R}^d$ such that the limit*

$$\Pi(t) := \lim_{N\to\infty} \frac{1}{N} \log \int_{\mathbb{R}^d} dP_N(x) e^{N\langle t,x\rangle} \tag{9.32}$$

exists and is finite for each $t \in \mathbb{R}^d$ as well as differentiable in t. Then $(\mathbb{R}^d, P_N, I)$, with rate function

$$I(x) = \sup_{t\in\mathbb{R}^d} \{\langle t,x\rangle - \Pi(t)\} \qquad\qquad (I = \Pi^*) \tag{9.33}$$

satisfy an LDP *in which I is convex and tight; in fact, differentiability is only needed for* (9.2).

Once again, if the distribution P_N of $X_N : \Omega_N \to \mathbb{R}^d$ originates in $\mathbb{P}_N \in \mathrm{Prob}(\Omega_N)$, then

$$\int_{\mathbb{R}^d} dP_N(x) e^{Ntx} = \langle e^{N\langle t, X_N\rangle} \rangle_{\mathbb{P}_N}. \tag{9.34}$$

Theorem 9.8 generalizes Cramér: if $d = 1$, $\mathbb{P}_N = q^N$, and $X_N = S_N$ as in Theorem 8.5, then

$$\langle e^{N\langle t, X_N\rangle} \rangle_{\mathbb{P}_N} = \langle e^{tE} \rangle_q^N, \tag{9.35}$$

so that N disappears from (9.32), no limit is needed, and hence $\Pi(t) = \Pi_q(t)$, cf. (8.6).

[114] We only give a version using strong assumptions, following Ellis (1995), Theorem 5.1. See Rassoul-Agha, §12.2, and Dembo & Zeitouni, §2.3 and §4.5.3, and den Hollander (2000), §V.2, for weaker assumptions and complete proofs.

Exercise 62 *Also rederive Sanov's theorem for finite A from Theorem 9.8.*

All this suggests that the proof of Theorem 9.8 is even more difficult than the proof of Theorem 8.5. This is indeed the case, and hence we only give a heuristic sketch of a somewhat convincing argument in $d = 1$, in which we also assume that $\Pi(t)$ is twice differentiable.[115]

First, introduce the probability measures

$$dP_N^t(x) := dP_N(x) \cdot \frac{e^{Ntx}}{\int_{\mathbb{R}^d} dP_N(y) e^{Nty}} \overset{N\to\infty}{\approx} dP_N(x) \cdot e^{N(tx-\Pi(t))}, \tag{9.36}$$

where the alleged approximation is based on the existence of (9.32). Second, using the assumption that Π is differentiable and computing naively (i.e. as a physicist) we find

$$\frac{d\Pi(t)}{dt} = \lim_{N\to\infty} \int dP_N^t(x) x = \lim_{N\to\infty} \langle X_N \rangle_{P_N^t}; \tag{9.37}$$

$$\frac{d^2\Pi(t)}{dt^2} = \lim_{N\to\infty} N(\langle X_N^2 \rangle_{P_N^t} - \langle X_N \rangle_{P_N^t}^2) = \lim_{N\to\infty} N \cdot \mathrm{Var}_{P_N^t}(X_N). \tag{9.38}$$

Therefore, if we also assume the existence of $d^2\Pi(t)/dt$, eq. (9.38) suggests that $\mathrm{Var}_{P_N^t}(X_N) \sim 1/N$, just like for averages of i.i.d. variables, so that the usual proof of the weak law of large numbers (rehearsed after Lemma 8.7) also applies here and yields that for each $\varepsilon > 0$,

$$\lim_{N\to\infty} P_N^t(|X_N - \langle X_N \rangle_{P_N^t}| < \varepsilon) = 1. \tag{9.39}$$

A formal computation (justified by the Radon–Nikodym theorem) therefore gives, for $x_0 \in \mathbb{R}$,

$$\begin{aligned} P_N(X_N \in (x_0-\varepsilon, x_0+\varepsilon)) &= \int_{x_0-\varepsilon}^{x_0+\varepsilon} dP_N(x) = \int_{x_0-\varepsilon}^{x_0+\varepsilon} dP_N^t(x) \frac{dP_N^(x)}{dP_N^t(x)} \approx \int_{x_0-\varepsilon}^{x_0+\varepsilon} dP_N^t(x)\, e^{-N(tx-\Pi(t))} \\ &\approx e^{-N(tx_0-\Pi(t))} P_N^t(X_N \in (x_0-\varepsilon, x_0+\varepsilon)). \end{aligned} \tag{9.40}$$

Now consider $I(x) = \sup_{t\in\mathbb{R}}\{tx - \Pi(t)\}$, cf. (9.33). If $d\Pi(t)/dt = x$ has a solution $t = t_0$, then, since (9.38) shows that $d^2\Pi(t)/dt^2 \geq 0$, the function $t \mapsto tx - \Pi(t)$ has at least a local maximum at $t = t_0$; but since the function in question is concave,[116] this maximum is even global, so that

$$t_0 x - \Pi(t_0) = \sup_{t\in\mathbb{R}}\{tx - \Pi(t)\} = I(x). \tag{9.41}$$

So if we now take $x_0 = d\Pi(t)/dt$ in (9.40), then using (9.37) and (9.39) we obtain

$$\lim_{N\to\infty} \frac{1}{N} \log P_N(X_N \in (x_0-\varepsilon, x_0+\varepsilon)) = -I(x). \tag{9.42}$$

A generous use of (9.6) then shows that $I(x)$ as defined by (9.33) is the rate function of an LDP. In fact, a rigorous version of this argument only proves the lower bound (9.2); the upper bound (9.1) follows from the exponential Chebyshev inequality, analogously to the proof of Cramér's theorem.

The simplest application of the Gärtner–Ellis theorem beyond i.i.d. variables is to Markov chains, see Definition 3.5 and surrounding text.[117] Using the Kolmogorov model, we realize

[115] Partly taken from Touchette (2009), Appendix C. The change of measure argument also occurs in rigorous proofs.

[116] The proof that $t \mapsto \Pi(t)$ is convex is practically the same as in Theorem 8.5; the limit $N \to \infty$ preserves convexity.

[117] Our presentation closely follows Dembo & Zeitouni, §3.1.

a real-valued stochastic process (X_n) with finite state space A via (8.1), given some (injective) "energy" function $E : A \to \mathbb{R}$ (one could also embed $A \subset \mathbb{R}$ straight away). Assume the (X_n) form an irreducible homogeneous Markov chain with corresponding irreducible stochastic matrix P_{ab} giving the transition probabilities, with Perron–Frobenius eigenvalue ρ (see Theorem 3.6). Then

$$P(t)_{ab} := P_{ab} e^{tE(b)} \tag{9.43}$$

is also an irreducible stochastic matrix, which has a unique Perron–Frobenius eigenvalue $\rho(t)$.

Theorem 9.9 *The "average energy" $S_N := \frac{1}{N}\sum_{n=1}^{N-1} X_n$ satisfies an* LDP *with tight rate function*

$$I(x) := \sup_{t\in\mathbb{R}}\{xt - \log\rho(t)\}; \qquad\qquad I = (\log\rho)^*. \tag{9.44}$$

Proof. By Theorem 9.8 we are ready if we can show that

$$\lim_{N\to\infty} \frac{1}{N} \log\langle e^{NtS_N}\rangle_P = \log\rho(t), \tag{9.45}$$

where the expectation value is with respect to the probability measure P on $A^{\mathbb{N}}$ defining the given Markov chain. Using (3.30) we may therefore compute

$$\langle e^{NtS_N}\rangle_P = \sum_{a_0} p(a_0) e^{tE(a_0)} P(t)^{N-1}_{a_0 a_{N-1}}. \tag{9.46}$$

Exercise 63 *Prove this.*

Eq. (9.45) therefore follows from Theorem 3.6, i.e. (3.33). Finally, since again by Theorem 3.6 the special eigenvalue $\rho(t)$ is a non-degenerate root of the characteristic polynomial of $P(t)$ it is clear from (9.43) that $\rho(t)$ and hence $\log\rho(t)$ exists for all $t \in \mathbb{R}$ and is C^1 in t. □

Exercise 64 *Check that Theorem 9.9 implies Theorem 8.1.*

Theorem 9.9 thus being a generalization of Cramér's theorem from i.i.d. variables to certain Markov chains, Sanov's theorem can equally well be generalized in the same direction. For finite A this can be done from Theorem 9.9 (and hence ultimately from the Gärtner–Ellis theorem), which we first need to generalize from $\mathbb{R}$ to $\mathbb{R}^d$; this is easily done with practically the same proof (indeed, Theorem 9.8 was already stated for $\mathbb{R}^d$). Instead of $E : A \to \mathbb{R}$ we now work with $E : A \to \mathbb{R}^d$, we have $x \in \mathbb{R}^d$ and $t \in \mathbb{R}^d$, in (9.43) we replace tE_b by $\langle t, E(b)\rangle$, where $\langle\cdot,\cdot,\rangle$ is the standard inner product in $\mathbb{R}^d$ (rather than some expectation value), and similarly, in (9.44) we replace xt by $\langle x,t\rangle$. We take $d = |A|$ and embed $\mathrm{Prob}(A)$ in $\mathbb{R}^d$ by arbitrarily ordering the elements $a \in A$ as $(a_1,\ldots,a_d)$ and hence identifying $p \in \mathrm{Prob}(A)$ with $(p(a_1),\ldots,p(a_d)) \in \mathbb{R}^d$. Then define

$$E(a) = \delta_a \in \mathrm{Prob}(A) \subset \mathbb{R}^d; \qquad E(a_1) = (1,0,\ldots,0),\ldots,E(a_d) = (0,\ldots,0,1), \tag{9.47}$$

that is, $E(a_i)$ is just the unit vector e_i in $\mathbb{R}^d$; the labeling $a \in A$ has been changed to $i \in \{1,\ldots,d\}$. Hence $X_n : A^{\mathbb{N}} \to \mathbb{R}^d$ is given by $X_n(s) = \delta_{s(n)}$, so that $S_N = L_N$ with $L_N(s) \in \mathbb{R}^d$.

Theorem 9.10 *For finite A, let $(A^{\mathbb{N}}, P)$ with $P \in \mathrm{Prob}(A^{\mathbb{N}})$ and random variables $(s_n : A^{\mathbb{N}} \to A)$ define an an irreducible homogeneous Markov chain with transition probabilities P_{ab}. The empirical measures (L_N) satisfy an* LDP *with tight rate function $I_P : \mathrm{Prob}(A) \to [0,\infty]$ given by*

$$I_P(p) = \sup_{u\gg 0}\left\{\sum_{j=1}^{d} p_j \log\left(\frac{u_j}{(uP)_j}\right)\right\}. \tag{9.48}$$

Here $u \gg 0$ means that $u_i > 0$ for each i, ad we write $p_j = p(a_j)$ as well as $(uP)_j = \sum_i u_i P_{ij}$.

Proof. By Theorem 9.9 and (9.47), which gives $\langle t, E(a_j)\rangle = t_j$, we obtain an LDP for large deviations from this average with tight rate function

$$I_P(p) = \sup_{t\in\mathbb{R}^d}\{\langle p,t\rangle - \log\rho(t)\}, \tag{9.49}$$

where $\rho(t)$ is the Perron–Frobenius eigenvalue of the matrix (where $i,j = 1,\dots,d = |A|$)

$$P(t)_{ij} = P_{ij}e^{t_j}. \tag{9.50}$$

As an exercise, let us recover the rate function $I(p) = D_q(p)$ in Theorem 7.3, cf. (5.12) - (5.13). In the i.i.d. case with prior $q \in \mathrm{Prob}(A)$, here given as $q \in \mathbb{R}^d$ via its components $q_i = q(a_i)$, we have

$$P_{ij} = q_j; \qquad\qquad P(t)_{ij} = q_j e^{t_j}. \tag{9.51}$$

The second matrix has an eigenvector $u(t)$ with components $u_i(t) = q_i e^{t_i}$ (no sum) and eigenvalue

$$\lambda(t) = \sum_i q_i e^{t_i}. \tag{9.52}$$

Since $u \gg 0$ (in the sense that $u_i > 0$ for each i), we can use (3.33) to compute $\rho(t)$, yielding

$$\rho(t) = \lambda(t). \tag{9.53}$$

Our earlier result (7.22), proved around (7.27), then gives $I_P = D_p$.

To avoid confusion we denote I_P as defined in (9.49) by I'_P and hence need to show that $I_P = I'_P$. For arbitrary but fixed p, we take an arbitrary $u \gg 0$ and show that the point $t \in \mathbb{R}^d$ defined by

$$t_j = \log\left(\frac{u_j}{(uP)_j}\right) \tag{9.54}$$

leads to $\rho(t) = 1$ and hence $\log\rho(t) = 0$ (to see that t_j is well defined, a simple corollary of the definition of an irreducible stochastic matrix is that $u \gg 0$ implies $uP \gg 0$). This gives

$$\langle p,t\rangle - \log\rho(t) = \langle p,t\rangle = \sum_{j=1}^d p_j \log\left(\frac{u_j}{(uP)_j}\right). \tag{9.55}$$

Hence (9.49) gives $I'_P(p) \geq \sum_{j=1}^d p_j \log\left(\frac{u_j}{(uP)_j}\right)$, and since u was arbitrary (9.48) implies

$$I'_P(p) \geq I_P(p). \tag{9.56}$$

Finally, $uP(t) = u$, as is easily checked from (9.54), and so using $v = u$ in (3.33) gives $\rho(t) = 1$.

Conversely, fix $t \in \mathbb{R}^d$ and take $u \gg 0$ to be the left eigenvector of $P(t)$ (in Theorem 3.6 u is unfortunately called p). Using (9.48), (9.50), $uP(t) = \rho(t)u$, and $\sum_j p_j = 1$, we find

$$\begin{aligned} I_P(p) &\geq \sum_{j=1}^d p_j \log\left(\frac{u_j}{(uP)_j}\right) = -\sum_{j=1}^d p_j \log\left(\frac{(uP)_j}{u_j}\right) = \sum_l p_j t_j - \sum_{j=1}^d p_j \log\left(\frac{(uP(t))_j}{u_j}\right) \\ &= \langle p,t\rangle - \log\rho(t). \end{aligned} \tag{9.57}$$

Since t was arbitrary, we may take the supremum over all $t \in \mathbb{R}^d$ to obtain $I_P(p) \geq I'_P(p)$. □

10 Applications to statistical physics

In view of the role of entropy in large deviation theory it should be no surprise that (so far) the main applications to physics have been to statistical physics and thermodynamics.[118]

10.1 Back to Boltzmann

To connect with the physics literature, for finite A, let us return to the Boltzmann entropy (5.19) - (5.20), This definition works, but as noted earlier, it requires an approximating sequence $p_N \to p$ with $p_N \in \mathrm{Prob}_N(A)$, which is inconvenient. But this can be avoided if we redefine s_B via

$$s'_B(p|q) := \lim_{\delta\to 0}\lim_{N\to\infty} \frac{\log q^N(T_{N,\delta}(p))}{N}, \tag{10.1}$$

where $T_{N,\delta}(p)$, like (5.33), is an even simpler δ-buffer around $T_N(p)$ as defined by (5.2), given by

$$T_{N,\delta}(p) := \{\sigma \in A^N \mid \|L_N(\sigma) - p\| < \delta\}, \tag{10.2}$$

where, at least for for finite A, any norm on $R^{|A|}$ may be used, such as the the ℓ^1-norm

$$\|f\|_1 := \sum_{a\in A} |f(a)|. \tag{10.3}$$

Instead of (5.10), we then obtain

$$e^{NS(p)-o(N)} \le |T_{N,\delta}(p)| \le e^{N(S(p)+\Delta(\delta))+o(N)}, \tag{10.4}$$

where the notation $\Delta(\delta)$ means that for any $\varepsilon > 0$ there exist $\delta > 0$ such that $\Delta(\delta) < \varepsilon$. From this, one infers that (10.1) in the limit coincides with (5.20), in that

$$s_B(p|q) = s'_B(p|q) = -D(p\|q). \tag{10.5}$$

The derivation of (10.5) is similar to (5.20), where of course (10.4) replaces (5.9). To prove the upper bound in (10.4), we use (5.27), the upper bound in (5.9) or (5.10) with $p \rightsquigarrow q$, and the fact that by continuity of $p \mapsto S(p)$ the bound $\|p-q\|_1 < \delta$ implies $|S(p) - S(q)| < \Delta(\delta)$, so that

$$|T_{N,\delta}(p)| \le e^{N(S(p)+\Delta(\delta))+o(N)}. \tag{10.6}$$

Exercise 65 *Prove this.*

To obtain the lower bound in (10.4), find a sequence (q_N) in $\mathrm{Prob}_N(A)$ with $\|p - q_N\| = O(1/N)$, so that for $N > (1/\delta)$ we obtain, using the lower bound in (5.10),

$$|T_{N,\delta}(p)| = \sum_{q\in \mathrm{Prob}_N(A):\|p-q\|<\delta} |T_{N,\delta}(q)| \ge |T_{N,\delta}(q_N)| \ge e^{NS(q_N)-o(N)}. \tag{10.7}$$

But as above, $|S(q_N) - S(p)| = \Delta(\delta) = \Delta(1/N) = o(1)$, so that $N\Delta(1/N) = o(N)$.

In the same spirit, given an (injective) "energy" function $E : A \to \mathbb{R}$ and associated random variable $S_N : A^{\mathbb{N}} \to \mathbb{R}$ defined by (4.39), we define the *Clausius entropy* s_C relative to a prior $q \in \mathrm{Prob}(A)$ as a function of $u \in \mathbb{R}$ (which in thermodynamics is the energy density) by

$$s_C(u|q) := \lim_{\delta\to 0}\lim_{N\to\infty} \frac{\log q^N(E_{N,\delta}(u))}{N}; \tag{10.8}$$

$$E_{N,\delta}(u) := \left\{\sigma \in A^N \mid |S_N(\sigma) - u| < \delta\right\}. \tag{10.9}$$

[118] See also Ellis (1985, 1995), Pfister (2002), Touchette (2009), Rassoul-Agha (2015), Dorlas (2021), etc.

We can compute this entropy from Theorem 8.1, notably from (8.32), which gives

$$s_C(x|q) = -I_q(x), \tag{10.10}$$

where the rate function I_q is given by (8.3) - (8.4) or, equivalently, by (8.5).

Exercise 66 *Prove this, making a case distinction $x \in (E_-, E_+)$, $x = E_\pm$, and $x \notin [E_-, E_+]$.*

Combined with eqs. (10.5) and (10.10), Theorem 8.1 implies that the Boltzmann entropy (10.1), or equivalently (5.23), and the thermodynamic Clausius entropy (10.8) are related by

$$s_C(x|q) = \sup_{p \in \mathrm{Prob}(A)} \{ s_B(p|q) \mid \langle E \rangle_p = x \}. \tag{10.11}$$

The simplest example is probably a baby model of a *paramagnet*, consisting of $N \to \infty$ frozen non-interacting spin-$\frac{1}{2}$ particles.[119] The dimension of space does not matter. An external magnetic field splits the ground state energy (taken to be zero for simplicity) and hence gives rise to two energy levels $\pm\varepsilon$ per particle. Hence the microstates are $\sigma \in 2^N$. In Theorem 8.1 we take

$$A = \{0,1\}; \qquad q = f; \qquad E(0) = -\varepsilon; \qquad E(1) = \varepsilon, \tag{10.12}$$

i.e. $q(0) = q(1) = \frac{1}{2}$, so that in (8.6) we have

$$\tilde{Z}_f(t) = \tfrac{1}{2}\left(e^{t\varepsilon} + e^{-t\varepsilon}\right) = \cosh(t\varepsilon); \qquad \Pi_f(t) = \log\cosh \mathrm{t}\varepsilon). \tag{10.13}$$

A computation starting from (8.5) quite similar to the one leading to (8.11) gives

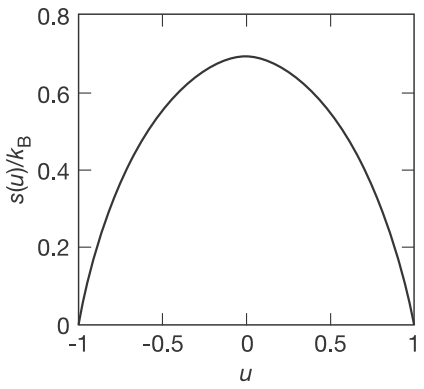
$$\begin{aligned} I_f(x) &= \sup_{t \in \mathbb{R}} \{ tx - \log\cosh(t\varepsilon) \} \\ &= \tfrac{1}{2}(1 + x/\varepsilon)\log(1 + x/\varepsilon) + \tfrac{1}{2}(1 - x/\varepsilon)\log(1 - x/\varepsilon) \qquad (x \in [-\varepsilon, \varepsilon]); \\ &= \infty \qquad\qquad (x \notin [-\varepsilon, \varepsilon]). \end{aligned} \tag{10.14}$$

Exercise 67 *Prove this.*

The corresponding (Clausius) entropy is given by (10.10), but to make it positive one adds a constant $\log|A| = \log 2$ as explained between (5.21) and (5.23); under the flat prior on A this corresponds to counting the number of states in A^N instead of using their probability. Thus we put

$$s(u) := -I_f(u) + \log 2 = -\frac{1+u}{2}\log\left(\frac{1+u}{2}\right) - \frac{1-u}{2}\log\left(\frac{1-u}{2}\right), \tag{10.15}$$

where we switched from x to u, as usual in thermodynamics, and taken $\varepsilon = 1$, so that (10.15) is valid *verbatim* for $u \in (-1,1)$, and by taking limits also for $u = \pm 1$; if $u \notin [-1,1]$, then $s(u) = -\infty$. Here is what the entropy (as a function of energy) looks like:[120]



[119] See Dorlas (2021), Chapters 19–21.

[120] Figure copied from Dorlas (2021), Figure 21.1, page 139.

Bypassing Theorem 8.1, eq. (10.15) may also be derived directly from the (re)definition

$$s(u) := \lim_{N\to\infty} \frac{1}{N} \log |E_N(u_N)|;$$

$$E_N(u_N) = \{\sigma \in A^N \mid S_N(\sigma) = u_N\}; \qquad S_N(\sigma) = \frac{1}{N} \sum_{n=0}^{N-1} E(\sigma_n), \tag{10.16}$$

with $\sigma_n \in \{0,1\}$ and $E(0) = -1$, $E(1) = 1$, and (u_N) is a sequence converging to $u \in (-1,1)$ such that $u_N = S_N(\sigma)$ for some $\sigma \in A^N$, i.e.,

$$u_N = \frac{2M - N}{N} \tag{10.17}$$

for some $M = 0, \ldots, N$ (so that M spins are in state $\sigma = 1$ with energy $+1$ and $N - M$ spins are in $\sigma = 0$ with energy -1). Hence

$$|E_N(u_N)| = \binom{N}{M}. \tag{10.18}$$

To compute $s(u)$ we use the following version of Stirling's formula, see also (5.8):[121]

$$\lim_{N\to\infty} \left(\frac{1}{N} \log(N!) - \log N \right) = -1. \tag{10.19}$$

To do so, we use a clever rearrangement by adding and subtracting the same terms, as follows:

$$\begin{aligned} \frac{1}{N} \log \binom{N}{M} = \frac{1}{N} \log \left(\frac{N!}{M!(N-M)!} \right) &= \frac{1}{N} \log(N!) - \log N - \frac{M}{N} \left(\frac{1}{M} \log(M!) - \log M \right) \\ &- \frac{N-M}{M} \left(\frac{1}{N-M} \log((N-M)!) - \log(N-M) \right) \\ &- \frac{M}{N} \log \left(\frac{M}{N} \right) - \frac{N-M}{N} \log \left(\frac{N-M}{N} \right). \end{aligned} \tag{10.20}$$

Using $M/N = (1+u_N)/2$ and $(N-M)/N = (1-u_N)/2$, as well as (10.19) and $u_N \to u$, we obtain

$$\lim_{N\to\infty} \frac{1}{N} \log \binom{N}{M} = -1 + 0 + 1 - \frac{1+u}{2} \log \left(\frac{1+u}{2} \right) - \frac{1-u}{2} \log \left(\frac{1-u}{2} \right) = s(u), \tag{10.21}$$

which recovers (10.15). See also the discussion after (10.10). Restoring ε then gives

$$s(u) = -\frac{1 + (u/\varepsilon)}{2} \log \left(\frac{1 + (u/\varepsilon)}{2} \right) - \frac{1 - (u/\varepsilon)}{2} \log \left(\frac{1 - (u/\varepsilon)}{2} \right). \tag{10.22}$$

Eq. (10.22) implies all thermodynamics properties of this baby paramagnet. In the absence of heat we have $du = T\,ds$ and hence the (inverse) temperature T is given by $T^{-1} = ds/du$, yielding

$$\frac{1}{T} = \frac{1}{2\varepsilon} \log \left(\frac{\varepsilon - u}{\varepsilon + u} \right); \qquad u = -\varepsilon \tanh \left(\frac{\varepsilon}{T} \right). \tag{10.23}$$

This also allows us to compute the magnetization m as a function of T, since for finite N we have

$$m = \frac{N - M) \cdot \mu + M \cdot (-\mu)}{N} = -\frac{\mu}{\varepsilon} u_N, \tag{10.24}$$

see (10.17), which of course is now given by $u_N = \varepsilon \cdot (2M - N)/N$. Here μ is the magnetic moment of the spins. Combining (10.23) and (10.24) we obtain the well-known formula

$$m = \mu \tanh \left(\frac{\varepsilon}{T} \right). \tag{10.25}$$

[121] We closely follow Dorlas (2021), Chapter 19.

10.2 Non-interacting models

Here is a different approach to the non-interacting systems just studied. We now replace the product probability q^N on A^N (for some $q \in \mathrm{Prob}(A)$) by a probability distribution of the kind

$$\mathbb{P}_{N,\beta}(\sigma) = \frac{1}{Z_N(\beta)} e^{-\beta H_N(\sigma)}; \qquad Z_N(\beta) := \sum_{\sigma \in A^N} e^{-\beta H_N(\sigma)}, \tag{10.26}$$

where $\beta > 0$ and $H_N : A^N \to \mathbb{R}$ is a function defined "uniformly" for each N. What this means depends on the kind of interaction, as explained below. In any case, the "pressure", defined as

$$p(\beta) = \lim_{N\to\infty} \frac{1}{N} \log Z_N(\beta), \tag{10.27}$$

should exist, and with it the "free energy" $f(\beta) := -p(\beta/\beta$, so that, in our earlier sense of $\approx$,

$$Z_N(\beta) \approx e^{Np(\beta)} = e^{-N\beta f(\beta)}. \tag{10.28}$$

In the simplest case, which encompasses both no interaction and so-called mean-field interactions, we take an injective function $E : A \to \mathbb{R}$ as before, with associated average $S_N : A^n \to \mathbb{R}$ defined by (4.39), and a further function $h : [E_-, E_+] \to \mathbb{R}$, where $E_\pm$ were defined in (8.22). This h is typically the restriction of some polynomial $h : \mathbb{R} \to \mathbb{R}$ (abusing notation), and we finally put

$$H_N : A^N \to \mathbb{R} : \qquad H_N = Nh(S_N). \tag{10.29}$$

The simplest nontrivial example is then given by

$$h(x) = x; \qquad h(S_N) = S_N; \qquad H_N(\sigma) = \sum_{n=0}^{N-1} E(\sigma_n). \tag{10.30}$$

In this case the partition function and the pressure are easily computed (do it!) as

$$Z_N(\beta) = \left(\sum_{a\in A} e^{-\beta E(a)} \right)^N; \qquad p(\beta) = \log \left(\sum_{a\in A} e^{-\beta E(a)} \right). \tag{10.31}$$

Indeed, the right-hand side of (10.27) is independent of N and no limit is needed.

We know from Cramér's theorem 8.1 how to compute the probability of the fluctuations of the energy $h(S_N)$ *with respect to* q^N, in which case we also have laws of large numbers to the effect that $h(S_N) \to \langle h(S_N) \rangle_q$ is various ways, at least for linear functions h. But in statistical physics we need the probability of the fluctuations of $h(S_N)$ *with respect to* $\mathbb{P}_{N,\beta}$, and as we shall see, already for quadratic functions h there may not even be a straightforward large of large numbers.

In the linear case (10.30) there are two ways to proceed, of which the first one is a bit easier.

1. We derive a LDP from the Gärtner–Ellis theorem 9.8 with $d = 1$ and for $X_N = S_N$. In (9.32) we substitute (9.34) with $\Omega_N = A^N$ and $\mathbb{P} = \mathbb{P}_{N,\beta}$. In terms of (10.31) this gives

$$\begin{aligned} \Pi(t) &= \lim_{N\to\infty} \frac{1}{N} \log \langle e^{NtX_N} \rangle_{P_N} = \lim_{N\to\infty} \frac{1}{N} \log \langle e^{NtS_N} \rangle_{\mathbb{P}_{N,\beta}} \\ &= \lim_{N\to\infty} \frac{1}{N} \log \left(\frac{1}{Z_N(\beta)} \sum_{\sigma\in A^N} e^{-N(\beta-t)S_N(\sigma)} \right) = p(\beta - t) - p(\beta). \end{aligned} \tag{10.32}$$

   Theorem 9.8 therefore applies: the random variables $X_N = S_N$ satisfy a LDP w.r.t. $\mathbb{P}_{N,\beta}$ with convex (and tight) rate function $I_E : \mathbb{R} \to [0,\infty]$, given as a Fenchel transform

$$I_E(u) = \sup_{t\in\mathbb{R}}\{tu - \Pi(t)\} = \sup_{t\in\mathbb{R}}\{tu - p(\beta - t)\} + p(\beta) = \beta u + p(\beta) - \inf_{t\in\mathbb{R}}\{tu + p(t)\}. \quad (10.33)$$

2. Alternatively, we may combine Cramér's Theorem 8.1 and part 2 of Varadhan's Theorem 9.5. In Varadhan's theorem we take P_N to be the probability measure on $\mathscr{X} = \mathbb{R}$ induced by $S_N : A^N \to \mathbb{R}$ from the flat prior f^N on A^N, where $f(a) = 1/|A|$ for each $a \in A$ as usual and hence $f^N(\sigma) = |A|^{-N}$ for $\sigma \in A^N$ (so that Cramér's Theorem applies), and also put:[122]

$$F(x) = -\beta x. \quad (10.34)$$

Exercise 68 *The aim of this exercise is to establish the desired* LDP *for the S_N w.r.t. $\mathbb{P}_{N,\beta}$.*

1. *Similarly to* (9.34)*, show that the probability measure Q_N on $\mathbb{R}$ defined in* (9.15) *with* (10.34) *equals the distribution of $S_N : A^N \to \mathbb{R}$ induced by the probability measure $\mathbb{P}_{N,\beta}$ on A^N.*

2. *Show that* (9.15) *and* (10.26) *are related by*

$$Z_N = \left(\frac{Z_N(\beta)}{|A|}\right)^N. \quad (10.35)$$

3. *Show that $\Pi_f(t)$ as defined in* (8.6) *with $q = f$ is related to $p(\beta)$ in* (10.27) *by*

$$\Pi_f(t) = p(-t) - \log|A|. \quad (10.36)$$

4. *According to Theorem 8.1 and especially* (10.36)*, the (S_N) satisfy a* LDP *with respect to $\mathbb{P}_N = f^N$ with rate function $I_f(u) = \Pi_f^*(u)$. Combine this with Theorem 9.5.2, especially with* (9.16) *where $I = I_f$, to prove the* LDP *stated above, with I_F in* (9.16) *given by $I_F = I_E$, as in* (10.33)*. Hint: use Fenchel duality* (A.19) *to compute the infimum in* (9.16)*.*

The equilibrium properties of the system are found by minimizing $I_E(u)$, which as we saw in (10.33), or indeed from the general setup, is a function of β. If I_E is convex, as is the case here, it has a unique minimum at some energy $u = u(\beta) = -p'(\beta)$.

Exercise 69 *The simplest physically relevant example is a model for paramagnetism, in which*

$$A = \{0,1\}; \qquad E(0) = -\varepsilon; \qquad E(1) = \varepsilon. \quad (10.37)$$

Show that $p(t) = \log\cosh(t\varepsilon)$, and consequently, that $I_E(u) = I_f(u) + \beta u + p(\beta)$, where

$$I_f(x) = \sup_{t\in\mathbb{R}}\{tx - \log\cosh(t\varepsilon)\} \quad (10.38)$$

$$= \tfrac{1}{2}(1+x/\varepsilon)\log(1+x/\varepsilon) + \tfrac{1}{2}(1-x/\varepsilon)\log(1-x/\varepsilon) \ \ (x \in [-\varepsilon,\varepsilon]);$$

$$= \infty \ \ (x \notin [-\varepsilon,\varepsilon]). \quad (10.39)$$

Computing $u(\beta) = -p'(\beta)$, for which the previous expression is irrelevant, conclude that

$$u(\beta) = -\varepsilon\tanh(\beta\varepsilon). \quad (10.40)$$

Using (10.24), this recovers (10.25). Hence the approaches in §10.1 and 10.2 are consistent; this is a special case of the *equivalence of the micro-canonical ensemble and the canonical ensemble*.

[122] Note that S_N takes values in $[E_-, E_+]$, so that $F : [E_-, E_+] \to \mathbb{R}$, which is bounded. Hence Theorem 9.5 applies.

10.3 The Curie–Weiss model

Mean-field models are defined by taking some polynomial $h : \mathbb{R} \to \mathbb{R}$ in (10.29). They may also be defined on $\mathbb{Z}^d$, but they are insensitive to the dimension d and may as well be defined on $\mathbb{Z}$ or $\mathbb{N}$ straight away. We only treat the case where $A = \{0,1\}$ with $E(0) = -1$ and $E(1) = 1$. It is then easier to work with $A = \{-1,1\}$ and $E(a) = a$. In that case, our S_N in (4.39) just comes down to

$$S_N(\sigma) = \frac{1}{N} \sum_{n=0}^{N-1} \sigma_n. \tag{10.41}$$

The Curie–Weiss model is defined by the function

$$h(x) = -(\tfrac{1}{2}Jx^2 + Bx), \tag{10.42}$$

where $B \in \mathbb{R}$ and $J > 0$ are constants. In other words, we have

$$H_N(\sigma) = -N(\tfrac{1}{2}JS_N(\sigma)^2 + BS_N(\sigma)) = -\frac{J}{2N} \sum_{m,n=0}^{N-1} \sigma_m\sigma_n - B\sum_{n=0}^{N-1} \sigma_n. \tag{10.43}$$

The Curie–Weiss model is an approximation to the Ising model (to be discussed later), in which for any given site $x \in \mathbb{Z}^d$ the pair interaction of σ_x *with its nearest neighbours* is approximated by the an interaction with the "block spin" S_N that averages the value of *all other spins*.

A finite-volume *ground state* is a minimizer $\sigma \in A^N$ of H_N, denoted by $\sigma^{(0)}$ (we suppress the role of N in this notation). [123] For $B = 0$, it is clear that either $\sigma_n^{(0)} = 1$ for all n, or $\sigma_n^{(0)} = -1$ for all n; we denote these by $\sigma_\pm^{(0)}$, respectively. This is an example of *spontaneous symmetry breaking* (SSB), in that the $\mathbb{Z}_s$-symmetry of the the Hamiltonian H_N given by $\sigma \mapsto -\sigma$ is not respected by the ground state; what does happen is that this symmetry maps $\sigma_+^{(0)}$ to $\sigma_-^{(0)}$ and *vice versa*. For $B \neq 0$ the spins align with the sign of B, i.e. the ground state is unique and given by $\sigma_+^{(0)}$ if $B > 0$ and by $\sigma_-^{(0)}$ if $B < 0$. In those cases the Hamiltonian itself already breaks the Z_2 symmetry, which maps S_N to $-S_N$ (this is called *explicit* as opposed to *spontaneous* symmetry breaking).

If more generally we define a *state* of the model as a probability measure on A^N, so that $\mathbb{P}_{N,\beta}$ is a state without further ado and $\sigma_\pm$ should be identified with the point measures $\delta_{\sigma_\pm} \equiv \delta_\pm$, then since A^N is just a finite set, it is easy to see that we have the (weak) limits

$$\lim_{\beta\to\infty} \mathbb{P}_{N,\beta} = \delta_0 := \tfrac{1}{2}(\delta_+ + \delta_-) \qquad (B = 0); \tag{10.44}$$

$$\lim_{\beta\to\infty} \mathbb{P}_{N,\beta} = \delta_+ \qquad (B > 0); \tag{10.45}$$

$$\lim_{\beta\to\infty} \mathbb{P}_{N,\beta} = \delta_- \qquad (B < 0); \tag{10.46}$$

$$\lim_{\beta\to 0} \mathbb{P}_{N,\beta} = f^N \qquad (\text{all cases}). \tag{10.47}$$

[123] The concept of a *state* was arguably first introduced in the context of thermodynamics (by Sadi Carnot),[124] and generally denotes a specification of the "relevant" facts of the matter about some physical system at some given time (and hence the concept of a state is predicated on the concept of "now", mistaken as this may well be). These facts ideally give a complete description of the system, such as the momentary positions and velocities (or momenta) of all particles in a gas; we earlier used the term *microstates* for this. An example is the exact configuration $\sigma \in A^N$. As in the main text, a special case of a microstate is a *ground state* for some given Hamiltonian H_N; this is just some σ that minimizes H_N (and as such may or may not be unique). For the same (lattice) systems, macrostates are for example the empirical measures L_N or the averages S_N (of spin or energy, etc.); in thermodynamics one typically chooses just a small number of parameters, such as the pressure, volume, and temperature of a gas. As a compromise between microstates and macrostates, Boltzmann and Gibbs saw states as *probability measures* on for example A^N (and more generally on the set of microstates of the system). And this, of course, is what the $P_{N,\beta}$ are examples of. Microstates $\sigma \in A^N$ are actually special cases of probability measures on A^N, since the former defines a point measure δ_σ.

Here f is the flat prior on $\{-1,1\}$. Neither the zero temperature limit state (10.44) nor its infinite T counterpart (10.44) breaks the $\mathbb{Z}_2$ symmetry; as a limit of the symmetric states $\mathbb{P}_{N,\beta}$ the former could not do so, and it is remarkable that for $h \neq 0$ the latter even *restores* the symmetry!

What happens at $0 < \beta < \infty$ and at $N \to \infty$? To answer this, we note that S_N is an *order parameter* for the $\mathbb{Z}_2$ symmetry, in that its expectation value vanishes in symmetric states and is non-zero in asymmetric states. For example,

$$\langle S_N \rangle_{\delta_\pm} = \pm 1; \qquad \langle S_N \rangle_{\delta_0} = 0; \qquad \langle S_N \rangle_{P_{\Lambda,\beta}} = 0. \tag{10.48}$$

So we study the random variables (S_N) as $N \to \infty$, not even knowing if a law of large numbers applies! Fortunately, this is a clean application of Theorem 9.5.2, which gives:

Proposition 10.1 *The random variables (S_N) satisfy an* LDP *w.r.t. the distribution induced by $\mathbb{P}_{N,\beta}$ with Hamiltonian* (10.43)*, with lsc (and tight)* but not necessarily convex *rate function*

$$I_{CW}(x) = I_0(x) + \beta h(x) - C \ \ (x \in [-1,1]); \qquad I_{CW}(x) = \infty \ \ (x \notin [-1,1]) \tag{10.49}$$

$$I_0(x) = \tfrac{1}{2}((1-x)\log(1-x) + (1+x)\log(1+x)); \qquad C = \inf_{y \in \mathbb{R}} \{I_0(y) + \beta h(y)\}. \tag{10.50}$$

Consequently, there are three cases, where Q_N is the probability measure of S_N on its range $[-1,1] \subset \mathbb{R}$ induced by the probability distribution $\mathbb{P}_{N,\beta}$ on A^N (cf. Theorem 9.5):[125]

- *$B = 0$ and $\beta \leq J^{-1}$ (i.e. $T \geq J$): I_{CW} has a unique minimum $I_{CW}(x_0) = 0$ at $x_0 = 0$, and*

$$\lim_{N \to \infty} Q_N = \delta_0. \tag{10.51}$$

  this is equivalent to $S_N \to 0$ as in the weak law of large numbers, i.e., for all $\varepsilon > 0$,

$$\lim_{N \to \infty} Q_N(S_N \in (-\varepsilon, \varepsilon)) = 1. \tag{10.52}$$

- *$B = 0$ and $\beta > J^{-1}$ (i.e. $T < J$): the model has a phase transitions in the sense that the rate function I_{CW} has two degenerate minima $I_{CW}(x_\pm) = 0$ at $x_\pm = m_\pm(\beta) \neq 0$ and hence*

$$\lim_{N \to \infty} Q_N = \tfrac{1}{2}(\delta_{m_+(\beta)} + \delta_{m_+(\beta)}). \tag{10.53}$$

- *$B \neq 0$ and arbitrary β and J: same as the first case, except that now the unique minimum $I_{CW}(x_0) = 0$ lies at some at $x_0 = m(\beta, B) \neq 0$, so that $\lim_{N\to\infty} Q_N = \delta_{m(\beta,B)}$.*

Proof. We see from (10.84) and (10.43) that S_N is distributed according to Q_N as defined in (9.15), with P_N the flat prior on $\{-1,1\}^N$ and $F(x) = -\frac{1}{2}Jx^2 - Bx$. Therefore, taking into account the switch from $A = \{0,1\}$ to $A = \{-1,1\}$, from (8.10) - (8.11) and (9.16) we obtain (10.49) - (10.50). Tightness of the rate function I_{CW} is clear from the fact that I_0 is lsc, F is even continuous, and the domain $[-1,1]$ of I_0 and hence of I_{CW} is compact (so that its closed subsets are compact). It is easiest to find the following results by plotting I_0 and F, and using the fact that for $x \in (-1,1)$,

$$I'_{CW}(x) = \tfrac{1}{2}\log\left(\frac{1+x}{1-x}\right) - \beta(Jx+B); \qquad I''_{CW}(x) = \frac{1}{1-x^2} - \beta J, \tag{10.54}$$

so that $I'_{CW}(x) = 0$ iff one of the most famous equations in mean field theory holds, namely

$$x = \tanh(\beta(Jx+B)). \tag{10.55}$$

[125] The limits (10.51) and (10.53) are taken in the weak topology of $\mathrm{Prob}(\mathbb{R})$ w.r.t. the Borel structure. We state without proof that for $\beta \leq J^{-1}$ we also have $S_N \to 0$ strongly with respect to the Gibbs measure on $A^{\mathbb{N}}$ induced by the finite-size probability measures $\mathbb{P}_{N,\beta}$. On the other hand, for $\beta > J^{-1}$ there is neither a weak nor a strong LLN!

- For $0 < \beta J \leq 1$ and any $B \in \mathbb{R}$, eq. (10.55) has a unique solution (which is $x = 0$ for $B = 0$ and whose sign otherwise equals the sign of B), which minimizes I_{CW}. See figure.[126]
- For $\beta J > 1$ and $B = 0$ one finds two solution $x_\pm$ of (10.55), related by $x_- = -x_+$, at each of which $I''_{CW}(x_\pm) > 0$, and hence I_{CW} has two degenerate minima (and no maxima).
- For $\beta J > 1$ and $B \neq 0$ there may be up to three solutions of (10.55); the number depends on the relative size of J and B; e.g. for $B \ll J$ there are three zeros. But only the one with the same sign as B (which always exists) corresponds to an absolute minimum of I_{CW}.

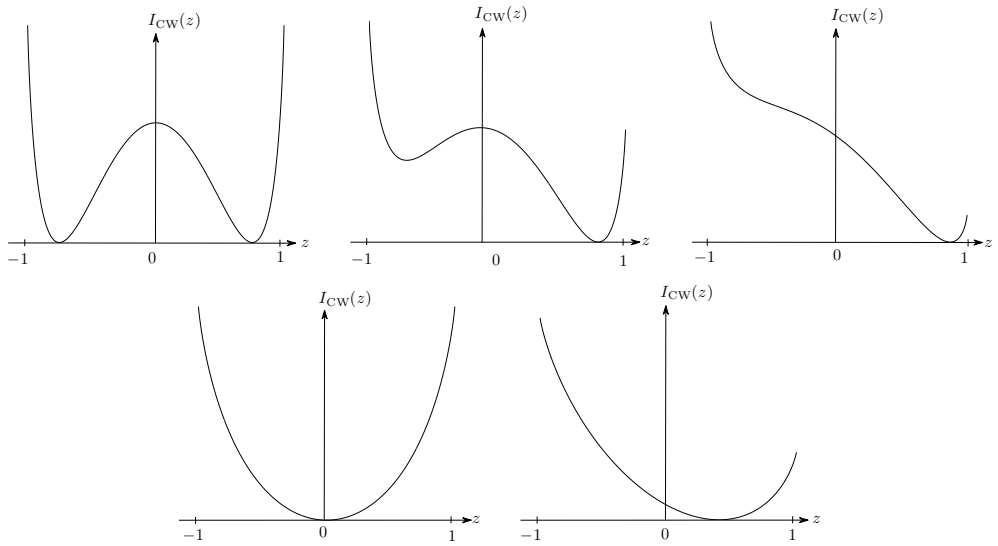


Figure 3.1. The Curie-Weiss rate function. Top plots have $\beta > J^{-1}$ while bottom plots have $\beta \leq J^{-1}$. Top left to right: $h = 0$, $0 < h < h_0(J,\beta)$, and $h > h_0(J,\beta)$. Bottom left to right, $h = 0$ and $h > 0$. The case $h < 0$ is symmetric to that of $h > 0$.

Exercise 70 *Prove these properties.*

The remaining claims then follow from the LDP. In the first case (i.e. $B \neq 0$ etc.), eq. (9.1) gives

$$\limsup_{N\to\infty} \frac{1}{N} \log Q_N(S_N \notin (m(\beta,B)-\varepsilon, m(\beta,B)+\varepsilon)) \leq -I_0((-\infty, m(\beta,B)-\varepsilon] \cup [m(\beta,B)+\varepsilon, \infty]) < 0, \tag{10.56}$$

since $I_{CW}(m(\beta,B) = 0$ is the unique minimum of $I_{CW} \geq 0$ and I_{CW} is continuous on $[-1,1]$ (since I_0 and F both are). Hence $Q_N(S_N \notin (m(\beta,B)-\varepsilon, m(\beta,B)+\varepsilon)) \to 0$, which implies (10.52). The equivalence between (10.52) and (10.51) follows from the Portmanteau theorem.[127]

In the second case (i.e. $B = 0$ and $\beta J > 1$), similar reasoning from (9.1) gives

$$\lim_{N\to\infty} Q_N(S_N \in ((m_+(\beta)-\varepsilon, m_+(\beta)+\varepsilon) \cup (m_-(\beta)-\varepsilon, m_-(\beta)+\varepsilon))) = 1. \tag{10.57}$$

[126] Copied from Rassoul-Agha, page 47, Figure 3.1.

[127] See Definition 4.8. This theorem applies because clearly $\partial((m(\beta,B)-\varepsilon, m(\beta,B)+\varepsilon)) = \{m(\beta,B)-\varepsilon, m(\beta,B)+\varepsilon\}$ and since $\varepsilon > 0$ we have $\delta_{m(\beta,B)}(\{m(\beta,B)-\varepsilon, m(\beta,B)+\varepsilon\}) = 0$.

The Portmanteau theorem then gives weak convergence of Q_N to some convex combination of the point measures $\delta_{m_\pm(\beta)}$; upon which the Z_2 symmetry of Q_N forces (10.53). □

Note that the rate function I_{CW} is *not convex* if $J\beta > 1$, i.e., when there is a phase transition.

Proposition 10.1 describes fluctuations in the average magnetization S_N, which is enough to determine the phase structure of the model, as stated. It is also interesting (and easy) to study energy fluctuations in the Curie–Weiss model. As both a step towards this and as a goal in itself, we note that in our model the pressure defined in (10.27) and hence the free energy exist and are computable via Theorem 9.5. Writing $s_0 = -I_0$, cf. (10.50), we obtain

$$p(\beta) := \lim_{N\to\infty} \frac{1}{N} \log Z_N(\beta) = - \inf_{x\in[-1,1]} \{I_0(x) + \beta h(x)\} = \sup_{x\in[-1,1]} \{s_0(x) - \beta h(x)\}, \tag{10.58}$$

where, according to (10.31), Z_N is given by

$$Z_N(\beta) = \sum_{\sigma\in A^N} e^{-\beta N h(S_N)}. \tag{10.59}$$

Theorem 9.4 now applies, because in our model the energy is a function of the spins whose rate function we know. Since $p(\beta)$ is minus the constant C in (10.50), see also (9.16), we obtain:

Proposition 10.2 *The random variables (h_N) satisfy an* LDP *as $N \to \infty$, with tight rate function*

$$I_E(u) = \inf_{x\in[-1,1]} \{I_0(x) + \beta h(x) \mid h(x) = u\} + p(\beta). \tag{10.60}$$

Hence for the corresponding entropy $s_E(u) := -I_E(u)$ we obtain

$$s_E(u) = \sup_{x\in[-1,1]} \{s_0(x) \mid h(x) = u\} - (\beta u + p(\beta)). \tag{10.61}$$

As in the previous section, we may alternatively compute I_E via Theorem 9.8; this once agin gives (10.33), where this time $p(\beta)$ is given by (10.58).

Exercise 71 *Carry this out and show that the rate function computed from* (9.16) *equals the one obtained from* (9.33)*; in other words, that* (10.33) *with* (10.58) *coincides with* (10.60) *with* (10.50)*.*

10.4 Local Gibbs measures

Implicitly, the two sections above were our first serious encounter with the *canonical ensemble* in classical statistical mechanics, which gives a systematic way to describe equilibrium states of classical spin systems (and also of continuous systems, which will not be treated in these notes) and the associated fluctuations of important physical quantities such as the energy. The mathematical formalism for this is provided by the theory of *Gibbs* measures, which (eventually on the basis of work by Gibbs around 1900) was developed in the 1960s by Dobrushin, Lanford, Ruelle, and others.[128] The following two aspects of this theory may be separated:

1. The definition of equilibrium states in infinite volume, based on the assumption that in finite volume such states are described by the canonical ensemble.

2. The study of large deviations of (notably) the energy in such states (which replace Bernoulli measures in this respect), and the identification of the role of entropy and free energy.

[128] See for example Simon (1993), Keller (1998), Ruelle (2004), and Georgii (2011); the latter also describes the link with large deviations, as do Rassoul-Agha & Seppäläinen (2015). See Georgii (1993) for a turbo treatment.

We start with the first point. We often generalize $\mathbb{N}$ to $\mathbb{Z}^d$ (where $d = 1,2,3,\ldots$) and in what follows $\Lambda \subset \mathbb{Z}^d$ *is always finite*. For our introductory treatment it is even enough to take one of:

$$\mathbb{N} \supset \Lambda = \Lambda_N := \{0,\ldots,N-1\}, \tag{10.62}$$

$$\mathbb{Z}^d \supset \Lambda = \Lambda_N := \{x = (x_1,\ldots,x_d) \in \mathbb{Z}^d \mid |x_i| \leq N, i = 1,\ldots,d\}, \tag{10.63}$$

so that the number of points $|\Lambda_N|$ in Λ_N is given by $|\Lambda_N| = N$ in the first case and

$$|\Lambda_N| = (2N+1)^d \tag{10.64}$$

in the second. For $\mathbb{Z}^d$, all large deviation theory so far then remains valid if we replace N by $|\Lambda_N|$ whenever some quantity is scaled by N or $1/N$. With A a finite set, as before, we often write

$$\Omega := A^{\mathbb{Z}^d}; \qquad\qquad \Omega_\Lambda := A^\Lambda, \tag{10.65}$$

and similarly $\Omega = A^{\mathbb{N}}$. Hence Ω_Λ is finite and is equipped with the σ-algebra $\mathscr{P}(\Omega_\Lambda)$. Measure theory on Ω is based on the *cylindrical* σ*-algebra* $\mathscr{F}$. We give the definition for $\mathbb{Z}^d$, but the adaptation to $\mathbb{N}$ should be obvious. Recall that $\omega_x : \Omega \to A$ is the evaluation map $s \mapsto s(x)$.

Definition 10.3 *Let A be a finite set.*[129]

1. *$\mathscr{F}$ is the smallest σ-algebra Ω for which the evaluation maps ω_x are measurable (with respect to the maximal σ-algebra $\mathscr{P}(A)$ on A) for all $x \in \mathbb{Z}^d$.*
2. *For* finite *$\Lambda \subset \mathbb{Z}^d$, $\mathscr{F}_\Lambda$ is the smallest σ-algebra for which ω_x is measurable for all $x \in \Lambda$.*
3. *For* finite *$\Lambda \subset \mathbb{Z}^d$, $\mathscr{T}_\Lambda$ is the smallest σ-algebra making all maps ω_x with $x \notin \Lambda$ measurable.*
4. *The* tail σ-algebra *$\mathscr{T}$ is defined by $\mathscr{T} := \bigcap_\Lambda \mathscr{T}_\Lambda$.*

Equivalently, the "local" σ-algebra $\mathscr{F}_\Lambda$ consist of all cylinder sets

$$[\Delta]_\Lambda := \{\omega \in \Omega \mid \omega_{|\Lambda} \in \Delta\} \qquad\qquad (\Delta \subset A^\Lambda), \tag{10.66}$$

where $\omega_{|\Lambda} : \Lambda \to A$ is the restriction of $\omega : \mathbb{Z}^d \to A$ to Λ. The the "quasi-local" σ-algebra $\mathscr{F}$ is generated by $\bigcup_\Lambda \mathscr{F}_\Lambda$, where Λ runs over all finite subsets of $\mathbb{Z}^d$. For finite A we have a bijection

$$\mathscr{P}(A^\Lambda) \cong \mathscr{F}_\Lambda \qquad\qquad \Delta \leftrightarrow [\omega_{|\Lambda} \in \Delta]. \tag{10.67}$$

Hence $\mathscr{F}_\Lambda$ is a finite σ-algebra, generated by "atoms" (i.e. its smallest mutually disjoint elements)

$$[\sigma] := \{\omega \in \Omega \mid \omega_{|\Lambda} = \sigma\}; \qquad\qquad (\sigma \in A^\Lambda), \tag{10.68}$$

in the sense that $\mathscr{F}_\Lambda$ consists of all unions and intersections of these atoms. A function $f : \Omega \to \mathbb{R}$ is $\mathscr{F}_\Lambda$-measurable iff it is constant on each atom $[\sigma]$, which means that $f(\omega)$ is independent of the values of $\omega : \mathbb{Z}^d \to A$ outside Λ. In that case, for any (probability) measure on $\mathscr{F}_\Lambda$ we have[130]

$$\int dP\, f = \sum_{\sigma \in A^\Lambda} P([\sigma]) f(\sigma), \tag{10.69}$$

[129] The definition is the same for Polish A, except that $\mathscr{P}(A)$ below is replaced by the Borel σ-algebra on A.

[130] This follows from integration theory. An integral $\int_\Omega dP(\omega)\, f(\omega)$ for some Σ-measurable function $f : \Omega \to [0,\infty]$ is defined as the supremum over all finite sums $\sum_i c_i P(B_i)$, where $0 \leq \sum c_i 1_{B_i} \leq f$ and $B_i \in \Sigma$, which may be taken to be disjoint. If Σ is finite, then it contains disjoint atoms A_i whose union is Ω. In that case $f = \sum_i c_i 1_{A_1}$ exactly and hence $\int dP\, f = \sum_i c_i P(A_i)$, where c_i equals the constant value of f on A_i. Note that any finite σ-algebra is a power set.

where we slightly incorrectly write $f(\sigma)$ for the constant value $f(\omega)$ at any $\omega \in [\sigma]$. Then:

A given physical theory is then specified by a family of (*fixed*) local *Hamiltonians*

$$H_\Lambda : A^\Lambda \to \mathbb{R}. \tag{10.70}$$

It is desirable to separate Λ from the "theory" that specifies the dynamics. This can indeed be done in two cases, of which we already discussed mean-field theories. The second case is meant to describe local interactions, as follows. A systematic way to coherently write down a large class of Hamiltonians H_Λ for different regions Λ comes from the concept of a *potential*:[131]

Definition 10.4 *A* potential $\Phi = (\Phi_M)_M$ *is a set of $\mathscr{F}_M$-measurable functions*

$$\Phi_M : \Omega \to \mathbb{R} \tag{10.71}$$

indexed by the set of all finite *subsets $M \subset \mathbb{Z}^d$, such that:*

1. *Φ_M only depends on $s_{|M}$ (so that we may also write $\Phi_M : A^M \to \mathbb{R}$).*
2. *Φ is translation-invariant in the sense that $\Phi_{M+x} = \Phi_M \circ \theta_x$ for each $x \in \mathbb{Z}^d$, where*

$$\theta_x : \mathbb{Z}^d \to \mathbb{Z}^d; \qquad \theta_x(y) = x + y. \tag{10.72}$$

3. *Φ satisfies the summability condition*

$$\|\Phi\| := \sum_{0 \in M \subset \mathbb{Z}^d} \sup_{\sigma \in \Omega} |\Phi_M(\sigma)| < \infty. \tag{10.73}$$

In many realistic models (like the Ising model) the sum in (10.73) only has a finite number of terms, so that the condition is automatically satisfied. Mean-field models have no potential.

In the simple case of *free boundary conditions*, the local Hamiltonians are then given by

$$H_\Lambda : A^\Lambda \to \mathbb{R}; \qquad H_\Lambda \equiv H_\Lambda^{\text{free}} = \sum_{M \subset \Lambda} \Phi_M. \tag{10.74}$$

Other boundary conditions will be studied later. In view of (10.71) it would be more precise to write $H_\Lambda : \Omega \to \mathbb{R}$ in (10.74), too, and then observe that because of condition (a) in Definition 10.4 indeed H_Λ only depends on $s_{|\Lambda}$. The archetypical example is the *Ising model*, in which the spins take values in $A = \{-1, 1\}$ and the potential is given by

$$\Phi_{\{x\}}(\sigma) = -B\sigma_x; \quad \Phi_{\{x,y\}}(\sigma) = -J\sigma_x\sigma_y \text{ (if } \|x-y\|_1 = 1); \quad \Phi_M = 0 \text{ (otherwise)}, \tag{10.75}$$

where $B \in \mathbb{R}$ (representing a possible external magnetic field) and $J > 0$ (in the ferromagnetic case) are constants, and $\|x-y\| = 1$ means that x and y are "nearest" neighbours, since the $\|\cdot\|_1$ norm on $\mathbb{R}^d \supset \mathbb{Z}^d$ is used (so "diagonal" neighbours are excluded). Thus

$$H_\Lambda = -J \sum_{x,y \in \Lambda, \|x-y\|_1 = 1} \sigma_x\sigma_y - B \sum_{x \in \Lambda} \sigma_x. \tag{10.76}$$

In statistical mechanics these (*fixed*) Hamiltonians are combined with a (*variable*) *temperature*

$$k_B T = \beta^{-1}, \tag{10.77}$$

[131] We (unusually) allow the empty set $M = \emptyset$, so that constants may be added to the Hamiltonian via $\Phi_\emptyset$.

where k_B is Boltzmann's constant taken to be unity in what follows, such that the Bernoulli measures q^Λ on A^Λ (and in the limit on $A^{\mathbb{N}}$ etc.) are replaced by the *local Gibbs measures*

$$P_{\Lambda,q,\beta}(\{\sigma\}) := \frac{1}{Z_\Lambda(\beta)} q^\Lambda(\sigma) e^{-\beta H_\Lambda(\sigma)} \ \ (\sigma \in A^\Lambda); \tag{10.78}$$

$$Z_{\Lambda,q}(\beta) := \sum_{\sigma \in A^\Lambda} q^\Lambda(\sigma) e^{-\beta H_\Lambda(\sigma)}. \tag{10.79}$$

The resemblance between (10.84) and Proposition 8.2, especially eq. (8.23), is striking.[132] We also introduce the global *pressure* $\Pi_{\Lambda,q}(\beta)$ and *free energy* $F_{\Lambda,q}(\beta)$ by

$$\Pi_{\Lambda,q}(\beta) := \log Z_{\Lambda,q}(\beta); \qquad\qquad F_{\Lambda,q}(\beta) := -\beta^{-1}\Pi_{\Lambda,q}(\beta), \tag{10.80}$$

so that

$$Z_{\Lambda,q}(\beta) = e^{-\beta F_{\Lambda,q}(\beta)}. \tag{10.81}$$

In case of a flat prior $q = f$ on A (i.e. $f(a) = 1/|A|$ for each $a \in A$), also each microstate $\sigma \in \Omega_\Lambda$ is *a priori* equiprobable, with probability

$$f^\Lambda(\sigma) = 1/|\Omega_\Lambda| = |A|^{-|\Lambda|}. \tag{10.82}$$

The Gibbs measure (10.79) modifies this equiprobability via the Boltzmann factor $e^{-\beta H_\Lambda(\sigma)}$, but of course at $\beta = 0$, i.e. $T = \infty$, we recover (10.82). Thus (high) temperature introduces (high) randomness; we will analyze this idea in more detail later on in connection with chaos. It is customary to omit the factor f^Λ in the above expressions, so that, writing $P_{\Lambda,q,\beta} = P_{\Lambda,\beta}$ etc.,

$$P_{\Lambda,\beta}(\{\sigma\}) = \frac{1}{Z_\Lambda(\beta)} e^{-\beta H_\Lambda(\sigma)}; \tag{10.83}$$

$$Z_\Lambda(\beta) = \sum_{\sigma \in A^\Lambda} e^{-\beta H_\Lambda(\sigma)} = e^{\Pi_\Lambda(\beta)} = e^{-\beta F_\Lambda(\beta)} \tag{10.84}$$

We recall that the relative entropy and the Shannon entropy are defined by

$$D_Q(P) = \sum_{\sigma \in \Omega_\Lambda} P(\sigma) \log\left(\frac{P(\sigma)}{Q(\sigma)}\right); \tag{10.85}$$

$$S_\Lambda(P) = -\sum_{\sigma \in \Omega_\Lambda} P(\sigma) \log P(\sigma), \tag{10.86}$$

where Ω_Λ is finite! For the flat prior $Q = f^\Lambda$ we infer from (10.82) that these are related by

$$D_{f^\Lambda}(P) = -S_\Lambda(P) + |\Lambda| \log |A|. \tag{10.87}$$

The following result is a special case of (7.23), in which we replace A by A^Λ, q by q^Λ, and f by $-\beta H_\Lambda$, but it is instructive to state and prove it directly:

[132] Dorlas (2021), Chapter 24, and others even justify the local Gibbs measures on this basis. Physically speaking Proposition 8.2 describes non-interacting particles; a very large system of interacting particles may be divided into N large subsystems whose components interact with each other whilst boundary interactions between the various subsystems may be neglected. Looking at each subsystem as a particle, eq. (8.23) suggests (10.84).

Proposition 10.5 *For any finite $\Lambda \subset \mathbb{Z}^d$, $H_\Lambda : \Omega_\Lambda \to \mathbb{R}$, $\beta \in \mathbb{R}$, and $q \in \mathrm{Prob}(A)$, we have*

$$\Pi_{\Lambda,q}(\beta) = -\inf_{P \in \mathrm{Prob}(\Omega_\Lambda)} \{\beta \langle H_\Lambda \rangle_P + D_{q^\Lambda}(P)\}, \tag{10.88}$$

where the infimum is a minimum that is uniquely attained at $P = P_{\Lambda,q,\beta}$. The flat prior $q = f$ gives

$$\Pi_\Lambda(\beta) = \sup_{P \in \mathrm{Prob}(\Omega_\Lambda)} \{S_\Lambda(P) - \beta \langle H_\Lambda \rangle_P\}, \tag{10.89}$$

where the supremum is uniquely attained at the Gibbs measure $P = P_{\Lambda,\beta}$, cf. (10.84)*.*

For $\beta > 0$, eqs. (10.88) and (10.89) evidently give the thermodynamical formulae "$F = E - TS$":

$$F_{\Lambda,q}(\beta) = \inf_{P \in \mathrm{Prob}(\Omega_\Lambda)} \{\langle H_\Lambda \rangle_P + \beta^{-1} D_{q^\Lambda}(P)\}; \tag{10.90}$$

$$F_\Lambda(\beta) = \inf_{P \in \mathrm{Prob}(\Omega_\Lambda)} \{\langle H_\Lambda \rangle_P - S_\Lambda(P)\}. \tag{10.91}$$

Thus at nonzero temperature *Gibbs measures minimize the free energy*, much as ground states (see below) minimize the energy at zero temperature. Compare also (5.12).

Exercise 72 *Prove Proposition 10.5.*

As a special case Proposition 8.2, in which we simply replace A by A^Λ and E by H_Λ, we see:

Proposition 10.6 *For any $U \in (\min_\sigma H_\Lambda(\sigma), \max_\sigma H_\Lambda(\sigma))$ there is a unique $\beta \in \mathbb{R}$ such that*

$$\langle H_\Lambda \rangle_{P_{\Lambda,\beta}} = U; \tag{10.92}$$

$$S_\Lambda(P_{\Lambda,\beta}) = \sup\{S_\Lambda(P) \mid P \in \mathrm{Prob}(\Omega_\Lambda), \langle H_\Lambda \rangle_P = U\}. \tag{10.93}$$

Proof. Also here a direct proof is instructive (though unnecessary). The function $\beta \mapsto \Pi_\Lambda(\beta)$ is smooth (in the finite system at hand it is even analytic), with first and second derivatives

$$\frac{d}{d\beta}\Pi_\Lambda(\beta) = -\langle H_\Lambda \rangle_{P_{\Lambda,\beta}}; \qquad \frac{d^2}{d\beta}\Pi_\Lambda(\beta) = \mathrm{Var}_{P_{\Lambda,\beta}}(H_\Lambda) \geq 0, \tag{10.94}$$

cf. footnote 72, with equality near the end iff H_Λ is constant (in which case the proposition is true in an empty way). Thus the continuous function $\beta \mapsto \langle H_\Lambda \rangle_{P_{\Lambda,\beta}}$ is strictly decreasing. Hence

$$\lim_{\beta \to -\infty} \langle H_\Lambda \rangle_{P_{\Lambda,\beta}} = \max_\sigma H_\Lambda(\sigma)); \qquad \lim_{\beta \to \infty} \langle H_\Lambda \rangle_{P_{\Lambda,\beta}} = \min_\sigma H_\Lambda(\sigma)), \tag{10.95}$$

from which the first claim follows. To prove (10.93), we now fix β at the value yielding (10.92) for given $U \in \mathbb{R}$, and use the variational principle (10.89) to estimate for any P with $\langle H_\Lambda \rangle_P = U$:

$$S_\Lambda(P) - \beta U = S_\Lambda(P) - \beta \langle H_\Lambda \rangle_P \leq S_\Lambda(P_{\Lambda,\beta}) - \beta \langle H_\Lambda \rangle_{P_{\Lambda,\beta}} = S_\Lambda(P_{\Lambda,\beta}) - \beta U, \tag{10.96}$$

where in the last step we used (10.92). Hence $S_\Lambda(P) \leq S_\Lambda(P_{\Lambda,\beta})$, whence (10.93). □

The point is now to study the limit $\Lambda \to \mathbb{N}$ or '$\Lambda \nearrow \mathbb{Z}^d$ in a suitable sense (here taken to be $\Lambda = \Lambda_N$ and $N \to \infty$), hoping to see for example phase transitions through lack of uniqueness of limiting measures, and other interesting phenomena. If these limits exist, a key role is played by

$$p(\beta) = \lim_{N \to \infty} \frac{1}{|\Lambda_N|} \log Z_{\Lambda_N}(\beta); \qquad f(\beta) := \lim_{N \to \infty} \frac{1}{|\Lambda_N|} F_{\Lambda_N} = -\frac{p(\beta)}{\beta}, \tag{10.97}$$

called the *pressure* and the *free energy*, respectively. It is hard to miss the formal analogy with (the Gärtner–Ellis) Theorem 9.8, which indeed will be explored in what follows (see e.g. §10.3).

10.5 Global Gibbs measures

We define global Gibbs measures as extensions of the local expressions (10.79) or (10.84) to $\Omega = A^{\mathbb{Z}^d}$, where for simplicity we still assume that A is finite (but the formalism is much more general).[133] The main feature will be that such extensions are not necessarily unique.

We first include *boundary conditions* in (10.74) and (10.84). As before, Λ *is a finite sublattice of* $\mathbb{Z}^d$. Fix $\eta \in \Omega$. For any Λ and $\sigma \in A^\Lambda$, define $\omega = \sigma\eta \in \Omega$ by $\omega(x) = \sigma(x)$ for $x \in \Lambda$ and $\omega(x) = \eta(x)$ for $x \notin \Lambda$ (some people therefore write $\sigma\eta_{|\Lambda^c}$ instead of $\sigma\eta$; our notation is easier provided we only write $\sigma\eta$ iff it is clear that $\sigma \in A^\Lambda$ and we realize that *within the expression* $\sigma\eta$ we interpret η as $\eta_{|\Lambda^c}$). Given a potential Φ (see Definition 10.4) and some Λ, extend (10.74) to

$$H_\Lambda^\eta(\sigma) = \sum_{M \cap \Lambda \neq \emptyset} \Phi_M(\sigma\eta), \tag{10.98}$$

i.e. the sum includes all finite $M \subset \mathbb{Z}^d$ that overlap with the given Λ. For example, for the Ising model we may have $x \in \Lambda$, $\|x-y\| = 1$, but $y \notin \Lambda$, so that (10.75) includes terms $-J\sigma_x\eta_y$, which are absent from (10.74). Here one may think of $\eta(x) = \pm 1$ for all x. Then introduce probabilities

$$P_{\Lambda,\beta}^\eta(\sigma) := \frac{1}{Z_\Lambda^\eta(\beta)} e^{-\beta H_\Lambda^\eta(\sigma)} \ (\sigma \in A^\Lambda); \qquad Z_\Lambda^\eta(\beta) = \sum_{\sigma \in A^\Lambda} e^{-\beta H_\Lambda^\eta(\sigma)} = e^{-\beta F_\Lambda^\eta(\beta)}; \tag{10.99}$$

$$P_{\Lambda,\beta}^\eta(B) := \sum_{\sigma \in A^\Lambda | \sigma\eta \in B} P_{\Lambda,\beta}^\eta(\sigma) = \frac{1}{Z_\Lambda^\eta(\beta)} \sum_{\sigma \in A^\Lambda} 1_B(\sigma\eta) e^{-\beta H_\Lambda^\eta(\sigma)} \qquad (B \in \mathscr{F}). \tag{10.100}$$

The expression $P_{\Lambda,\beta}^\eta(\sigma)$ only depends on $\sigma \in A^\Lambda$ and $\eta_{|\Lambda^c}$ (with $\Lambda^c := \mathbb{Z}^d \backslash \Lambda$), since $H_\Lambda^\eta(\sigma)$ and hence also $Z_\Lambda^\eta(\beta)$ only depend on $\eta_{|\Lambda^c}$. For later use, note that $P_{\Lambda,\beta}^\omega(\{\sigma\eta\})$ equals $P_{\Lambda,\beta}^\eta(\sigma)$ iff $\eta = \omega$ outside Λ, and vanishes otherwise. Using (10.74), we symbolically use "$\eta =$ free" to mean

$$P_{\Lambda,\beta}^{\text{free}}(\sigma) := \frac{1}{Z_\Lambda^{\text{free}}(\beta)} e^{-\beta H_\Lambda^{\text{free}}(\sigma)}; \qquad Z_\Lambda^{\text{free}}(\beta) = \sum_{\sigma \in A^\Lambda} e^{-\beta H_\Lambda^{\text{free}}(\sigma)}. \tag{10.101}$$

In the following definition Φ and β (which is often absorbed into Φ or H_Λ) are fixed (although the dependence on Φ is suppressed in the notation). The tail σ-algebra $\mathscr{T}_\Lambda$ that will appear was defined in Definition 10.3, and the ensuing conditional expectation was recalled in general in (4.26).

Definition 10.7 *A probability measure* $P_\beta \in \text{Prob}(\Omega)$ *is a* Gibbs measure *at inverse temperature* $\beta \in \mathbb{R}$ *iff for all finite* $\Lambda \subset \mathbb{Z}^d$ *and* $B \in \mathscr{F}$ *(i.e. the cylindrical* σ*-algebra on* $\Omega = A^{\mathbb{Z}^d}$*) we have*

$$P_\beta(B \mid \mathscr{T}_\Lambda)(\omega) = P_{\Lambda,\beta}^\omega(B), \tag{10.102}$$

for P_β*-almost each* $\omega \in \Omega$*, or, equivalently (given the definition of a conditional expectation):*

$$P_\beta(B) = \int_\Omega dP_\beta(\omega)\, P_{\Lambda,\beta}^\omega(B). \tag{10.103}$$

Eq. (10.102) or (10.103) is called the DLR-*equation* (after Dobrushin, Lanford, and Ruelle, who proposed it). This looks obscure. One simplification arises if we take $B = [\sigma]_\Lambda$ for some $\sigma \in A^\Lambda$.

[133] Georgii (2011) and Friedli & Velenik (2018) are excellent introductions to Gibbs measures; for a very useful summary see also Van Enter, Fernández, & Sokal (1993). All these authors use the formalism of *specifications*, which we avoid, at some cost to the generality of our treatment and at greater cost, unfortunately, to the proofs we give.

In that case the condition $\sigma\omega \in B$ holds for arbitrary $\omega \in \Omega$, and since for given $\sigma \in A^\Lambda$ the expression $P^\omega_{\Lambda,\beta}([\sigma]) = P^\omega_{\Lambda,\beta}(\sigma)$ only depends on ω_{Λ^c} (as already noted), eq. (10.103) reads

$$P_\beta([\sigma]) = \pi_\Lambda^{-1} P_\beta(\sigma) = \int_{\Omega_{\Lambda^c}} dP_{\Lambda^c}(\omega) P^\omega_{\Lambda,\beta}(\sigma), \tag{10.104}$$

with $P_{\Lambda^c} = \pi_{\Lambda^c}^{-1} P_\beta$); the mere *existence* of some probability measure P_{Λ^c} on Ω_{Λ^c} for which (10.104) holds, for each finite $\Lambda \subset \mathbb{Z}^d$ and $\sigma \in \Omega_\Lambda$, is equivalent to the conditions in Definition 10.7.[134]

Another special case that may clarify the meaning of (10.103) is $B = \{\sigma\eta\}$ for fixed $\sigma \in \Omega_\Lambda$ and $\eta \in \Omega$, but unfortunately, like for any point, $P_\beta(B) = 0$ in that case. To remedy this, we replace $\mathbb{Z}^d$ by a finite sublattice $\Lambda' \subset \mathbb{Z}^d$ that contains Λ (i.e., $\Lambda \subset \Lambda'$), and replace condition (10.103) by

$$P_{\Lambda'}(B) = \sum_{\omega \in \Omega_{\Lambda'}} P_{\Lambda'}(\omega) P^\omega_{\Lambda,\beta}(B). \tag{10.105}$$

If we now put $B = \{\sigma\eta\}$, where now $\eta \in \Omega_{\Lambda'}$, we see from (10.100) that only terms in which $\omega_{|\Lambda'\setminus\Lambda} = \eta$ contribute to the sum in (10.105), which therefore comes down to

$$P^{(\Lambda')}_\beta(\sigma\eta) = \sum_{\omega \in \Omega_{\Lambda'} | \omega_{|\Lambda'\setminus\Lambda} = \eta} P^{(\Lambda')}_\beta(\omega) P^\eta_{\Lambda,\beta}(\sigma). \tag{10.106}$$

Using the elementary conditional expectation $P(B \mid C) = P(B \cap C)/P(C)$, this is the same as:[135]

$$P^{(\Lambda')}_\beta(\omega_{|\Lambda} = \sigma \mid \omega_{|\Lambda'\setminus\Lambda} = \eta) = P^\eta_{\Lambda,\beta}(\sigma). \tag{10.107}$$

In words: the conditional probability that ω coincides with σ on Λ given its value η outside Λ equals the local Gibbs probability of σ for the Hamiltonian with boundary condition η.

Exercise 73 *In the above situation:*

1. *Using* (10.98)*, where $H_{\Lambda'}$ has free boundary conditions (so $\eta =$ "free"), that the difference $H_{\Lambda'}(\sigma\eta) - H_\Lambda(\sigma\eta)$ is independent of $\sigma \in A^\Lambda$ for any η.*[136]

2. *Use this independence to show that eq.* (10.107) *is solved by*

$$P^{(\Lambda')}_\beta(\omega) = \frac{1}{Z_{\Lambda'}} e^{-\beta H_{\Lambda'}(\omega)} \ \ (\omega \in A^{\Lambda'}); \qquad Z_{\Lambda'} = \sum_{\omega' \in A^{\Lambda'}} e^{-\beta H_{\Lambda'}(\omega)}. \tag{10.108}$$

3. *Now assume that $\Lambda' \subset \mathbb{Z}^d$, and show that* (10.98) *satisfies the independence condition in point 1 for any boundary condition $\eta' \in \Omega$ (not to be confused with η in $\sigma\eta$).*

4. *Use this fact to show that eq.* (10.107) *is solved by $P^{(\Lambda')}_\beta = P^{\eta'}_{\Lambda',\beta}$.*

Although (10.107) is only valid for *finite* Λ' (as otherwise $P^{(\Lambda')}_\beta(\omega_{|\Lambda'\setminus\Lambda} = \eta) = 0$), the point is that:

A global Gibbs measure is defined by specifying conditional probabilities for all finite sublattices Λ of $\mathbb{Z}^d$ and boundary conditions outside these sublattices in the form of local Gibbs measures.

[134] See Ruelle (2004), §1.5.

[135] The left-hand side is short for $P^{(\Lambda')}_\beta(B \mid C)$ with $B = \{\omega \in \Omega_{\Lambda'} \mid \omega_{|\Lambda} = \sigma\}$ and $C = \{\omega \in \Omega_{\Lambda'} \mid \omega_{|\Lambda'\setminus\Lambda} = \eta\}$.

[136] Note that the independence condition in point 1 is reasonable (except in mean-field models), since $H_{\Lambda'}(\sigma\eta) - H_\Lambda(\sigma\eta)$ contains the interactions between spins in $\Lambda'\setminus\Lambda$, which are given by η.

Another way to (re)write (10.102) - (10.103) is to realize that each $\Lambda \subset \mathbb{Z}^d$ induces a factorization

$$\Omega \xrightarrow{\cong} \Omega_\Lambda \times \Omega_{\Lambda^c}; \qquad \omega \mapsto (\omega_{|\Lambda}, \omega_{|\Lambda^c}), \tag{10.109}$$

where $\Lambda^c = \mathbb{Z}^d \backslash \Lambda$. Any two measures μ_1, μ_2 on measure spaces X_1, X_2 induce a product measure $\mu_1 \times \mu_2$ on the Cartesian product $X_1 \times X_2$ by extending $\mu_1 \times \mu_2(B_1 \times B_2) := \mu_1(B_1)\mu_2(B_2)$, where $B_i \subset X_i$ (assumed measurable).[137] Any positive measurable function $f : X_1 \times X_2 \to \mathbb{R}^+$ induces a further measure $f \cdot (\mu_1 \times \mu_2)$ on $X_1 \times X_2$. In this construction, take $\Lambda \subset \mathbb{Z}^d$ finite, $X_1 = \Omega_\Lambda$, $X_2 = \Omega_{\Lambda^c}$, $\mu_1 = \delta_\Lambda$ (i.e. the counting measure $\delta_\Lambda(\sigma) = 1$ for each $\sigma \in \Omega_\Lambda$), $\mu_2 = \pi_{\Lambda^c}^{-1} P$ for some $P \in \mathrm{Prob}(\Omega)$, where $\pi_{\Lambda^c} : \Omega \to \Omega_{\Lambda^c}$ is projection onto the second coordinate, and finally

$$f = P_{\Lambda,\beta} \qquad P_{\Lambda,\beta}(\omega) = P_{\Lambda,\beta}^{\omega_{|\Lambda^c}}(\omega_{|\Lambda}), \tag{10.110}$$

cf. (10.99) and subsequent comment. The DLR-equation (10.102) or (10.103) then comes down to

$$P_\beta = P_{\Lambda,\beta} \cdot (\delta_\Lambda \times \pi_{\Lambda^c}^{-1} P_\beta). \tag{10.111}$$

Warning:[138] The local Gibbs measures $P_{\Lambda,\beta}^\eta$ are not the marginals $\pi_\Lambda^{-1} P_\beta$ of some global Gibbs measure P_β, not even for free boundary conditions, and not even of some other local Gibbs measure $P_{\Lambda',\beta}^{\eta'}$ for some finite $\Lambda' \supset \Lambda$. Thus although *in principle* a global Gibbs measure P_β could be defined by its marginals, as in the Kolmogorov extension lemma 3.3, *in practice* such a definition is useless since these marginals can only be computed or known otherwise if we already know P_β.

Having said this, local Gibbs measures do converge to global ones! Here is the situation:[139]

Theorem 10.8 *Fix a potential Φ and an inverse temperature $\beta \in \mathbb{R}$.*

1. *For any boundary condition $\eta \in \Omega$, or a free boundary condition in which H_Λ^η is given by* (10.74)*, the sequence $(P_{\Lambda_N,\beta}^\eta)_N$ of local Gibbs measures on Ω_{Λ_N} has a subsequence that converges to some global Gibbs measure P_β in the following sense: for any finite Λ and from sufficiently large N such that $\Lambda \subset \Lambda_N$ onwards we have, cf.* (3.8)*,*

$$\pi_\Lambda^{-1} P_\beta = \lim_{N\to\infty} \pi_{\Lambda_N \Lambda}^{-1} P_{\Lambda_N,\beta}^\eta. \tag{10.112}$$

   In particular, there exists a limit Gibbs state for free boundary conditions, denoted by P_β^{free}.

2. *The set $\mathscr{G}(\beta\Phi)$ of all (global) Gibbs states is the closed convex hull of these limit states.*[140]

3. *The set $\mathscr{G}(\beta\Phi)$ is a nonempty compact convex subspace of* $\mathrm{Prob}(\Omega)$.

4. *The extreme boundary $\partial_e \mathscr{G}(\beta\Phi)$ consists of all $P \in \mathscr{G}(\beta\Phi)$ that are trivial on $\mathscr{T}$.*

5. *The extreme boundary $\partial_e(\mathscr{G}(\beta\Phi) \cap \mathrm{Prob}^{\mathbb{Z}^d}(\Omega))$ consists of all $P \in \mathscr{G}(\beta\Phi)$ that are trivial on $\mathscr{I}$ (and hence are ergodic Gibbs measures); in other words, we have*

$$\partial_e(\mathscr{G}(\beta\Phi) \cap \mathrm{Prob}^{\mathbb{Z}^d}(\Omega)) = \mathscr{G}(\beta\Phi) \cap \partial_e \mathrm{Prob}^{\mathbb{Z}^d}(\Omega). \tag{10.113}$$

[137] See e.g. Dudley (1989), §4.4.

[138] We owe this comment to Friedli & Velenik (2018), pp. 269–270.

[139] See Ruelle (2004), Theorem 1.9; Georgii (2011), Theorems (14.5) and (14.15); Rassoul-Agha (2015), Theorem 7.24; We use the weak topology on $\mathrm{Prob}(\Omega)$ for convergence of measures, see Definition 4.8. The Portmanteau theorem is very useful here, since for $\Omega = A^{\mathbb{Z}^d}$ we have $P_n \to P$ weakly iff $P_n(B) \to P(B)$ for any cylinder sets $B \subset \Omega$. Indeed, in the product topology on Ω cylinder sets are clopen and hence satisfy $B^- \backslash \mathring{B} = \emptyset$, so that $P(B^- \backslash \mathring{B}) = 0$ for any P.

[140] Like $\mathrm{Prob}^T(X)$ for compact metrizable X, the set $\mathscr{G}(\beta\Phi)$ is a *simplex*, see Definition 4.10.

Recall that $\mathscr{T}$ is the tail σ-algebra, cf. Definition 10.3. The invariant σ-algebra $\mathscr{I}$ and the last claim in general will now be explained.[141] So far, we just defined ergodicity in the setting of dynamical systems (X,P,T), where (X,Σ,P) is a probability space and $T : X \to X$ is measurable. But if T is invertible this amounts to a $\mathbb{Z}$-action on X, given by $n : x \mapsto T^n x$. A more abstract setting of ergodic theory is then given by measurable G-actions on a measure space, which we always take to be a probability space (X,Σ,P), in which P is G-invariant.[142] We take $G = \mathbb{Z}^d$ and its action on $\Omega = A^{\mathbb{Z}^d}$ the shift action induced by the left-action (10.72) of $\mathbb{Z}^d$ onto itself, i.e.,

$$(y \cdot \omega)_x := \omega_{x+y}. \tag{10.114}$$

Cf. the shift (2.5), which (defined on $A^{\mathbb{Z}}$) is a special case; but we also consider general compact metrizable spaces Ω, with the ensuing Borel structure Σ. This class of spaces overlaps with $A^{\mathbb{Z}^d}$ provided A is compact and $A^{\mathbb{Z}^d}$ carries the product topology (in which it is compact by Tychonoff's theorem).[143] We say that a measurable set $B \in \Sigma$ is *G-invariant* iff $x \cdot B = B$ for all $x \in G$, where

$$x \cdot B = \{x \cdot \omega, \omega \in B\}. \tag{10.115}$$

These G-invariant measurable sets form the *invariant σ-algebra* $\mathscr{I}$, and clearly measurable functions are G-invariant iff they are $\mathscr{I}$-measurable. Instead of Definition 4.1 we say that the G-action is *ergodic* iff $B \in \mathscr{I}$ implies $P(B) = 0$ or $P(B) = 1$. Proposition 4.3 then remains valid if we replace $f(Tx) = f(x)$ by $f(x \cdot \omega) = f(\omega)$ for all $x \in G$. For compact Ω it is enough to require this for all *continuous* functions f, so that the G-action is ergodic iff any G-invariant function $f \in C(\Omega)$ is constant, where G-invariance now means that $f(x \cdot \omega) = f(\omega)$ for all $x \in G$. We only state the ergodic theorem, in a form suggested by (4.23), for $G = \mathbb{Z}^d$ acting on $\Omega = A^{\mathbb{Z}^d}$:[144]

Theorem 10.9 *Let Ω be compact metrizable space equipped with both a $\mathbb{Z}^d$-action and a $\mathbb{Z}^d$-invariant probability measure P. For any $f \in L^1(\Omega,P)$, the limit*

$$f^*(\omega) := \lim_{N\to\infty} \frac{1}{|\Lambda_N|} \sum_{x\in\Lambda_N} f(x \cdot \omega) \tag{10.116}$$

exists pointwise P-a.e. and equals a G-invariant function $f^* \in L^1(\Omega,P)$, given by

$$f^* = E_P(f \mid \mathscr{I}). \tag{10.117}$$

In particular, if the G-action is ergodic, then for P-almost every $\omega \in \Omega$ we have

$$\lim_{N\to\infty} \frac{1}{|\Lambda_N|} \sum_{x\in\Lambda_N} f(x \cdot \omega) = \int_\Omega dP\, f, \tag{10.118}$$

[141] For our purposes Mackey (1974) is a nice introduction to the relevant parts of ergodic theory.

[142] This more *abstract* setting is not more *general* than the previous one, since T need not be invertible and hence may not define a group action. In the abstract setting one requires G to be *amenable* (and *a priori* locally compact). This means that every continuous G-action on a compact metrizable space Ω has an invariant probability measure, and may also be defined by the property that for any compact subset $K \subset G$ and $\delta > 0$ there is a Borel set $B \subset G$ with compact closure such that $\mu(B\Delta(K \cdot B)) < \delta\mu(B)$, where μ is the (left) Haar measure on G. Every compact group and every locally compact abelian group is amenable (but there are others). In the ergodic theorem sequences of the type $\Lambda_N \subset \mathbb{Z}^d$ are replaced by so-called *Følner sequences*, defined as compact subsets $B_N \subset G$ that (at least for sufficiently large N) have the above property $\mu(B_N\Delta(K \cdot B_N)) < \delta\mu(B_N)$ for every compact $K \subset G$ and $\delta > 0$.

[143] For any set T and any space A the *product topology* on $\Omega = A^T$ is the coarsest topology that makes all evaluations map $A^T \to A$, $\omega \mapsto \omega_x$, continuous; it is generated by the cylinder sets (like the σ-algebra $\mathscr{F}$). For $T = \mathbb{Z}^d$ the product topology is metrizable, for example by the metric $d(\omega,\omega') := \lambda^{-\inf\{\|x\|, x\in\mathbb{Z}^d, \omega_x\neq\omega'_x\}}$, where $\|x\| := \max_{i=1,\dots d}\{|x^i|\}$ or any other norm, and any $\lambda > 1$, with the convention or theorem that $\inf\emptyset = \infty$ as usual, so that $d(\omega,\omega) = 0$. For $d = 1$ and $A = \{0,1,\dots,|A|-1\}$ an inequivalent metric that nonetheless gives the same topology would be $d'(\omega,\omega') = \sum_{n=-\infty}^{\infty} \lambda^{-|n|}|\omega_n - \omega'_n|$, again for any $\lambda > 1$. See Katok & Hasselblatt (1995), §1.9.

[144] See Georgii (2011), Theorem 14.A8. For more general (amenable) groups and spaces see Lindenstrauss (2001).

or, equivalently, for P-almost every $\omega \in \Omega$ as weak convergence of probability measures,

$$\lim_{N\to\infty} \frac{1}{|\Lambda_N|} \sum_{x\in\Lambda_N} \delta_{x\cdot\omega} = P. \tag{10.119}$$

Even for finite A the case $\Omega = A^{\mathbb{Z}^d}$ is rich enough to give rise to a strange structure of the compact convex set $K = \mathrm{Prob}^{\mathbb{Z}^d}(\Omega)$ of invariant probability measures on Ω: it is always a *Poulsen complex*, in the sense that $\partial_e K$ is (weakly) dense in K (which for finite Ω would be unheard of).[145]

Here is an example of this phenomenon, for $d = 1$ and $A = \{0,1\}$, so that, seen as a dynamical system (Ω, P, T), the map T is the shift map $T = S$, (2.5), now defined for all $n \in \mathbb{Z}$. For some given S-invariant $P \in \mathrm{Prob}^S(\Omega)$, we construct ergodic probability measures $P_N \in \partial_e \mathrm{Prob}^S(\Omega)$ such that $P_N \to P$ (weakly). This construction is based on writing $\mathbb{Z}$ as a disjoint union, for given $N > 0$,

$$\mathbb{Z} = \bigsqcup_{k\in\mathbb{Z}} I_k; \qquad\qquad I_k := [k(2N+1) - N, k(2N+1) + N], \tag{10.120}$$

so that each blok I_k in the partition has length $2N+1$. This in turn induces a decomposition

$$\{0,1\}^{\mathbb{Z}} = \times_{k\in\mathbb{Z}} \{0,1\}^{I_k}. \tag{10.121}$$

Given $P \in \mathrm{Prob}^S(\Omega)$, for fixed $N > 0$ we then define $P_{\Lambda_N} \in \mathrm{Prob}(\Omega_{\Lambda_N})$, with $\Lambda_N = [-N, N]$ and $\Omega_\Lambda = A^\Lambda$, as usual, i.e. by $P_{\Lambda_N}(\sigma) = P([\sigma]) = P(\{\omega \in \Omega \mid \omega_{|\Lambda_N} = \sigma_{|\Lambda_N}\})$ for any $\sigma \in \Omega_{\Lambda_N}$. Noting that $\Lambda_N = I_0$, we then transfer P_{Λ_N} to any Ω_{I_k} by translation, that is, define

$$P_{I_k}^{(k,N)}(\sigma') = P_{I_0}(S^{k(2N+1)}\sigma'), \tag{10.122}$$

where $\sigma' \in \Omega_{I_k}$ and hence $S^{k(2N+1)}\sigma' \in \Omega_{I_0}$. Finally, define our measures P_N on Ω via

$$P_N := \frac{1}{2N+1} \sum_{n=-N}^{N} S^{-n} Q_N; \qquad\qquad Q_N = \times_{k\in\mathbb{Z}} P_{I_k}^{(k,N)}. \tag{10.123}$$

The reason for this move is that the product measure Q_N is not S-invariant, but merely periodic, in that $S^{2N+1}Q_N = Q_N$ by construction (translation by an block I_0 of length $2N+1$).

Exercise 74 *Prove that P_N is S-invariant.*

Since "$\lim_{N\to\infty}[-N,N] = \mathbb{Z}$", it is natural that $P_N \to P$. A proof uses the fact that $\cup_M \mathscr{F}_{\Lambda_M}$ is dense in $C(\Omega)$, and so to prove $P_N \to N$ weakly it suffices that $|P_N(f) - P(f)| \to 0$ for all M and all $f \in \mathscr{F}_{\Lambda_M}$ (whose elements are continuous!). Indeed, for $N > M$ one shows that

$$|P_N(f) - P(f)| \le 2 \cdot \frac{2M+1}{2N+1} \|f\|_\infty. \tag{10.124}$$

The proof that P_N is ergodic (cf. Definition 4.1) relies on a nice lemma (of we we only need half):

Lemma 10.10 *Let (X, P, T) a dynamical system with X compact and T invertible, and consider*

$$f_{\Lambda_N}(x) := \frac{1}{2N+1} \sum_{n=-N}^{N} f(T^n x), \tag{10.125}$$

for $f \in L^2(X,P)$. Then P is ergodic iff for all f in a dense subspace of $L^2(X,P)$ we have

$$\lim_{N\to\infty} \langle |f_{\Lambda_N}|^2 \rangle_P = |\langle f \rangle_P|^2. \tag{10.126}$$

[145] See Georgii (2011), Theorem (14.12), and Simon (2011), Example 9.7, reviewed below. See also Definition 4.10.

Exercise 75 *Prove this lemma, and then check that* (10.126) *with* $P \rightsquigarrow P_N$ *holds.*

Using (10.123) then gives $\lim_{L\to\infty}\langle f^2_{\Lambda_L}\rangle_{P_N} = \langle f\rangle^2_{P_N}$, which is (10.126). Since M was arbitrary and $\cup_M \mathscr{F}_{\Lambda_M}$ is dense in $L^2(\Omega, P')$ for any probability measure P' defined on the cylindrical σ-algebra generated by the $\mathscr{F}_{\Lambda_M}$, P_N is ergodic by Lemma 10.10. Since $P_N \to P$ and $P \in \mathrm{Prob}^S(\{0,1\}^{\mathbb{Z}})$ was arbitrary, we have shown that the compact convex set $\mathrm{Prob}^S(\{0,1\}^{\mathbb{Z}})$ is a Poulsen simplex!

As in the local case, the unbiased Bernoulli measure $P = f^{\mathbb{Z}^d}$ is a Gibbs measure at $\beta = 0$ (i.e. $T = \infty$). To see this, take $\Lambda' \subset \mathbb{Z}^d$ finite, and a generic cylinder set defined by $\eta' \in A^{\Lambda'}$:

$$B = [\eta'] := \{\omega \in \Omega \mid \omega_{|\Lambda'} = \eta'\} \in \mathscr{F}_{\Lambda'}. \tag{10.127}$$

Then on the *Ansatz* $P_0 = f^{\mathbb{Z}^d}$ the left-hand side of (10.103) equals

$$f^{\mathbb{Z}^d}([\eta']) = \prod_{x\in\Lambda'} q(\eta'_x) = \prod_{x\in\Lambda'} |A|^{-1} = |A|^{-|\Lambda'|}. \tag{10.128}$$

Now we make a (logically unnecessary but pedagogically useful) case distinction:

1. If $\Lambda \cap \Lambda' = \emptyset$, then $P^\omega_{\Lambda,0}([\eta']) = 1$, since

   $$Z^\omega_\Lambda(0) = |A^\Lambda| = |A|^{|\Lambda|} \tag{10.129}$$

   in (10.99), whilst the second sum over σ in (10.100) also includes all $\sigma \in A^\Lambda$. The right-hand side of (10.103) therefore equals

   $$\int_{\omega_{\Lambda'}=\eta'} df^{\mathbb{Z}^d}(\omega) = f^{\mathbb{Z}^d}([\eta']) = |A|^{-|\Lambda'|}. \tag{10.130}$$

2. If $M := \Lambda \cap \Lambda' \neq \emptyset$, then (10.129) of course remains true, but this time the second sum over σ in (10.100) is constrained to those $\sigma \in A^\Lambda$ that satisfy $\sigma\omega \in [\eta']$, which implies

   $$\sigma_x = \eta'_x \quad (x \in M); \qquad\qquad \omega_x = \eta'_x \quad (x \in \Lambda'\backslash M). \tag{10.131}$$

   By the first equation the second sum in in (10.100) has only $|A^{\Lambda\backslash M}| = |A|^{|\Lambda|-|M|}$ terms, so that

   $$P^\omega_{\Lambda,0}([\eta']) = \frac{1}{Z^\omega_\Lambda(0)} \sum_{\sigma\in A^\Lambda | \sigma\omega\in[\eta']} e^0 = \frac{|A|^{|\Lambda|-|M|}}{|A|^{|\Lambda|}} = |A|^{-|M|}. \tag{10.132}$$

   The second equation in (10.131), which the reader should verify from the explanation preceding (10.98), implies that the integral over ω on the right-hand side of (10.103) is constrained to $\omega \in \Lambda'\backslash M$, and hence equals $|A|^{-(|\Lambda'|-|M|)}$ times the result of (10.132), so that we obtain

   $$\int_\Omega dP_\beta(\omega)\, P^\omega_{\Lambda,\beta}([\eta']) = |A|^{-(|\Lambda'|-|M|)} \cdot |A|^{-|M|} = |A|^{-|\Lambda'|}, \tag{10.133}$$

   and once again we see that $P_0 = f^{\mathbb{Z}^d}$ is a Gibbs measure. Note that case 1 follows from 2.

Stationary Markov chains also fall under this scope, at least under a condition that is even stronger than irreducibility, namely strict positivity ($P_{ab} > 0$) of all transition probabilities (3.26) (and finite A). In view of Theorem 3.6 this gives a unique stationary probability $P \in \mathrm{Prob}(A^{\mathbb{N}})$ such that the underlying stochastic process has joint probabilities (3.30). Realizing the process on $\Omega = A^{\mathbb{N}}$ (cf. Theorem 3.4), which we now interpret as a spin chain, the pair potential defined by

$$\Phi_{n,n+1}(\sigma) := -\log P_{\sigma(n),\sigma(n+1)}; \qquad\qquad \Phi_M = 0 \ \ (M \neq \{n, n+1\}), \tag{10.134}$$

turns out to have a unique global Gibbs measure at $\beta = 1$, which equals P. This only works if the Markov chain in question is indexed by $n \in \mathbb{Z}$ instead of $n \in \mathbb{N}$, which extension can trivially (and uniquely) be achieved by exploiting stationarity, cf. (3.26): we add random variables X_n also for $n < 0$ and define the ensuing "new" joint probabilities bu putting, assuming $m > 0$ and $k > -m$,

$$P(X_{-m} = a_{-m}, \dots, X_k = a_k, \dots) := P(X_0 = a_{-m}, \dots, X_{k+m} = a_k, \dots). \tag{10.135}$$

Short of proving this,[146] we just verify that P as given by (3.30) satisfies the DLR-equation (10.103), and even this we only do in a special case (which clearly shows what is going on): we take

$$B = (X_0 = a_0, \dots, X_N = a_N) \equiv \{\omega \in \Omega \mid X_0(\omega) = a_0, \dots, X_N(\omega) = a_N\}, \tag{10.136}$$

and take $\Lambda = \Lambda_M = \{-M, \dots, M\}$ for some $M > N$, cf. (10.62). The constraint $\sigma\omega \in B$ that appears in (10.103) with (10.100) then amounts to $\sigma_n = a_n$ for $n = 0, \dots, N$, and no constraint on ω at all. In computing (10.100) we therefore sum only over σ_{-M} to σ_{-1} and over σ_{N+1} to σ_M. Using (10.100), (10.134), and (10.98) we find

$$\sum_{\sigma \in A^{\Lambda_M} \mid \sigma\omega \in B} e^{-H^{\omega}_{\Lambda_M}(\sigma)} = P^{M+1}_{\omega_{-M-1}a_0} P_{a_0a_1} \cdots P_{a_{N-1}a_N} P^{M-N+1}_{a_N\omega_{M+1}}; \tag{10.137}$$

$$Z^{\omega}_{\Lambda_M}(1) = \sum_{\sigma_{-M}, \dots, \sigma_M} P_{\omega_{-M-1}\sigma_{-M}} \cdots P_{\sigma_M\omega_{M+1}} = P^{2M+2}_{\omega_{-M-1}\omega_{M+1}}. \tag{10.138}$$

To obtain (10.137) it may be useful to note the intermediate step where the left-hand side equals

$$\sum_{\sigma_{-M}, \dots, \sigma_{-1}; \sigma_{N+1}, \dots, \sigma_M} P_{\omega_{-M-1}\sigma_{-M}} P_{\sigma_{-M}\sigma_{-M+1}} \cdots P_{\sigma_{-1}a_0} P_{a_0a_1} \cdots P_{a_{N-1}a_N} P_{a_N\sigma_{N+1}} \cdots P_{\sigma_M\omega_{M+1}}.$$

Hence for B in (10.136), $\Lambda = \Lambda_M$ ($M > N$), $\beta = 1$, and Φ from (10.134) the r.h.s. of (10.103) equals

$$\int_\Omega dP_1(\omega)\, P^{\omega}_{\Lambda_M,1}(B) = P_{a_0a_1} \cdots P_{a_{N-1}a_N} \int_\Omega dP_1(\omega)\, \frac{P^{M+1}_{\omega_{-M-1}a_0} P^{M-N+1}_{a_N\omega_{M+1}}}{P^{2M+2}_{\omega_{-M-1}\omega_{M+1}}}. \tag{10.139}$$

We are checking that (10.103) holds for $P_1 = P$, the probability measure for the given Markov chain. Thus we put $P_1 = P$ in (10.139) and use the fact that stationarity of the Markov chain implies that the right-hand side of (10.139) is independent of M (as long as $M > N$). We may therefore let $M \to \infty$, and use (3.36) from Theorem 3.6 (which applies because we assumed that $P_{ab} > 0$). After some further analysis (of the second expression below) which we omit,[147] this gives

$$P^{M+1}_{\omega_{-M-1}a_0} \to p(a_0); \qquad \frac{P^{M-N+1}_{a_N\omega_{M+1}}}{P^{2M+2}_{\omega_{-M-1}\omega_{M+1}}} \to \frac{p(\omega_{M+1})}{p(\omega_{M+1})} = 1. \tag{10.140}$$

Thus (10.139) equals $p(a_0)P_{a_0a_1} \cdots P_{a_{N-1}a_N}$, which by (3.30) equals $P(B)$. We conclude that $P_1 = P$ satisfies (10.103) and hence is a Gibbs measure.

[146] See e.g. Keller (1998), §5.4, Rassoul-Agha (2015), Example 7.21, or Brémaud (2020), Theorem 10.1.9.

[147] See Keller, Proof of Theorem 5.4.1.

10.6 Large deviations and Fenchel duality for translation-invariant measures

We now extend Proposition 10.5 and the closely related duality (7.22) - (7.23) to infinite systems. This is technically difficult and we just outline the main steps and ideas.[148] In the next subsection we then give a large deviations perspective on these results, generalizing Sanov's theorem and Cramér's theorem to Gibbs measures. This theory has achieved a remarkable degree of perfection.

The relative entropy is a good place to start. In the context of ergodic theory (and Theorem 10.8 shows that Gibbs measures do fit this context), the original definition (7.16) - (7.17) is problematic in view of Lemma 4.4: for if p in this definition is ergodic and q is T-invariant (or, in the Gibbs context, $\mathbb{Z}^d$-invariant as we often assume), then $D_q(p)$ is either zero (if $q = p$) or infinite (if $q \perp p$). Fortunately, this can be remedied by a limit construction, which also clarifies the relationship between the relative entropy and the Kolmogorov–Sinai entropy.

For a general measure space (Ω, Σ) we fix a prior $Q \in \mathrm{Prob}(\Omega)$; if $\Omega = A^{\mathbb{Z}^d}$ (with $\Sigma = \mathscr{F}$) this will initially be the prior $Q = f^{\mathbb{Z}^d}$ induced by the flat prior $q = f$ on A, and later will be a Gibbs measure. Since we need to vary–and therefore explicitly include–the σ-algebra Σ on which P and Q are defined, we restore our original notation $D(P\|Q)$, cf. (7.4), and add Σ as a suffix. Thus:

$$D_\Sigma(P\|Q) := \int_\Omega dP \log \left(\frac{dP}{dQ} \right)_\Sigma \quad \text{if } P \ll Q \text{ on } \Sigma; \tag{10.141}$$

$$D_\Sigma(P\|Q) := \infty \quad \text{otherwise}, \tag{10.142}$$

where P and Q *are defined on* Σ, and also the Radon–Nikodym derivative dP/dQ *is defined with respect to* Σ and hence *must be measurable for* Σ. For example, if $\Omega = A^{\mathbb{Z}^d}$ (for finite A) and $\Sigma = \mathscr{F}_\Lambda$, in which case we write D_Λ for $D_{\mathscr{F}_\Lambda}$, for general $Q \in \mathrm{Prob}(A^{\mathbb{Z}^d})$ and $P \ll Q$ we have

$$\left(\frac{dP}{dQ} \right)_{\mathscr{F}_\Lambda} (\omega) = \frac{P([\omega_{|\Lambda}])}{Q([\omega_{|\Lambda}])}; \tag{10.143}$$

$$D_\Lambda(P\|Q) = \sum_{\sigma \in A^\Lambda} P([\sigma]) \log \left(\frac{P([\sigma])}{Q([\sigma])} \right), \tag{10.144}$$

cf. (10.68) for $[\sigma]$, which also implies the notation $[\omega_{|\Lambda}] = \{\omega' \in A^{\mathbb{Z}^d} \mid \omega'_{|\Lambda} = \omega_{|\Lambda}\}$ used here.

Exercise 76 *Prove* (10.143) *from* (10.69) *and the definition of the Radon–Nikodym derivative*[149]

And similarly for any measure space (Ω, Σ) whose underlying σ-algebra Σ is generated by a finite (measurable) partition of Ω. And even in general, we may write:[150]

$$D_\Sigma(P\|Q) = \sup_\Pi \sum_{U \in \Pi} P(U) \log \left(\frac{P(U)}{Q(U)} \right), \tag{10.145}$$

where the supremum is taken over all finite Σ-measurable partitions Π of Ω. Compare (2.57)!

Of course, for fixed Σ this refined relative entropy has the properties stated in Proposition 5.1.1. and is lsc, but it now acquires an additional monotonicity property if we vary Σ:

Proposition 10.11 *If* $\Sigma_1 \subset \Sigma_2$ *are* σ*-algebras on* Ω*, then* $D_{\Sigma_1}(P\|Q) \leq D_{\Sigma_2}(P\|Q)$.

[148] Details may be found in Keller (1998), Chapter 5; Ruelle (2004), Introduction and Chapters 3 and 4; Gerorgii (2011), Chapter 15; Rassoul-Agha (2015), Chapters 7 and 8.

[149] See footnote 54.

[150] See Georgii (2011), §15.1.

Since D_Σ will be a negative entropy, see eg. (10.153) below, and/or its earlier (and also later) identification as a rate function, and $\Sigma_1 \subset \Sigma_2$ means that the information in Σ_2 is finer (greater) than the information in Σ_1 (think of the extreme case $\Sigma_1 = \{\Omega, \emptyset\}$), this proposition states that the entropy of less information is greater than entropy of more information, as expected.

Exercise 77 *Before proving this proposition and applying it to statistical mechanics, let us show its power in a different context, namely the one discussed at the end of chapter 7. So let*

$$\Omega = A = B \times C, \tag{10.146}$$

for finite sets B and C for simplicity, let Σ_1 consist of all sets $\Delta \times C$, where $\Delta \subset B$, and let $\Sigma_2 = \mathscr{P}(\Omega)$. Thus Σ_1 is the coarsest σ-algebra on Ω that makes all functions that just depend on the first coordinate $b \in B$ measurable, where Σ_2 makes all functions on Ω measurable. Let $p \in \mathrm{Prob}(A)$ have marginals $r \in \mathrm{Prob}(B)$ and $s \in \mathrm{Prob}(C)$, see (7.35)*, and likewise let $q \equiv p' \in \mathrm{Prob}(A)$ have marginals $r' \in \mathrm{Prob}(B)$ and $s' \in \mathrm{Prob}(C)$. Suppose that $p \ll q$.*

1. *Explain the relationship between p restricted to Σ_1 and r (and likewise for p' and r').*
2. *Show that $(dp/dp')_{\Sigma_1} = \pi_1^*(dr/dr')_{\mathscr{P}(B)}$, where $\pi_1 : B \times C \to B$ is projection onto the first coordinate, and hence for $g : B \to \mathbb{R}$ the pullback $\pi_1^* g : B \times C \to \mathbb{R}$ is just $\pi_1^* g(b,c) = g(b)$.*
3. *Infer that $D_{\mathscr{P}(B)}(r\|r') \le D_{\mathscr{P}(B\times C)}(p\|p')$, or $D(P_X\|Q_X) \le D(P_{X,Y}\|Q_{X,Y})$, cf.* (7.38)*.*
4. *Prove this property also without invoking Proposition 10.11.*

Proof of Proposition 10.11. The proof relies on switching between dP/dQ defined with respect to Σ_2 and w.r.t. Σ_1. Denoting the former by ν_2 and the latter by ρ_1, they are related by

$$\rho_1 = E_Q(\rho_2 \mid \Sigma_1), \tag{10.147}$$

as can be checked from the definitions of the Radon–Nikodym derivative and the conditional expectation, cf. footnote 54 and (4.24), respectively. Eq. (10.147) leads to the identity

$$D_{\Sigma_1}(P\|Q) = \langle \varphi(\rho_1) \rangle_Q = \langle \varphi \circ E_Q(\rho_2 \mid \Sigma_1) \rangle_Q, \tag{10.148}$$

in terms of the (strictly) convex function $\varphi(x) = x \log x$. Jensen for conditional expectations gives

$$\langle \varphi \circ E_Q(\rho_2 \mid \Sigma_1) \rangle_Q \le \langle E_Q(\varphi \circ \rho_2 \mid \Sigma_1) \rangle_Q = \langle \varphi(\rho_2) \rangle_Q = D_{\Sigma_2}(P\|Q), \tag{10.149}$$

where we used (4.24). □

Exercise 78 *Show that* (10.143) *satisfies* (10.147) *with $\Sigma_1 = \mathscr{F}_{\Lambda_1}$ and $\Sigma_2 = \mathscr{F}_{\Lambda_2}$ for $\Lambda_1 \subset \Lambda_2$.*

We now relate this to the Kolmogorov-style entropies for dynamical systems defined in §2, we first generalize the latter to $\mathbb{Z}^d$-actions T on the given space Ω, assuming P to be $\mathbb{Z}^d$-invariant.[151] Given a partition $\pi = \{U_a, a \in A\}$ of X, cf. (2.28), for any finite sublattice $\Lambda \subset \mathbb{Z}^d$, e.g. the rectangles Λ_N in (10.63), instead of (2.30) we now refine π to

$$\pi^\Lambda = \{U_\sigma, \sigma \in A^\Lambda\}; \qquad\qquad U_\sigma := \bigcap_{x\in\Lambda} T^{-x} U_{\sigma_x}. \tag{10.150}$$

[151] Using the letter T may be confusing but actually clarifies the analogy. Seen as a name of a $\mathbb{Z}^d$-action on Ω (called X in §2), we have a measurable map $T : \mathbb{Z}^d \times \Omega \to \Omega$ satisfying the axioms of a group action; we write $x\omega$ for $T(x,\omega)$. If $T : X \to X$ is a map as in §2, now written $T : \Omega \to \Omega$, and T is invertible, the ensuing $\mathbb{Z}$-action is given by $x\omega = T^x\omega$.

In $d=1$, $\Lambda=\{0,\dots,N-1\}$ reproduces (2.30). On the other hand, take $\Omega=A^{\mathbb{Z}^d}$ with $\mathbb{Z}^d$-action

$$(x\omega)_y=\omega_{x+y}, \tag{10.151}$$

and take π to be the "canonical" partition of $A^{\mathbb{Z}^d}$ with elements

$$U_a=\{\omega\in A^{\mathbb{Z}^d}\mid \omega_0=a\}. \tag{10.152}$$

If we write $D_\Lambda(P)\equiv D_{\mathscr{F}_\Lambda}(P\|f^{\mathbb{Z}^d})$, it follows from (2.35) and (10.141) that

$$D_\Lambda(P)=-H_P(\pi^\Lambda)+|\Lambda|\log|A|, \tag{10.153}$$

cf. (1.28). To see this, note that the elements of π^Λ consist of the cylinder sets

$$[\sigma]=\{\omega\in A^{\mathbb{Z}^d}\mid \omega_{|\Lambda}=\sigma\}, \tag{10.154}$$

where σ runs through A^Λ, so that

$$H_P(\pi^\Lambda)=-\sum_{\sigma\in A^\Lambda}P([\sigma])\log P([\sigma]). \tag{10.155}$$

The Radon–Nikodym derivative in (10.141) by definition must be $\mathscr{F}_\Lambda$-measurable and satisfy

$$P(B)=\int_B dQ(\omega)\,\frac{dP}{dQ}(\omega), \tag{10.156}$$

for any $B\in\mathscr{F}_\Lambda$, i.e. for any cylinder set $[\sigma]$ as in (10.154). By (10.143), for $Q=f^{\mathbb{Z}^d}$ this equals

$$\frac{dP}{dQ}(\omega)=\frac{P([\omega_{|\Lambda}])}{Q([\omega_{|\Lambda}])}=|A|^{|\Lambda|}P([\omega_{|\Lambda}]), \tag{10.157}$$

cf. (10.143). Since $P\ll f^{\mathbb{Z}^d}$ for any P, we may use (10.141). Since P is defined on $\mathscr{F}_\Lambda$, we obtain:

$$\begin{aligned}D_\Lambda(P)&=\int_\Omega dP(\omega)\,\log\Big(|A|^{|\Lambda|}P([\omega_{|\Lambda}])\Big)=|\Lambda|\log|A|+\int_\Omega dP(\omega)\,\log P([\omega_{|\Lambda}])\\&=|\Lambda|\log|A|+\sum_{\sigma\in A^\Lambda}P([\sigma])\log P([\sigma])=|\Lambda|\log|A|-H_P(\pi^\Lambda).\end{aligned} \tag{10.158}$$

This is just (10.153) and closes our intermezzo on Kolmogorov's entropy, although the following lemma (on our way to the thermodynamic limit) is similar to Kolmogorov's inequality (2.46).[152]

Lemma 10.12 *For $\Omega=A^{\mathbb{Z}^d}$, $P\in\mathrm{Prob}(\Omega)$, any prior $Q=q^{\mathbb{Z}^d}$, and any finite $\Lambda_1,\Lambda_2\subset\mathbb{Z}^d$ one has*

$$D_{\Lambda_1}(P\|Q)+D_{\Lambda_2}(P\|Q)\le D_{\Lambda_1\cup\Lambda_2}(P\|Q)+D_{\Lambda_1\cap\Lambda_2}(P\|Q). \tag{10.159}$$

Exercise 79 *Prove this lemma.*

This lemma (with $\Lambda_1\cap\Lambda_2=\emptyset$) makes the following adaptation of Fekete's lemma relevant:[153]

[152] We follow Georgii (2011), Proposition (15.10).

[153] Lemma 2.2 has the inequality in the opposite direction and has an infimum where Lemma 10.13 has a supremum. Indeed, the former applies to the Kolmogorov–Sinai entropy, which is the negative of the relative entropy to which the latter applies, cf. (10.153).

Lemma 10.13 *Suppose we have numbers $a_\Lambda \in [0,\infty]$ indexed by rectangles $\Lambda \subset \mathbb{Z}^d$, i.e. finite sublattices of the form $\Lambda = \prod_{i=1}^d [m_i, n_i]$ with $m_i, n_i \in \mathbb{Z}$ (and intervals taken in $\mathbb{Z}$), that:*

1. *satisfy $a_{\Lambda_1} + a_{\Lambda_2} \le a_{\Lambda_1+\Lambda_2}$ whenever $\Lambda_1 \cap \Lambda_2 = \emptyset$;*
2. *are translation-invariant in the sense that $a_{\Lambda+x} = a_\Lambda$ for all $x \in \mathbb{Z}^d$.*

Then $\lim_{N\to\infty} a_{\Lambda_N}/|\Lambda_N|$ exists in $[0,\infty]$, and equals $\sup_\Lambda a_\Lambda/|\Lambda|$, taken over all rectangles.[154]

Proof.[155] Take $C < \sup_\Lambda a_\Lambda/|\Lambda|$, so there exists some Λ such that $a_\Lambda/|\Lambda| > C$. For given N, fill Λ_N with as many translates of Λ as possible, say k_N copies, so that

$$\lim_{N\to\infty} |\Lambda_N|/k_N|\Lambda| = 1. \tag{10.160}$$

Using this property, iterating properties 1 and 2, and the property $k_N|\Lambda| \le |\Lambda_N|$ then gives

$$\sup_\Lambda \frac{a_\Lambda}{|\Lambda|} \ge \liminf_{N\to\infty} \frac{a_{\Lambda_N}}{|\Lambda_N|} = \liminf_{N\to\infty} \frac{a_{\Lambda_N}}{k_N|\Lambda_N|} \ge \frac{a_\Lambda}{|\Lambda|} > C. \tag{10.161}$$

Letting $C \to \sup_\Lambda a_\Lambda/|\Lambda|$ gives $\liminf_{N\to\infty} \frac{a_{\Lambda_N}}{|\Lambda_N|} = \sup_\Lambda \frac{a_\Lambda}{|\Lambda|}$, and trivially we also have

$$\limsup_{N\to\infty} \frac{a_{\Lambda_N}}{|\Lambda_N|} \le \sup_\Lambda \frac{a_\Lambda}{|\Lambda|}; \qquad \limsup_{N\to\infty} \frac{a_{\Lambda_N}}{|\Lambda_N|} \ge \liminf_{N\to\infty} \frac{a_{\Lambda_N}}{|\Lambda_N|} = \sup_\Lambda \frac{a_\Lambda}{|\Lambda|}. \tag{10.162}$$

Hence $\liminf_{N\to\infty} \frac{a_{\Lambda_N}}{|\Lambda_N|}$ and $\limsup_{N\to\infty} \frac{a_{\Lambda_N}}{|\Lambda_N|}$ both equal $\sup_\Lambda \frac{a_\Lambda}{|\Lambda|}$; hence they are equal to each other, so that $\lim_{N\to\infty} \frac{a_{\Lambda_N}}{|\Lambda_N|}$ exists and equals their common value $\sup_\Lambda \frac{a_\Lambda}{|\Lambda|}$. □

Proposition 10.14 *For $Q = q^{\mathbb{Z}^d}$ and any translation-invariant $P \in \mathrm{Prob}^{\mathbb{Z}^d}(A^{\mathbb{Z}^d})$ the expression*

$$d_Q(P) := \lim_{N\to\infty} \frac{D_{\Lambda_N}(P\|Q)}{|\Lambda_N|} \tag{10.163}$$

exists and equals the supremum of $D_\Lambda(P\|Q)/|\Lambda|$ over all rectangles Λ. Furthermore:

1. *The function $P \mapsto d_Q(P)$ is lsc and affine.*
2. *For any $c \ge 0$ the level sets $\{P \in \mathrm{Prob}^{\mathbb{Z}^d}(A^{\mathbb{Z}^d}) \mid d_Q(P) \le c$ are compact.*

We call $d_Q(P)$ the *mean relative entropy* (over the lattice $\mathbb{Z}^d$). So far we defined it only for translation-invariant probability measures $P \in \mathrm{Prob}^{\mathbb{Z}^d}(A^{\mathbb{Z}^d})$, but if necessary d_Q may be extended to all $P \in \mathrm{Prob}(A^{\mathbb{Z}^d})$ or even all $P \in C(A^{\mathbb{Z}^d})^*$ by complementing (10.163) with

$$d_Q(P) = \infty; \qquad (P \notin \mathrm{Prob}^{\mathbb{Z}^d}(A^{\mathbb{Z}^d})). \tag{10.164}$$

The existence of the limit in (10.163) follows at once from Lemmas 10.12 and 10.13. The limit function $d_Q(\cdot)$ being lsc follows from the fact each approximant is lsc and the limit being a supremum. This property remains valid if we add (10.164). It is surprising, however, that $d_Q(\cdot)$ is not just *convex*, like its approximants, but even *affine*.

Exercise 80 *Prove concavity of $P \mapsto d_Q(P)$. (i.e. the oppositie inequality to convexity).*

[154] See (10.63) for the definition of the special rectangles Λ_N.

[155] See Rassoul-Agha, page 86, or Georgii (2011), Lemma (15.11).

We omit the technical proof that its level sets are compact.[156] □

The function d_Q appears as the rate function in a version of Sanov's theorem.[157] This originally applied to the case $\Omega = A^{\mathbb{N}}$, cf. Theorem 7.3, and described large fluctuations of the empirical measure (7.1), which lies in $\mathrm{Prob}(A)$. This is easily generalized to $\Omega = A^{\mathbb{Z}^d}$, in which case we put

$$L_N(\omega) = \frac{1}{|\Lambda_N|} \sum_{x \in \Lambda_N} \delta_{\omega(x)}, \tag{10.165}$$

and for a given prior $Q = q^{\mathbb{Z}^d}$ we then have $L_N(\omega) \to q$ for Q-a.e. $\omega \in \Omega$, weakly in $\mathrm{Prob}(A)$. The function D_q is then the rate function for large fluctuations of L_N around its limit value q.

A LDP for probability measures on $A^{\mathbb{Z}^d}$ instead of A starts from the *empirical field*

$$R_N(\omega) := \frac{1}{|\Lambda_N|} \sum_{x \in \Lambda_N} \delta_{\omega \circ \theta_x}, \tag{10.166}$$

where $(\omega \circ \theta_x)_y := \omega_{x+y}$, cf. (10.72), which takes values in $\mathrm{Prob}(\Omega)$. Since the $\mathbb{Z}^d$-action on $\Omega = A^{\mathbb{Z}^d}$ is ergodic, if A is compact (so that Ω is compact in its canonical product topology) the (generalized) ergodic theorem (10.119) then gives weakly for $q^{\mathbb{Z}^d}$-a.e. $\omega \in \Omega$,

$$\lim_{N \to \infty} R_N(\omega) = q^{\mathbb{Z}^d}. \tag{10.167}$$

Unfortunately, $R_N(\omega)$ depends on values of ω outside Λ_N and hence is not $\mathscr{F}_{\Lambda_N}$-measurable. Moreover, it is not translation invariant. Both issues are remedied at one stroke, as follows.

First, restrict ω to Λ_N and then extend $\omega_{|\Lambda_N} : \Lambda_N \to A$ from Λ_N to all of $\mathbb{Z}^d$ by making it periodic: that is, we fill up $\mathbb{Z}^d$ with copies of Λ_N by writing $z = (z_1, \ldots, z_d) \in \mathbb{Z}^d$ uniquely as $z = x + y$ with $x \in \Lambda_N$ and $y_i = k_i \cdot (2N+1)$ for $i = 1, \ldots, d$ and $k_i \in \mathbb{Z}$, and defining

$$\omega_N(z) := \omega(x). \tag{10.168}$$

This only depends on $\omega_{|\Lambda_N}$ and hence gives the duly $\mathscr{F}_{\Lambda_N}$-measurable *periodized empirical field*

$$\tilde{R}_N(\omega) := \frac{1}{|\Lambda_N|} \sum_{x \in \Lambda_N} \delta_{\omega_N \circ \theta_x}. \tag{10.169}$$

By the Portmanteau theorem, $P_N \to P$ weakly in $\mathrm{Prob}(\Omega)$ iff $P_N(B) \to P(B)$ in $\mathbb{R}$ for any cylinder set $B \in \mathscr{F}_\Lambda$ (see footnote 139). Since $|R_N(\omega)(B) - \tilde{R}_N(\omega)(B)| \to 0$,[158] eq. (10.167) gives

$$\lim_{N \to \infty} \tilde{R}_N(\omega) = q^{\mathbb{Z}^d}. \tag{10.170}$$

We are now in a position to state Sanov's theorem for lattices (where A is Polish):[159]

Theorem 10.15 *The periodized empirical field* (10.169) *on* $\Omega = A^{\mathbb{Z}^d}$ *with a prior* $Q = q^{\mathbb{Z}^d} \in \mathrm{Prob}(\Omega)$ *for some* $q \in \mathrm{Prob}(A)$*, satisfies a* LDP *with tight rate function* $I = d_Q$.

[156] See Georgii (2011), Proposition (15.14), or Rassoul-Agha, Proposition 6.8.

[157] We follow Rassoul-Agha, Theorem 6.13. The result is valid for general Polish spaces A (keep finite A in mind).

[158] If $B \in \mathscr{F}_\Lambda$ then, since $\omega(x+y) = \omega_N(x+y)$ whenever $x+y \in \Lambda_N$, for fixed $x \in \Lambda_N$ we have $\delta_{\omega \circ \theta_x}(B) = \delta_{\omega_N \circ \theta_x}(B)$ provided $x+y \in \Lambda_N$ for all $y \in \Lambda$. The points $x \in \Lambda_N$ for which this fails (i.e. for which there is some $y \in \Lambda$ for which $x+y \notin \Lambda_N$) mus lie within a distance $\mathrm{diam}(\Lambda)$ of the boundary of Λ_N and hence the number of such points divided by $|\Lambda_N|$ goes to zero as $N \to \infty$. The ensuing convergence $|R_N(\omega)(B) \to \tilde{R}_N(\omega)(B)| \to 0$ is even uniform in ω.

[159] The ergodic theorem leading to (10.167) and (10.170) require A to be compact, but Theorem 10.15 does not.

We omit the proof, which is very lengthy even by the standards of large deviation theory.[160]

Our next job is the construction of an infinite-volume limit of the pressure for some potential Φ, see Definition 10.4 (we have included translation invariance in the definition).

Proposition 10.16 *For a general prior $Q = q^{\mathbb{Z}^d}$ and $\beta \in \mathbb{R}$, the* pressure *defined by*

$$\pi_q(\beta\Phi) := \lim_{N\to\infty} \frac{1}{|\Lambda_N|} \log Z^\eta_{\Lambda_N,q}(\beta); \tag{10.171}$$

$$Z^\eta_{\Lambda,q}(\beta) := \left\langle e^{-\beta H^\eta_\Lambda} \right\rangle_Q = \sum_{\sigma\in A^\Lambda} q^\Lambda(\sigma) e^{-\beta H^\eta_\Lambda(\sigma)}, \tag{10.172}$$

where H^η_Λ in (10.98)*, exists, is finite, and is independent of the boundary condition $\eta \in \Omega$.*

Here free boundary conditions, with Hamiltonian (10.74), are included. See also (8.6) and (10.99), and (10.79). For a flat prior $q = f$ one usually omits $q^\Lambda(\sigma) = |A|^{-|\Lambda|}$ from (10.172), so that

$$\pi(\beta\Phi) := \lim_{N\to\infty} \frac{1}{|\Lambda_N|} \log \sum_{\sigma\in A^{\Lambda_N}} e^{-\beta H^\eta_{\Lambda_N}(\sigma)} = \pi_f(\beta\Phi) + \log|A|. \tag{10.173}$$

Proof. We prove this proposition for pair potentials with finite range and free boundary conditions; the general case involves some more limiting arguments but is based on the same idea.[161] Thus

$$H_\Lambda(\sigma) = \sum_{x\in\Lambda} \Phi_{\{x\}}(\sigma_x) + \sum_{y\in\Lambda, \|x-y\|\le R} \Phi_{\{x,y\}}(\sigma_x, \sigma_y), \tag{10.174}$$

where $R < \infty$, $\|\cdot\|$ is any norm on $\mathbb{R}^d$ (restricted to $\mathbb{Z}^d$), and $\Phi_{\{x,y\}}$ depends on $x-y$ only (since Φ is translation-invariant). In particular, the sum in (10.73) is finite and the Ising model (10.75) clearly falls within this class. For simplicity we put $\beta = 1$ and take the flat prior, so that we use

$$H_\Lambda = \sum_{M\subset\Lambda} \Phi_M; \qquad Z_\Lambda = \sum_{\sigma\in A^\Lambda} e^{-H_\Lambda(\sigma)}; \qquad \pi(\Phi) = \lim_{N\to\infty} \frac{1}{|\Lambda_N|} \log Z_{\Lambda_N}. \tag{10.175}$$

Take $N \in \mathbb{N}_*$ and the ensuing cube Λ_N, see (10.63), so that $|\Lambda_N| = (2N+1)^d$. Then for any $n \in \mathbb{N}_*$ pick n disjoint cubes $C_k(N) = \prod_{i=1}^d [m_i, n_i]$ $(k = 1, \ldots, n)$, with equal sides $n_i - m_i = 2N+1$ for each $i = 1, \ldots, d$. Their union $\cup_{k=1}^n C_k(N)$ has volume $|\cup_{k=1}^n C_k(N)| = n(2N+1)^d$. Then

$$\lim_{N\to\infty} \left(\frac{1}{|\Lambda_N|} \log Z_{\Lambda_N} - \frac{1}{|\cup_{k=1}^n C_k(N)|} \log Z_{\cup_{k=1}^n C_k(N)} \right) = 0, \tag{10.176}$$

uniformly in n. For $N_1, N_2 \in \mathbb{N}_*$ the region $\Lambda_{N_1 N_2}$ consists of $n_1 = N_2^d$ copies of Λ_{N_1}, suitably translated to become disjoint (and similarly it consists of $n_2 = N_1^d$ copies of Λ_{N_2}). Since convergence in (10.176) is uniform in n, for any $\varepsilon > 0$ can find N_0 such that for all $N_1 \ge N_0$ and any (fixed) N_2,

$$\left| \frac{1}{|\Lambda_{N_1}|} \log Z_{\Lambda_{N_1}} - \frac{1}{|\Lambda_{N_1 N_2}|} \log Z_{\Lambda_{N_1 N_2}} \right| < \varepsilon/2, \tag{10.177}$$

[160] See Georgii (2011), Theorem (15.45) or Rassoul-Agha, Theorem 6.13. We follow the former in stating the theorem just for the periodized empirical field and not for the empirical field (10.166). The latter states it for both R_N and $\tilde{R}_N$ using the same rate function d_Q defined by (10.163) - (10.164), but this is suspicious for R_N, since the samples $R_N(\omega)$ are not translation invariant. Although the limit measure $q^{\mathbb{Z}^d}$ is translation invariant, fluctuations around this limit for finite N need not be, and yet d_Q gives these zero probability because of (10.164). This is only appropriate for $\tilde{R}_N$, since each sample $\tilde{R}_N(\omega)$ is translation invariant and hence $\tilde{R}_N$ cannot fluctuate into non-translation invariant measures.

[161] We follow Dorlas (2021), based on unpublished lectures by N.M. Hugenholtz (in which however the factor R^d in (10.182) is missing). See Ruelle (2004), Theorem 3.4, Georgii (2011), Theorem (15.30), or Rassoul-Agha, Proposition 6.14, for the general case. The special case of the Ising model is also instructive, see Friedli & Velenik (2018), §3.2.1.

and likewise with N_1 and N_2 swapped. By a standard $\varepsilon/2$ argument, for $N_1, N_2 \geq N_0$,

$$\left| \frac{1}{|\Lambda_{N_1}|} \log Z_{\Lambda_{N_1}} - \frac{1}{|\Lambda_{N_2}|} \log Z_{\Lambda_{N_2}} \right| < \varepsilon. \tag{10.178}$$

Exercise 81 *Prove this.*

Hence $(|\Lambda_N|^{-1} \log Z_{\Lambda_N})_N$ is a Cauchy sequence in $\mathbb{R}$, which must converge, namely to $\pi(\Phi)$.

It remains to prove (10.176). The idea is that the boundary interactions between the cubes in the union $\cup_{k=1}^n C_k(N)$ divided by the volume vanish in the limit, whilst the interactions within the cubes cancel the first term in (10.176). To make this precise, we decompose the energy as

$$H_{\cup_{k=1}^n C_k(N)}(\sigma) = \sum_{k=1}^n H_{C_k(N)}(\sigma) + \sum_{k,l=1, l\neq k}^n I_{kl}(\sigma); \tag{10.179}$$

$$I_{kl}(\sigma) = \sum_{x\in C_k(N), y\in C_l(N), \|x-y\|\leq R} \Phi_{\{x,y\}}(\sigma_x, \sigma_y). \tag{10.180}$$

Recalling that $\Phi_{\{x,y\}}$ depends on $\|x-y\|$ only, the finite expression

$$M(x) := \sum_{y\in\mathbb{Z}^d, \|x-y\|\leq R} \max_{\sigma_x\sigma_y\in A} \{|\Phi_{\{x,y\}}(\sigma_x, \sigma_y)|\} \tag{10.181}$$

in fact does not depend on x, so $M(x) = M$. For fixed k, now count the number of pairs (x,y) for which $x \in C_k(N)$ and $\|x-y\| \leq R$, so that certainly all $y \in C_l(N)$, $l \neq k$ with $\|x-y\| \leq R$ are included, as required by (10.179) - (10.180). For simplicity assume $R \in \mathbb{N}$. For any cube C_L whose sides have length $L \gg R$ there are at most $L^d - (L-2R)^d$ points $x \in C_L$ within a distance R from the boundary of C_L; the others cannot interact with points outside C_L. Each of the former points can interact with at most R^d points $y \notin C_L$. With $L = 2N+1$ in our case, this implies the bound.

$$\left| \sum_{l=1, l\neq k}^n I_{kl}(\sigma) \right| \leq R^d (L^d - (L-2R)^d) M. \tag{10.182}$$

Using the inequality $L^d - (L-2R)^d \leq 2dRL^{d-1}$, this gives

$$\left| \sum_{k,l=1, l\neq k}^n I_{kl}(\sigma) \right| \leq 2ndR^{d+1}(2N+1)^{d-1} M. \tag{10.183}$$

Furthermore, translation invariance gives

$$\sum_{\sigma\in A^{\cup_{k=1}^n C_k(N)}} e^{-\sum_{k=1}^n H_{C_k(N)}(\sigma)} = (Z_{\Lambda_N})^n. \tag{10.184}$$

Exercise 82 *Prove this.*

Combining all of this, and recalling that

$$|\cup_{k=1}^n C_k(N)| = n(2N+1)^d = n|\Lambda_N|, \tag{10.185}$$

we obtain

$$\left| \frac{1}{|\Lambda_N|} \log Z_{\Lambda_N} - \frac{1}{|\cup_{k=1}^n C_k(N)|} \log Z_{\cup_{k=1}^n C_k(N)} \right| \leq \frac{2dR^{d+1}M}{2N+1}, \tag{10.186}$$

since the n in (10.183) cancels the n in (10.185) whilst the $(2N+1)^{d-1}$ in (10.183) combines with the $(2N+1)^d$ in (10.185) to leave the factor $(2N+1)^{-1}$ in (10.186). This proves (10.176). □

Having infinite-volume versions of the entropy and pressure, respectively, we also seek an infinite-volume version of the Fenchel duality (7.22) - (7.23). The main issue is to figure out the infinite-volume analogue of the term $\langle f\rangle_P$ in (7.23), and its relationship to Φ. To this end, define

$$f_\Phi : \Omega \to \mathbb{R}; \qquad\qquad f_\Phi := \sum_{0\in M\subset\mathbb{Z}^d} \frac{\Phi_M}{|M|}, \tag{10.187}$$

where M is finite. For example, for the Ising model (10.75) we have

$$f_\Phi(\sigma) = -J \sum_{x\in\mathbb{Z}^d, \|x\|_1=1} \sigma_x\sigma_0 - B\sigma_0. \tag{10.188}$$

For any $P \in \mathrm{Prob}^{\mathbb{Z}^d}(A^{\mathbb{Z}^d})$ and boundary condition $\eta \in \Omega$ (including $\eta =$ free) we then have:[162]

$$\langle f_\Phi\rangle_P = \lim_{N\to\infty} \frac{1}{|\Lambda_N|} \langle H^\eta_{\Lambda_N}\rangle_P, \tag{10.189}$$

Thus f_Φ is the average energy per site, as already suggested by the definition. Confusingly, this object leads some authors to define the local Hamiltonians with free boundary conditions by

$$H'_\Lambda := \sum_{x\in\Lambda} f_\Phi \circ \theta_x, \tag{10.190}$$

cf. (10.72), which may not coincide with (10.74), although for the Ising model it does. However,[163]

$$\lim_{N\to\infty} \frac{1}{|\Lambda|} \|H^\eta_{\Lambda_N} - H'_{\Lambda_N}\|_\infty = 0, \tag{10.191}$$

and hence in studying the energy as $N\to\infty$ one may use any expression; this is the key to (10.189).

Similarly, using f_Φ instead of Φ gives rise to an alternative versions of the pressure, namely

$$\pi'_q(f) := \lim_{N\to\infty} \frac{1}{|\Lambda_N|} \log Z'_{\Lambda_N,q}(f); \tag{10.192}$$

$$Z'_{\Lambda,q}(f) := \left\langle e^{\sum_{x\in\Lambda} f\circ\theta_x} \right\rangle_Q = \sum_{\sigma\in A^\Lambda} q^\Lambda(\sigma) e^{\sum_{x\in\Lambda} f\circ\theta_x(\sigma)}. \tag{10.193}$$

where $f \in C_b(\Omega)$ (i.e. $f\in C(\Omega)$ if Ω is compact). Fortunately, it can be shown that

$$\pi_q(\Phi) = \pi'_q(-f_\Phi). \tag{10.194}$$

We return to Fenchel duality.[164] In terms of the pressure π'_q, a version close to (7.22) - (7.23) is

$$d_Q(P) = \sup_{f\in C(\Omega)} \{\langle f\rangle_P - \pi'_q(f)\}; \tag{10.195}$$

$$\pi'_q(f) = \sup_{P\in\mathrm{Prob}^{\mathbb{Z}^d}(\Omega)} \{\langle f\rangle_P - d_Q(P)\}, \tag{10.196}$$

[162] See Georgii (2011), Theorem (15.23). This is even true for sequences (η_N) of boundary conditions.

[163] See Rassoul-Agha, Lemma 8.2, for a more detailed statement and proof of (10.191).

[164] Here we assume that A is finite, in which case $f:\Omega\to\mathbb{R}$ is continuous iff it is *quasi-local*, i.e., a sup-norm-limit of *local functions*, which in turn are defined as the functions f that are $\mathscr{F}_\Lambda$-measurable for some finite $\Lambda\subset\mathbb{Z}^d$, which is the case if $f=\pi^*_\Lambda f_\Lambda$ for some $f_\Lambda:\Omega_\Lambda\to\mathbb{R}$ (so that $f(\omega)$ only depends on $\omega_{|\Lambda}$); local functions are continuous.

where recall that $d_Q(P)$ has only been defined for translation-invariant probabilities $P \in \mathrm{Prob}^{\mathbb{Z}^d}(\Omega)$, cf. Proposition 10.14, where evidently also $Q = q^{\mathbb{Z}^d} \in \mathrm{Prob}^{\mathbb{Z}^d}(\Omega)$ for any prior $q \in \mathrm{Prob}(A)$. Hence

$$\pi_q(\beta\Phi) = -\inf_{P \in \mathrm{Prob}^{\mathbb{Z}^d}(\Omega)} \{\beta \langle f_\Phi \rangle_P + d_Q(P)\}; \tag{10.197}$$

or, in terms of the entropy $s(P) := -d_{f^{\mathbb{Z}^d}}(P)$ and the free energy $f(T) = -T\pi_f(\Phi/T)$ and $T > 0$,

$$f(T) = \inf_{P \in \mathrm{Prob}^{\mathbb{Z}^d}(\Omega)} \{\langle f_\Phi \rangle_P - Ts(P)\}. \tag{10.198}$$

Proof of (10.195) - (10.196). These imply all other versions. The key is Corollary 9.6, in which we replace $\mathbb{N}$ by $\mathbb{Z}^d$ and hence N by $|\Lambda_N|$ as appropriate, and take $\mathscr{X} = C(\Omega)^*$ with closed convex subspace $\mathrm{Prob}^{\mathbb{Z}^d}(\Omega)$, and $\mathscr{Y} = C(\Omega)$. The underlying LDP comes from Theorem 10.15 and we do not need to assume (9.27) since we already proved that the pressure exists. Moreover, we note that

$$\sum_{x \in \Lambda_N} f \circ \theta_x = |\Lambda_N| \langle f \rangle_{R_N}, \tag{10.199}$$

so that the partition function (10.193) may be rewritten accordingly:

$$Z'_{\Lambda_N,q}(f) = \left\langle e^{\sum_{x \in \Lambda_N} f \circ \theta_x} \right\rangle_{q^{\mathbb{Z}^d}} = \left\langle e^{|\Lambda_N| \langle f \rangle_{R_N}} \right\rangle_{q^{\Lambda_N}} = \int_{\mathrm{Prob}(\Omega)} dP_N(x) e^{|\Lambda_N| \langle y,x \rangle}, \tag{10.200}$$

where P_N is the probability measure on $\mathrm{Prob}(\Omega)$ induced by $R_N : \Omega \to \mathrm{Prob}(\Omega)$ and furthermore

$$q^{\mathbb{Z}^d} \in \mathrm{Prob}(\Omega); \qquad y = f_\Phi \in C(\Omega); \qquad x = R_N; \qquad \langle y,x \rangle = \langle y \rangle_x. \tag{10.201}$$

To complete the justification for invoking Corollary 9.6, we resolve the tension between Theorem 10.15 using the periodized empirical field $\tilde{R}_N$, whereas the definition of the pressure (10.172) with (10.199) uses the empirical field R_N. The argument between (10.169) and (10.170) implies that

$$\begin{aligned} \pi'_q(f) &= \lim_{N \to \infty} \frac{1}{|\Lambda_N|} \log \left\langle e^{\sum_{x \in \Lambda} f \circ \theta_x} \right\rangle_Q = \lim_{N \to \infty} \frac{1}{|\Lambda_N|} \log \left\langle e^{|\Lambda_N| \langle f \rangle_{R_N}} \right\rangle_Q \\ &= \lim_{N \to \infty} \frac{1}{|\Lambda_N|} \log \left\langle e^{|\Lambda_N| \langle f \rangle_{\tilde{R}_N}} \right\rangle_Q, \end{aligned} \tag{10.202}$$

so that in (10.200) we may replace R_N by $\tilde{R}_N$, and in its wake replace P_N by the probability measure $\tilde{P}_N$ on $\mathrm{Prob}(\Omega)$ induced by $\tilde{R}_N : \Omega \to \mathrm{Prob}(\Omega)$. Thus (10.195) - (10.196) is a special case of the dual pair $I = \Pi^*$ and $\Pi = I^*$ in Corollary 9.6, with $I = d_Q$ and $\Pi = \pi'_q$. □

10.7 Large deviations and variational principle for Gibbs measures

In this section our goal is to explain the remarkable relationships between the topics in the title; we omit most proofs.[165] To be on the safe side we assume that A is compact and metrizable,[166] and as usual we write $\Omega = A^{\mathbb{Z}^d}$, as well as $\mathrm{Prob}^{\mathbb{Z}^d}(\Omega)$ for the (compact convex) space of translation-invariant probability measure on Ω (in the weak topology). Also, $\mathscr{G}(\beta\Phi)$ is the (convex compact)

[165] The Bibliographical Notes to Chapter 15 of Georgii (2011) contain credits for these results. Briefly, the key variational principle is due to Dobrushin and to Lanford & Ruelle, whereas the large deviation results originate with Lanford, Ellis, and others, including also Georgii (1993) himself. References for this chapter are Keller (1998), Ruelle (2004), Georgii (2011), and Rassoul-Agha (2015).

[166] Most results are valid for more general Polish spaces, or even more generally, see especially Georgii (1993, 2011).

set of all (global) Gibbs states for given potential Φ at inverse temperature β, based on the flat prior $Q = f^{\mathbb{Z}^d}$ (the results are easily generalized to more general priors $Q = q^{\mathbb{Z}^d}$, though). The structure of this set was outlined in Theorem 10.8. In particular, the unique Gibbs measure P_β^{free} from part 1 of Theorem 10.8 is translation invariant and plays the role of a reference Gibbs measure.[167]

The following equivalence of quite different things is one of the main results of the theory.[168]

Theorem 10.17 *Let $P \in \text{Prob}^{\mathbb{Z}^d}(\Omega)$. The following statements are equivalent:*

1. $P \in \mathscr{G}(\beta\Phi)$.
2. *P attains the infimum in* (10.197)*, so that, in terms of the entropy $s(P) := -d_{f^{\mathbb{Z}^d}}(P)$,*
$$\pi_f(\beta\Phi) = -\beta\langle f_\Phi\rangle_P + s(P). \tag{10.203}$$
3. $d(P\|P_\beta^{\text{free}}) = 0$.
4. *$d(P\|P_\beta) = 0$ for all translation-invariant Gibbs measures $P_\beta \in \mathscr{G}^{\mathbb{Z}^d}(\beta\Phi)$.*

This theorem is an infinite-volume version of Proposition 10.5, with the vast difference that in finite volume the (local) Gibbs measure for free boundary conditions is unique; each different boundary condition comes with its own version of Proposition 10.5. In infinite volume (global) Gibbs measures need not be unique, but according to the above theorem they all lead to the same pressure (as we had already seen), and, perhaps more unexpectedly, they all have zero mean relative entropy "distance" to the reference Gibbs measure P_β^{free}. More generally, it can be shown that

$$d(P\|P_\beta) = -s(P) + \beta\langle f_\Phi\rangle_P + \pi_f(\beta\Phi), \tag{10.204}$$

for any $P \in \text{Prob}^{\mathbb{Z}^d}(\Omega)$ and translation-invariant Gibbs measure $P_\beta \in \mathscr{G}(\beta\Phi)$, for some given potential Φ. Since the right-hand side of (10.204) is independent of the choice of $P_\beta \in \mathscr{G}(\beta\Phi)$, the equivalence between statements 2 and 4 in Theorem 10.17 already follows. We also see that

$$d(P'_\beta\|P_\beta) = 0 \tag{10.205}$$

for any two translation-invariant Gibbes measures (as always, defined for the same potential Φ).

Eq. (10.204) follows from the finite-volume case plus limiting arguments.[169] For simplicity we take $P_\beta = P_\beta^{\text{free}}$. Using (10.101), (10.141), (10.144), and (10.74), for finite $\Lambda \subset \mathbb{Z}^d$ we have

$$D_\Lambda(P\|P_\beta^{\text{free}}) = -S_\Lambda(P) + \beta\langle H_\Lambda^{\text{free}}\rangle_{P_\Lambda} + \log Z_\Lambda^{\text{free}}(\beta), \tag{10.206}$$

where for $\sigma \in A^\Lambda$ we put

$$P_\Lambda(\sigma) := P([\sigma]); \qquad S_\Lambda(P) := -\sum_{\sigma\in A^\Lambda} P_\Lambda(\sigma)\log P_\Lambda(\sigma). \tag{10.207}$$

Using (10.189), (10.194), (10.172), and of course Proposition 10.14, we obtain (10.204).

Exercise 83 *Prove* (10.206).

[167] It may or may not be the only element of $\mathscr{G}(\beta\Phi)$.

[168] See Ruelle (2004), Theorem 4.2; Rassoul-Agha, page 127; Georgii (2011), Theorems (15.30) and (15.39).

[169] See Rassoul-Agha, Theorem 8.3.

We now turn to large deviations, first of the periodized empirical field $\tilde{R}_N$, and then of the energy. Theorem 10.15 describes the fluctuations of $\tilde{R}_N$ against a prior (and hence limit value) $Q = q^{\mathbb{Z}^d}$. Continuing the story (10.204), using a Gibbs measure as prior (etc.) similarly gives:[170]

Theorem 10.18 *For any prior $P_\beta \in \mathscr{G}^{\mathbb{Z}^d}(\beta\Phi)$, the periodized empirical field* (10.169) *satisfies a* LDP *with tight rate function $I(P) = d(P\|P_\beta)$; this function is the same for all $P_\beta \in \mathscr{G}^{\mathbb{Z}^d}(\beta\Phi)$.*

The last claim is evident from (10.204), according to which we could equivalently have defined

$$I(P) = -s(P) + \beta\langle f_\Phi\rangle_P + \pi_f(\beta\Phi). \tag{10.208}$$

This shows that whereas for a flat prior $Q = f^{\mathbb{Z}^d}$ the rate function was essentially minus the Shannon entropy $s(P)$, changing the prior to a Gibbs measure $Q = P_\beta$ induces some extra terms that are of course determined by the potential Φ, like P_β itself. The zeroes of $I(P)$ form the compact convex space $\mathscr{G}^{\mathbb{Z}^d}(\beta\Phi)$ of translation-invariant Gibbs measures, which may be degenerate.

Using the contraction principle (Theorem 9.4) we also obtain a LDP for the energy.[171] Compared to Cramér's Theorem 8.1, the random variables S_N in (1.36) and (8.1) are now replaced by

$$h_N := \frac{H^\eta_{\Lambda_N}}{|\Lambda_N|}, \tag{10.209}$$

see (10.98), although because of (10.189) the functions $h_N = f_\Phi$ give the same result.

Theorem 10.19 *For any prior $P_\beta \in \mathscr{G}^{\mathbb{Z}^d}(\beta\Phi)$, the energies (h_N) satisfy a* LDP *with rate function*

$$I_E : \mathbb{R} \to [0,\infty]; \qquad I_E(u) = \inf\{I(P), P \in \mathrm{Prob}^{\mathbb{Z}^d}(\Omega) \mid \langle f_\Phi\rangle_P = u\}, \tag{10.210}$$

with $I_E(u) = \infty$ if no P exists with $\langle f_\Phi\rangle_P = u$. Furthermore, I_E is convex and tight, and we have

$$I_E(u) = \sup_{t\in\mathbb{R}}\{tx - \pi_f((\beta - t)\Phi)\} + \pi_f(\beta\Phi) = -\inf_{t\in\mathbb{R}}\{\pi_f(t\Phi) + tu\} + \beta u + \pi_f(\beta\Phi). \tag{10.211}$$

Finally, the unique zero of I_E lies at $u = \langle f_\Phi\rangle_{P_\beta}$.

This is similar to Proposition 10.2 and Exercise 71, and similar comments apply: we may rewrite

$$I_E(u) = s_\beta(u) - s_{\mathrm{eq}}(u); \qquad s_\beta(u) = s_{\mathrm{eq}}(u(\beta)) + \beta(u - u(\beta)), \tag{10.212}$$

where $s_{\mathrm{eq}}(u)$ is the first term on the right-hand side of (10.211), identified with the equilibrium entropy at a temperature $\beta(u)$ uniquely determined by u via

$$\langle f_\Phi\rangle_{P_{\beta(u)}} = u, \tag{10.213}$$

whilst

$$s_{\mathrm{eq}}(u(\beta)) = \pi_f(\beta\Phi) + \beta\langle f_\Phi\rangle_{P_\beta}, \tag{10.214}$$

cf. (10.203), in which we write $\langle f_\Phi\rangle_{P_\beta} = u(\beta)$; in contrast with (10.213), this time the energy $u(\beta)$ is determined by the inverse temperature β, rather than the other way round. Thus

$$I_E(u(\beta)) = 0, \tag{10.215}$$

[170] See Georgii (2011), Theorem (15.45), for a proof.

[171] Theorem 10.19 is a special case of Corollary (15.48) in Georgii (2011).

confirming the last claim in Theorem 10.19. All in all, we may interpret $-I_E = s_{\mathrm{eq}} - s_\beta$ as a nonequilibrium entropy, normalized so that it vanishes at the equilibrium (energy) value of its argument, given the background Gibbs measure at given $T^{-1} = \beta$.

We close this chapter with a very general and powerful large deviation result, due to Kifer (1990). Here $(\Omega, \mathscr{F}, Q)$ is a probability space, X is a compact metric space, and

$$\zeta_N : \Omega \to \mathrm{Prob}(X) \tag{10.216}$$

is a sequence of random variables. Here one may think of the following special cases:

- $\Omega = A^{\mathbb{N}}$, $X = A$, and $\zeta_N = L_N$ (the empirical measure), see (5.1);
- $X = \Omega = A^{\mathbb{Z}^d}$ and $\zeta_N = R_N$, as in (10.166), or $\zeta_N = \tilde{R}_N$ as in (10.169);
- $X = \Omega$ with continuous (or Borel) map $T : X \to X$ and the *occupational measure*, cf. (4.29),

$$\zeta_N(x) := \frac{1}{N} \sum_{n=0}^{N-1} \delta_{T^n x}. \tag{10.217}$$

For any $f \in C(X)$ one then defines the *partition function* and *pressure* by

$$Z_N(f) := \left\langle e^{r_N \langle f \rangle_{\zeta_N}} \right\rangle_Q = \int_\Omega dQ(\omega)\, e^{r_N \int_X d\zeta_N(\omega) f}; \tag{10.218}$$

$$\pi_Q(f) := \lim_{N\to\infty} \frac{1}{r_N} \log Z_N(f), \tag{10.219}$$

where r_N is a suitable rate like $r_N = N$ or $r_N = |\Lambda_N|$, and the existence of the limit must be assumed:

Theorem 10.20 (Kifer) *If the pressure* (10.219) *exists and is finite for each $f \in C(X)$, then:*

1. *The map $f \mapsto \pi(f)$ is convex and continuous from $C(X)$ with sup-norm to $\mathbb{R}$, and gives rise to a Fenchel dual pair (in which the conjugate $I_Q : \mathrm{Prob}(X) \to [0,\infty]$ is convex and lsc):*

$$I_Q(P) = \sup_{f \in C(X)} \{\langle f \rangle_P - \pi_Q(f)\}; \tag{10.220}$$

$$\pi_Q(f) = \sup_{P \in \mathrm{Prob}(X)} \{\langle f \rangle_P - I_Q(P)\}. \tag{10.221}$$

2. *The* LDP *upper bound* (9.1) *holds with rate function* (10.220)*, that is, for closed $F \subset \mathrm{Prob}(X)$,*

$$\limsup_{N\to\infty} \frac{1}{r_N} \log Q(\zeta_N \in F) \leq -I(F). \tag{10.222}$$

3. *Under further assumptions the corresponding lower* LDP *bound* (9.2) *also holds.*[172]

Proof. We just prove the fairly easy upper bound.[173] The case $I(F) \leq 0$ is trivial, since the left-hand side of (10.222) is ≤ 0. Suppose $0 < I(F) < \infty$. Pick $\varepsilon > 0$. Then

$$F \subset \{P \in \mathrm{Prob}(X) \mid I_Q(P) > I_Q(F) - \varepsilon\} = \bigcup_{f \in C(X)} \Gamma_\varepsilon(f). \tag{10.223}$$

[172] These assumptions are: there exists a countable set of functions (f_n) in $C(X)$ with dense linear span S, each with $\|f_n\|_\infty = 1$, such that any element $f \in S$ gives rise to a unique $P \in \mathrm{Prob}(X)$ for which the supremum in (10.221) is attained (such a probability measure may be called an *equilibrium state* for f).

[173] We follow Kifer (1990). His proof of the lower bound is prohibitively difficult and unsuitable for a course.

Exercise 84 *Prove this.*

Hence the open sets $\Gamma_\varepsilon(f)$ cover F, and since F (being a closed subset of the weakly compact set $\mathrm{Prob}(X)$) is compact, it has a finite subcover $(\Gamma_\varepsilon(f_a))_{a\in A}$. Using (10.223) and (8.76) gives

$$\begin{aligned} Q(\zeta_N \in F) &\le \sum_{a\in A} Q(Z_n \in \Gamma_\varepsilon(f_a)) = \sum_{a\in A} Q(\{\zeta_N \in \mathrm{Prob}(X) \mid \langle f_a\rangle_{\zeta_N} > \pi_Q(f_a) + I_Q(F) - \varepsilon\}) \\ &\le \sum_{a\in A} e^{-r_N(\pi_Q(f_a)+I_Q(F)-\varepsilon)} \langle e^{r_N\langle f_a\rangle_{\zeta_N}}\rangle_Q = e^{-r_N(I_Q(F)-\varepsilon)} \sum_{a\in A} e^{-r_N\pi_Q(f_a)} Z_N(f_a). \quad (10.224) \end{aligned}$$

Taking $(1/r_N)$ times the logarithm of this inequality and taking $\limsup_{N\to\infty}$ makes the very last two terms cancel in view of (10.221), leaving $-(I_Q(F)-\varepsilon)$. Letting $\varepsilon \to 0$ then gives (10.222).

The case $I(F) = \infty$ is similar; instead of $\Gamma_\varepsilon(f)$ we now take

$$\Gamma^N(f) := \{P \in \mathrm{Prob}(X) \mid \langle f\rangle_P - \pi_Q(f) > N\}, \quad (10.225)$$

and likewise show that $F \subset \bigcup_{f\in C(X)} \Gamma^N(f)$. The same argument gives $-N$ instead of $-(I_Q(F)-\varepsilon)$, and so letting $N \to \infty$ gives $\limsup_{N\to\infty} \frac{1}{r_N} \log Q(\zeta_N \in F) \le -\infty$, hence $= -\infty$. □

11 Thermodynamic formalism

The above theory is a special case of what is called *thermodynamic formalism*. The goal of this formalism is to unify statistical mechanics with the theory of dynamical systems, especially chaotic ones. We already saw some such connections, especially via Kolmogorov's approach to dynamical systems and the ensuing concept of (Kolmogorov–Sinai) entropy, but the thermodynamic formalism takes this a step further by defining entropy and pressure in a topological (as opposed to a measure-theoretic) context. As in the case of statistical mechanics, these gives rise to a concept equilibrium states via a variational principle.[174] Although the theory works for general $\mathbb{Z}^d$-actions, we here restrict ourselves to $d = 1$ and hence assume we have a topological space X, assumed to be compact and metrizable to be on the safe side,[175] and a continuous map $T : X \to X$. This map gives rise to a (semigroup) action $\mathbb{N} \times X \to X$ by $(n,x) = T^n x$; if T is invertible (with continuous inverse) this gives a (group) action $\mathbb{Z} \times X \to X$, since in that case we may also take $n < 0$ (of course, for subsets $B \subset X$ the set $T^{-n}B$ is defined also for non-invertible T).

The first object is the *topological entropy*. This is defined like the Kolmogorov–Sinai entropy (2.57), except that we now start with a finite *open cover* $X = \bigcup_a U_a$ by sets

$$\alpha = \{U_a, a \in A\} \quad (11.1)$$

instead of a finite *measurable partition* of X; and, for any open cover $\gamma = \{C_i, i \in I\}$ of X (so that $X = \bigcup_i C_i$), in the absence of some given probability measure P, simply assign the same probability

$$P(C_i) = 1/|\gamma| \quad (11.2)$$

[174] Canonical texts include Keller (1998) and Ruelle (2004). Viana & Oliveira (2016) also contains an introduction. The setting of statistical mechanics is recovered by taking $X = \Omega = A^{\mathbb{Z}^d}$, on which $\mathbb{Z}^d$ acts by translation, cf. (10.114).

[175] Some technical arguments also require T to be *expansive*, in the sense that there exists $\delta > 0$ such that if $d(T^n x, T^n y) \le \delta$ for all n, then $x = y$, where d is a metric on X (this is a step towards chaos).

to each element C_i of the cover.[176] Thus parallel to (2.35), (2.49), and (2.57) we put

$$H(\gamma) := -\sum_{C\in\gamma} P(C)\log P(C) = \log|\gamma|; \tag{11.3}$$

$$h(\alpha) := \lim_{N\to\infty}\frac{1}{N}H(\alpha^N); \qquad\qquad h(X,T) := \sup_{\alpha} h(\alpha), \tag{11.4}$$

where for given initial open cover γ the refinements γ^N are defined as in (2.43), and the supremum is taken over all finite covers of X (the arguments for existence of the limit and the supremum are practically the same as for the measure-theoretical case). The first key result, then, is

$$h(X,T) = \sup_{P\in\mathrm{Prob}(X)} h(P,X,T), \tag{11.5}$$

where $h(P,X,T)$ is the Kolmogorov–Sinai entropy (2.57). This may be understood from the idea that the flat probability measure (11.2) maximizes the entropy and hence cannot be improved upon.

We now define the topological pressure. For $f\in C(X)$ and any finite open cover (11.1), put

$$Z_N(f,\alpha) := \inf_{\gamma\subset\alpha^N}\left\{\sum_{C\in\gamma}\sup_{x\in C} e^{\sum_{n=0}^{N-1} f(T^n x)}\right\}; \qquad \pi(f,\alpha) := \lim_{N\to\infty}\frac{1}{N}\log Z_N(f,\alpha), \tag{11.6}$$

where the infimum is taken over finite subcovers γ of α^N. It then turns out that every sequence (α_n) of open covers whose diameter goes to zero as $n\to\infty$ has the same limit,[177] so that we put

$$\pi(X,T,f) := \lim_{n\to\infty}\pi(f,\alpha_n). \tag{11.7}$$

This defines the *topological pressure* of X (as a topological space) relative to $f\in C(X)$. Note that

$$\pi(X,T,0) = h(X,T), \tag{11.8}$$

since $Z_N(0,\alpha) = |\alpha^N|$. Topological pressure and minus metric entropy are then Fenchel dual:

$$\pi(X,T,f) = \sup_{P\in\mathrm{Prob}^T(X)}\{\langle f\rangle_P + h(X,P,T)\}; \tag{11.9}$$

$$h(X,P,T) = -\sup_{f\in C(X)}\{\langle f\rangle_P - \pi(X,T,f)\} = \inf_{f\in C(X)}\{\pi(X,T,f) - \langle f\rangle_P\}, \tag{11.10}$$

provided $P\mapsto h(X,P,T)$ is finite and usc on $P\in\mathrm{Prob}^T(X)$. In (11.9) one may equivalently take the supremum over the ergodic T-invariant probability measures only. Combined with (11.8), eq. (11.5) is the special case $f=0$ of (11.9). As in statistical mechanics, a measure $P\in\mathrm{Prob}^T(X)$ that attains the supremum in (11.9) is called an *equilibrium state*. Such states exist because X and hence $\mathrm{Prob}(X)$ are compact; using some further assumptions one may even define Gibbs states. Kifer's Theorem 10.20 provides a large deviations perspective on the duality (11.9) -(11.10).

The example of a Markov chain in §10.5 also suggests vast generalizations.[178] Any stochastic process indexed by $\mathbb{N}$ or $\mathbb{Z}$ with a finite state space A, seen in the Kolmogorov representation, defines a generalized one-dimensional spin system with local "Hamiltonians"

$$H_F(\sigma) = -\log p_F(\sigma), \tag{11.11}$$

[176] This assumes that α has no subcover of smaller cardinality. In general one should pass to such a minimal subcover $\alpha'\subset\alpha$, and define $H(\alpha)=\log|\alpha'|$.

[177] The diameter of a finite cover is the supremum of the diameter of its elements, defined by the metric.

[178] See for example Beck & Schlögl (1993).

cf. (3.16), where $F \subset \mathbb{N}$ (or $\mathbb{Z}$) is finite; if we allow the value $H_F = \infty$ this even makes sense if $p_F(\sigma) = 0$. One may also introduce a parameter $\beta \in \mathbb{R}$ and define new probabilities

$$p_{F,\beta}(\sigma) := \frac{1}{Z_F(\beta)} e^{-\beta H_F(\sigma)}; \qquad Z_F(\beta) := \sum_{\sigma \in A^F} e^{-\beta H_F(\sigma)}. \tag{11.12}$$

Introducing β may be a useful too in studying the probabilities p_F and hence the overall probability $P \in \mathrm{Prob}(A^{\mathbb{N}})$; for example, letting $\beta \to 0$ flattens the probability distribution, whereas for $\beta \to \infty$ only the largest probabilities (corresponding to the smallest values of H_F) survive. One then has the entire formalism of thermodynamics at one's disposal. This strategy turns out to be especially useful in the study of (certain) chaotic dynamical systems.

12 Entropy and Kolmogorov randomness

Since entropy is closely related to randomness, it is no surprise that the most sophisticated concept of randomness known to date, namely *Kolmogorov randomness*, aka *algorithmic randomness*, is related to entropy.[179] Kolmogorov's problem (of which his new notion of randomness was supposed to be a solution), which was noticed already by Laplace and perhaps even earlier probabilists, was that, specializing to a 50-50 Bernoulli process for simplicity, any binary string σ of length N has probability $P(\sigma) = 2^{-N}$ and any (infinite) binary sequence x has probability $P(x) = 0$, although say $\sigma = 0011010101110100$ looks much more random than $\sigma = 111111111111111$. In other words, their *probabilities* say little or nothing about the *randomness* of individual outcomes.

Imposing statistical properties helps but is not enough to guarantee randomness. We first explain this in base 10. We call a real number $x \in [0,1]$ *Borel normal* if in its decimal expansion

$$x = \lim_{N\to\infty} \sum_{n=0}^{N} s_n 10^{-n-1}, \tag{12.1}$$

where $s_n \in 10 = \{0,1,\ldots,9\}$, and hence $s \in 10^{\mathbb{N}}$, each decimal string $\sigma \in 10^*$ has (asymptotic) frequency $10^{-\ell(\sigma)}$ in s, so that each digit $0,\ldots,9$ occurs 10% of the time ($\ell(\sigma) = 1$), each block 00 to 99 occurs 1% of the time ($\ell(\sigma) = 2$), etc. [180] More precisely and more generally, in any base $b = \{0,\ldots,b-1\}$, a b-valued sequence $s \in b^{\mathbb{N}}$ is Borel normal if for any b-valued string $\sigma \in b^*$,

$$\lim_{N\to\infty} \frac{1}{N} |\{n \in N \mid s_{|n} \in b^*\sigma\}| = b^{-\ell(\sigma)}, \tag{12.2}$$

and hence $x \in [0,1]$ is Borel normal in base b if the b-valued sequence $s \in b^{\mathbb{N}}$ in its b-ary expansion

$$x = \lim_{N\to\infty} \sum_{n=0}^{N} s_n b^{-n-1} \tag{12.3}$$

[179] This section is mainly based on Cover & Thomas (2006), chapter 14. The standard reference for algorithmic randomness of stringsis Li & Vitányi, (2008), of which §8.6 discusses some connections with entropy. See also Grünwald & Vitányi (2003). Algorithmic randomness of sequences is the main topic of Calude (2002) and Downey & Hirschfeldt (2010), but this is a matter of emphasis: all three books discuss both strings and sequences (we repeat our convention that a *string* σ is finite row of bits or letters from some finite alphabet, whereas a *sequence* s is an infinite one). For first introductions Grünwald & Vitányi (2008) and Dasgupta (2011) are recommended. The pointwise results in the main text are more advanced; see Towsner (2020) for a survey and original references. Franklin & Porter (2020) is a recent survey of various aspects of algorithmic randomness. The place of algorithmic randomness in a broader spectrum of theories of randomness is discussed for example in Porter (2012), Eagle (2019), and Landsman (2020). For the history of the subject see van Lambalgen (1987), Porter (2012), as well as the 'History and References' sections at the end of each chapter in Li & Vitányi, (2008).

[180] The literature on Borel normality is large. Khoshnevisan (2006) is a nice introduction.

is Borel normal in the above sense (this may depend on b!).[181] For example, for $b = 10$ consider

$$0.12345678910111213141516171819202122232425262 9\ldots$$
$$0.23571131719232931374143475359616771737983899 7\ldots \quad (12.4)$$

The first is *Champernowne's number*, which just lists all positive integers, whereas the second is the *Copeland–Erdös number*, which lists all primes.[182] Both are Borel normal!

Exercise 85 *Prove that Champernowne's number is Borel normal.*

Once you know the pattern behind these numbers, it is obvious how they continue. What about

$$0.30927562832084531584652001027797235612923012605863\ldots? \quad (12.5)$$

This looks really random, but these are the first few digits of π after the first million ones! Without this information, no one would recognize their origin. Unlike the previous two numbers, the decimal expansion of π is merely *conjectured* to be Borel normal, but this has been empirically verified in billions of decimals, so let us assume it is. Thus Borel randomness cannot be a good notion of randomness, since all three numbers can be generated by short formulae or algorithms, and on any reasonable intuition about randomness, this means that they can't be random. We may also argue that these numbers are *computable*, which seems the very opposite to being *random*.

Kolmogorov's paradoxical idea was to define randomness *through* computability! Roughly speaking, his idea was that a string σ is random iff the shortest computer program p that computes σ is about the length of σ itself (in which case p must simply store σ in its memory and print it):

An object O is random iff the shortest computable description of O is O itself.

This idea apparently exceeds the world of binary strings, although it is limited to objects for which one has a notion of computability. Indeed, adding *computable*, or at least some other definition of what is meant by a "description" is essential in view of *Berry's paradox*:

> The Berry number is the smallest positive integer that cannot be described in less than eighteen words.

The paradox, then, is that on the one hand this number must exist, since only finitely many integers can be described in less than eighteen words and hence the set of such numbers must have a lower bound, while on the other hand Berry's number cannot exists by its own definition. This is, of course, one of innumerable paradoxes of natural language (like the liar's paradox).

To make this idea precise we assume basic familiarity with the theory of computation, or at least with the concept of a Turing machine, or at the very least with computers and computer programs. A Turing machine T is nothing but a particular physical model of a computer, which in turn may be seen as a computable (partial) function $f : D_f \to \mathbb{N}$, whose domain $D_f \subset \mathbb{N}$ consists of all $n \in \mathbb{N}$ for which $f(n) \in \mathbb{N}$ is defined. Using some (computable) bijection $\mathbb{N} \cong 2^*$ we may identify $n \in \mathbb{N}$ as the binary code of a computer program p run by T, in which case $n \in D_f$ iff T terminates (or "halts") on the corresponding program p (written $T(p)\!\downarrow$) and produces $f(n)$ or $T(p)$ as its output; using the same (or some other) bijection also on this end, we may and will consider $T(p)$ to be a binary string, now written as $\sigma \in 2^*$. Thus $T(p) = \sigma$ and we consider the

[181] The b-ary expansion of $x \in [0,1]$ is unique unless $x = p/b^k$ for some $k > 0$ and $p \in \mathbb{N}$ satisfies $0 < p < b^k$. But rationals are not Borel normal anyway, so this lack of uniqueness does not jeopardize the definition.

[182] These numbers are supposed to have an infinite number of digits, so that we precede them by 0. to get a finite real.

Kolmogorov complexity of binary strings.[183] For technical reasons,[184] we assume that the codes p of all programs for which T halts (i.e. produces an output $T(p) = \sigma$) form a prefix set; as before, this means that if $T(p)\downarrow$ and $T(q)\downarrow$ for $p, q \in 2^*$, then p cannot be a prefix of q (or *vice versa*).

As in the usual case, one can prove that there are universal prefix Turing machines U, in the sense that for any prefix Turing machine T_n (based on some computable enumeration (T_n) of all prefix Turing machines) one has $T_n(p) = U(\langle n, p\rangle)$, where $\langle\cdot,\cdot,\rangle$ is some computable map from $\mathbb{N} \times 2^*$ into a prefix subset of 2^*. We often omit "prefix" in what follows, but it is always meant.

Definition 12.1 *Let U be a universal prefix Turing machine.*

1. *The* Kolmogorov complexity *of $\sigma \in 2^*$ (relative to U) is given by*

$$K_U(\sigma) := \min\{\ell(p), U(p) = \sigma\} \tag{12.6}$$

2. *The* conditional Kolmogorov complexity *of $\sigma \in 2^*$ (relative to U) is given by*

$$K_U(\sigma \mid \ell(\sigma)) := \min\{\ell(p), U(\langle \ell(\sigma), p\rangle) = \sigma\}. \tag{12.7}$$

How do these definitions depend on the choice of the universal prefix Turing machine U?

Proposition 12.2 *For any two universal prefix Turing machines U, V there is a constant $C(U,V) \in \mathbb{N}$ such that for all $\sigma \in 2^*$ we have*

$$|K_U(\sigma) - K_V(\sigma)| < C(U,V), \tag{12.8}$$

and similarly (for a different constant for the conditional Kolmogorov complexity.

Proof (sketch). First assume V is merely a (prefix) Turing machine (so not necessarily universal, like U). Then any program p_V for V that produces σ, i.e., $V(p_V) = \sigma$, can be transferred to U via a program $q_{U,V}$ on U that simulates V, so that $p_U = q(U,V)p_V$ with length $\ell(p_U) = \ell(q_{U,V}) + \ell(p_V)$ produces σ via U. Hence $K_U(\sigma) \leq K_V(\sigma) + \ell(q_{U,V})$ uniformly in σ. If V is also universal this works the other way round, too, and one obtains (12.8) with $C(U,V) = q_{U,V} + q_{V,U}$. □

In view of this it is common practice to omit the label U and just talk about *the* Kolmogorov complexity $K(\sigma)$ of strings σ, which, then, is defined up to a σ-independent constant. As such, we may define a string to be *c-random* (relative to U) for some $c \in \mathbb{N}$ if

$$K_U(\sigma) \geq \ell(\sigma) - c. \tag{12.9}$$

This is a useful criterion for *strings* whose length far exceed the U-dependent constant c, and hence it enables us to define randomness of *sequences* independently of U:

Definition 12.3 *A sequence $s \in 2^{\mathbb{N}}$ is* random *if there is a constant $c \in \mathbb{N}$ such that for all $N > 1$,*

$$K(s_{|N}) \geq N - c. \tag{12.10}$$

More precisely, one should say that for any universal prefix Turing machine U there is a constant $c(U)$ such that $K_U(s_{|N}) \geq N - c(U)$, but the point is that although the (minimum) *value* of constant $c(U)$ does depend on U, by Proposition 12.2 the *existence* of such a constant does not. Definition 12.3 (and its reformulation by Martin-Löf that we shall not discuss here) forms the basis of the theory of algorithmic randomness. To get an idea about this definition, as well as about Kolmogorov complexity in general, we analyze the case of strings in some more detail.

[183] One may equally well look at integers or strings over some finite alphabet A.

[184] Namely: (1) a good relationship with coding theory, and (2): a smooth extension of the theory from strings to sequences. Briefly, if $T(p) = \sigma$ we may regard p as a code-word for σ, i.e. $C(\sigma) = p$, which is decoded by T, and so a prefix Turing machine produces a uniquely decodable code.

Proposition 12.4 *There are constants c, c' such that, for all $\sigma \in 2^*$,*

$$K(\sigma) \leq \ell(\sigma) + 2\log_2 \ell(\sigma) + c; \tag{12.11}$$

$$K(\sigma \mid \ell(\sigma)) \leq \ell(\sigma) + c'. \tag{12.12}$$

Proof (sketch). In the worst case, storing σ costs $\ell(\sigma)$ bits. This is preceded by some σ-independent general printing instruction of length c. But the set of all $\sigma \in 2^*$ should be turned into a prefix set, for otherwise the programs p thus constructed do not form a prefix. A simple (but expensive) way to accomplish this is to double each digit in σ and close with 01 (for example, $\sigma = 1010$ becomes 1100110001), which is then read as a stop sign. This doubling act costs $2\log \ell(x)$ extra bits, which, absorbing the extra two bits 01 in c, gives (12.11).

If $\ell(\sigma)$ is given to U, we could use the same map $\langle \cdot, \cdot, \rangle : \mathbb{N} \times 2^* \to 2^*$ (whose image is a prefix set) that was mentioned before Definition 12.1; this would effectively incorporate $\ell(\sigma)$ into the printing instruction. Hence the need for the above move disappears.[185] □

Combining (12.9), (12.10), and (12.11), we see that s is random if $K(s_{|N}) \approx N$ as $N \to \infty$, and more precisely,[186] one can show that s is random iff

$$N + O(1) \leq K(s_{|N}) \leq N + K(N) + O(1), \tag{12.13}$$

where $K(N)$ is defined as in (12.6) but now with $N \in \mathbb{N}$ instead of $\sigma \in 2^*$ as the output. Like $K(\sigma)$, this makes sense up to a U-dependent but N-independent constant, which disappears into the $O(1)$ terms. Conistency with (12.11) comes from the estimate $K(N) \leq 2\log_2 N + c$, which is obvious: if the program p is given the length $\ell(N) \approx \log_2 N$ of the binary expansion N then at worst it needs to store these bits (giving it a length $\log_2 N$), and if not, providing $\ell(N)$ takes another $\log_2 N$ bits. In fact, the upper bound in (12.13) "wins",[187] since it can also be shown that s is random iff:

$$\lim_{N \to \infty} (K(s_{|N}) - N) = +\infty. \tag{12.14}$$

Similarly, a string σ is random if $K(\sigma) \approx \ell(\sigma)$, at least for long strings. On the other hand, long strings that are computable from short programs have $K(\sigma) \approx \log_2(\ell(\sigma))$, since the length of σ plus some short instructions are sufficient for some efficient program p to compute it. Also,

$$|\{\sigma \in 2^* \mid K(\sigma) < k\}| < 2^k, \tag{12.15}$$

since we may simply count the number of programs p of length $\ell(p) < k$ to be $\sum_{n=0}^{k-1} 2^n = 2^k - 1$. On the other hand, the total number of strings σ of length $\ell(\sigma) \leq k$ is $2^{k+1} - 1$ (this is just the previous number plus 2^k), and hence only a fraction $\frac{2^k - 1}{2^{k+1} - 1} < 1/2$ (which for large k of course approaches $1/2$) of these strings can be described in a way that saves just one bit! More generally, the fraction of strings of length $\ell(\sigma) \leq k$ that can be described saving at least n bits is $\frac{2^{k-n+1}}{2^{k+1} - 1} < 2^{-n}$.

Hence the large majority of strings is random, and the same is true for sequences:[188]

Theorem 12.5 (Martin-Löf) *Almost every sequence $s \in 2^{\mathbb{N}}$ is random with respect to $f^{\mathbb{N}}$.*

[185] Using a more efficient (recursive) coding of $\ell(\sigma)$, one may improve the term $2\log_2 \ell(\sigma)$ to $\log_2^* \ell(\sigma)$, where, for $n \in \mathbb{N}$, one defines $\log_2^* n = \log_2 n + \log_2 \log_2 n + \log_2 \log_2 \log_2 n + \cdots$, where the last term is the last possible *positive* term (e.g. $\log_2^* 7$ has the above three terms); but this makes no difference to the relevant asymptotics.

[186] See Li & Vitányi, (2008), page 220.

[187] See Calude (2002), Theorem 6.38 (attributed to Chaitin).

[188] See Calude (2002), Theorem 6.31. This is proved using Martin-Löf's own reformulation of randomness.

Here f is the flat prior on $A = 2$ and $f^{\mathbb{N}}$ is the associated Bernoulli probability measure on $2^{\mathbb{N}}$. If we define $x \in [0,1]$ to be random iff the sequence s in its binary expansion (2.10) is random, it follows from Exercise 14 that almost every $x \in [0,1]$ is random with respect to Lebesgue measure. Borel already proved this for Borel normality in 1909, and Theorem 12.5 implies his result:[189]

Proposition 12.6 *1. A random sequence is Borel normal in any base.*

2. A random sequence contains any string infinitely often.

3. A random sequence s satisfies the strong law of large numbers, in the sense that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} s_n = 1. \tag{12.16}$$

The second part follows from the first, as is easily proved by contradiction. So does the third, but we will prove it directly after we have introduced the appropriate technique. In any case, we see a considerable improvement over the usual strong law of large numbers (SLLN): whereas the latter merely states that (12.16) is true for $f^{\mathbb{N}}$-almost every sequence $2 \in 2^{\mathbb{N}}$, Proposition 12.6.3 *explicitly identify* sequences for which (12.16) holds, namely the random ones; and Theorem 12.5 adds that, consistent with the usual SLLN, these form a subset R of $2^{\mathbb{N}}$ of probability $f^{\mathbb{N}}(R) = 1$.

On the other hand one may wonder how "explicit" this identification is:[190]

Theorem 12.7 (Chaitin) *If $s \in \underline{2}^{\mathbb{N}}$ is random, then any consistent and sufficiently comprehensive mathematical theory T (like ZFC) can compute only finite many digits of s.*

This excludes defining a random number by somehow listing its digits, but some can be described by a formula. The most famous example is Chaitin's number Ω, or more precisely Ω_U,[191] which is the halting probability of some fixed universal prefix Turing machine U, given by

$$\Omega_U := \sum_{\sigma \in 2^* | U(\sigma)\downarrow} 2^{-\ell(\sigma)}. \tag{12.17}$$

There is also an analogue of this theorem for random strings:[192]

Theorem 12.8 *For any consistent and sufficiently comprehensive mathematical theory T (like ZFC) there is a constant $C \in \mathbb{N}$ such that T cannot prove any sentence of the form $K(\sigma) > C$ (although infinitely many such sentences are true), and as such T can only prove (Kolmogorov) randomness of finitely many strings (although infinitely many strings* are *in fact random).*

[189] For details and proofs see Calude (2002), Corollary 6.32 in §6.3 and almost all of §6.4.

[190] More precisely, if $s \in \underline{2}^{\mathbb{N}}$ is random, only finitely many true statements of the form: 'the n'th bit s_n of s equals its actual value' (i.e. 0 or 1) are provable in T. See Calude (2002), Theorem 8.7, which is stated for Chaitin's Ω but whose proof holds for any random sequence. As in Gödel's theorems, one also assumes that T is formalized as an axiomatic-deductive system in which proofs could in principle be carried out mechanically by a computer. See Chaitin (1987) for his own presentation and analysis of his two incompleteness theorems, i.e. our theorems 12.7 and 12.8. Raatikainen (1998) also gives a detailed presentation of the second theorem, including a critique of Chaitin's ideology.

[191] There exists a U for which not a single digit of Ω_U can be known, see Calude (2002), Theorem 8.11.

[192] The proof is based on the existence of a computably enumerable (c.e.) list $\mathsf{T} = (\tau_1, \tau_2, \ldots)$ of the theorems of T, and on the fact that after Gödelian encoding by numbers, theorems of any given grammatical form can be computably searched for in this list and will eventually be found. In particular, there exists a program p such that $p(n)$ halts iff there exists a string σ for which $K(\sigma) > n$ is a theorem of T. If there is such a theorem the output is $p(n) = \sigma$, where σ appears in the first such theorem of the kind (according to the list T). By definition of $K(\cdot)$, this means that $K(\sigma) \leq |P| + |n|$. Now suppose that no C as in the above statement of the theorem exists. Then there is $n \in \mathbb{N}$ large enough that $n > \ell(p) + |n|$ and there is a string $\sigma \in 2^*$ such that T proves $K(\sigma) > n$. Since T is consistent and hence sound (i.e., it only proves true theorems) this is actually true, which gives a contradiction between $K(\sigma) > n > \ell(p) + |n|$ and $K(\sigma) \leq \ell(p) + |n|$; this contradiction can be made more dramatic by taking n such that $n >> \ell(p) + |n|$. Note that this proof shows that a *proof* in T of $K(\sigma) > n$ (if true) would also *identify* σ.

At last, we now turn to the connection between Kolmogorov complexity and entropy! So far, we only have the resources to understand the following result for $A = 2$, but we state the general result. To this end we do note that for finite sets A one may extend Kolmogorov complexity $K(\sigma)$ to $\sigma \in A^*$, for example by using a computable bijection $A^* \cong 2^*$ (since both are $\cong \mathbb{N}$).

Theorem 12.9 *For any finite set A and prior $p \in \mathrm{Prob}(A)$, and any integer $N > 0$, we have*

$$S_2(p) \leq \frac{1}{N} \sum_{\sigma \in A^N} p^N(\sigma) K(\sigma \mid N) \leq S_2(p) + \frac{(|A|-1)\log_2 N}{N} + O(1/N), \tag{12.18}$$

so that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{\sigma \in A^N} p^N(\sigma) K(\sigma) \equiv \lim_{N\to\infty} \frac{1}{N} \langle K \rangle_{p^N} = S_2(p). \tag{12.19}$$

In other words, for large N the "average" complexity $\langle K \rangle_{p^N}$ of strings in A^N goes like $NS_2(p)$, which for $A = 2$ and the flat prior $p = f$ is just N (since $S_2(f) = 1$), as we already knew.

Eq. (12.19) follows from (12.18) in view of the fact that $K(\sigma \mid N) - K(\sigma) = O(\log_2 N)$, as explained in the proof of Proposition 12.4; hence (12.19) also holds for $K(\sigma \mid N)$ instead of $K(\sigma)$.

Proof. Theorem 6.3 suggests looking at a program $p = p_\sigma$ reaching the minimum in (12.7) as an encoder of σ, so that $\ell(p_\sigma) = K(\sigma \mid N)$ is the length $C(\sigma)$ of the codeword. That this suggestion is correct follows by noting that,[193] being a subset of the prefix set of all p, the shortest σ-producing programs p_σ thus defined form a prefix set in 2^* and hence their lengths satisfy

$$\sum_{\sigma \in A^N} e^{-\ell(p_\sigma)} = \sum_{\sigma \in A^N} e^{-K(\sigma|N)} \leq 1, \tag{12.20}$$

i.e. the Kraft inequality. The lower bound in (12.18) then follows from the lower bound in (6.17).

We prove the upper bound first for $A = 2$. In that case it follows from the estimate

$$K(\sigma \mid N) \leq NS_2\left(\frac{1}{N}\sum_{n=0}^{N-1} \sigma_n\right) + \log_2 N + c. \tag{12.21}$$

Granting this for the moment, we take the expectation value of this inequality under p^N and use Jensen's inequality for the concave function S_2, followed by $\langle \sigma_n \rangle_{p^N} = \langle \sigma_0 \rangle_p = p$. This gives

$$\sum_{\sigma \in A^N} p^N(\sigma) S_2\left(\frac{1}{N}\sum_{n=0}^{N-1} \sigma_n\right) \leq S_2\left(\frac{1}{N}\sum_{n=0}^{N-1} \sum_{\sigma \in A^N} p^N(\sigma)\sigma_n\right) = S_2\left(\frac{1}{N}\sum_{n=0}^{N-1} p\right) = S_2(p). \tag{12.22}$$

To derive (12.21), take $\sigma \in 2^N$ and let $k = \sum_{n=0}^{N-1} \sigma_n$, so that σ contains k copies of 1. Storing k in a program p (that knows N) takes $\log_2 k \leq \log_2 N$ bits. There are $\binom{N}{k}$ binary strings of length N with k copies of 1, which may be listed in some computable way. Identifying σ from these possibilities by its list number requires p to also store this list number, which is at most $\binom{N}{k}$ and hence this takes at most $\log_2 \binom{N}{k}$ bits. Hence the length of p is at most

$$\ell(p) \leq \log_2 N + \log_2 \binom{N}{k} + c'. \tag{12.23}$$

Eq. (12.21) then follows from the upper bound in (5.9), see exercise below. □

[193] Note that we use Lemma 6.2 with $A \rightsquigarrow A^N$, $a \rightsquigarrow \sigma$, and $p \rightsquigarrow p^N$.

Exercise 86 *Explain that for $A = 2$ we have $|T_N(k/N)| = \binom{N}{k}$. On this basis, extend the proof to finite sets A: instead of the number $\frac{1}{N}\sum_{n=0}^{N-1}\sigma_n = \frac{k}{N}$, which defines an element $p \in \mathrm{Prob}(2)$ via $p(1) = k/N$, use the empirical measure* (5.1)*, and instead of $\binom{N}{k}$, use $|T_N(k/N)|$, cf.* (5.2)*.*

Theorem 12.9 describes the asymptotics of the *average* Kolmogorov complexity of *strings*. The pointwise ergodic theorem (in a computable context) eventually leads to a *pointwise* version of (12.19) for *sequences*.[194] We first extend the notion of randomness for sequences $s \in A^{\mathbb{N}}$ to randomness with respect to a prior $p \in \mathrm{Prob}(A)$ and associated Bernoulli measure $p^{\mathbb{N}}$ (as we will see, the original definition then corresponds to the flat prior $p = f$). We refine Definition 12.3 to:[195]

Definition 12.10 *A sequence $s \in A^{\mathbb{N}}$ is* p-random *if there is $c \in \mathbb{N}$ such that for all $N > 1$,*

$$K(s_{|N}) \geq -\log_2 p^N(s_{|N}) - c. \tag{12.24}$$

For $A = 2$ and $p = f$ this recovers (12.10), since $f^N(\sigma) = |A|^{-N}$ for all $\sigma \in A^N$. For general (finite) A and $p = f$, one has $K(s_{|N}) \geq N\log_2(|A|) - c$, which is also correct since $N\log_2(|A|)$ is the length of $\sigma \in A^N$ *measured in bits* (as opposed to: A-valued digits). It is also interesting to notice that even in the original case $A = 2$ and $p = f$, we *could* have stated Definition 12.3 in the form (12.24), which shows that even though Kolmogorov randomness at first sight appears to be an entirely non-probabilistic concept, the flat prior $p = f$ is somehow lurking in the background.

Definition 12.11 *A real $x \in \mathbb{R}$ is* computable *if there is a computable function $f : \mathbb{N} \to \mathbb{Q}$ such that for each $n \in \mathbb{N}$,*

$$x \in \left[\frac{f(n)-1}{n}, \frac{f(n)+1}{n}\right]. \tag{12.25}$$

A probability distribution $p \in \mathrm{Prob}(A)$ is computable *if each $p(a) \in [0,1]$ is a computable real.*

Theorem 12.12 *For all computable probabilities $p \in \mathrm{Prob}(A)$ and all p random sequences $s \in A^{\mathbb{N}}$,*

$$\lim_{N\to\infty} \frac{1}{N} K(s_{|N}) = S_2(p). \tag{12.26}$$

This is similar in spirit to the Shannon–McMillan–Breiman theorem (5.41). The result can be clarified via the Asymptotic Equipartition Property, see (5.34) and subsequent text, which suggests that if s is p-random, then $s_{|N}$ lies in the generic set whose elements all have probability

$$p^N(s_{|N}) \approx e^{-NS(p)} = 2^{-NS_2(p)}, \tag{12.27}$$

so that (12.24) gives $K(s_{|N}) \geq NS_2(p) - c$. This gives the lower bound in (12.18), but now pointwise. The upper bound was proved (for general A, as in the exercise) by replacing the empirical measure $L_N(\sigma) \in \mathrm{Prob}(A)$ by its average p, which also works for $\sigma = s_{|N}$ and p-random s.

[194] See Towsner (2020) and Landsman (2023),and references therein.

[195] Experts will notice that our term "p-random" conflicts with the use of "1-random" etc. elsewhere.

A Convexity

The concept of convexity pervades thermodynamics and associated theories of entropy.

Definition A.1 *Let X be a vector space.*

1. *A subset $C \subset X$ is* convex *if for any $x, y \in C$ we also have $\lambda x + (1-\lambda)y \in C$ for all $\lambda \in (0,1)$ (in other words, the straight line segment between x and y entirely lies in C).*

   Equivalently, for any finite set of probabilities $(p_i)_{i\in I}$, i.e. $p_i \geq 0$ and $\sum_i p_i = 1$, any any set $(x_i)_{i\in I}$ of points in $\mathscr{D}_f$ (i.e. as many points as there are probabilities), we have $\sum_{i\in I} p_i x_i \in C$.

2. *Let $C \subset X$ be convex and $f : C \to (-\infty, \infty] = \mathbb{R} \cup \{\infty\}$ some function taking values in the (semi-) extended reals. The* domain *$\mathscr{D}_f$ of f is $\mathscr{D}_f = \{x \in C \mid f(x) < \infty\}$.*

3. *Such a function f is called* convex *if for all $x, y \in \mathscr{D}_f$ and $\lambda \in (0,1)$ we have*

$$f(\lambda x + (1-\lambda)y)) \leq \lambda f(x) + (1-\lambda) f(y). \tag{A.1}$$

   It is strictly convex *if this holds with strict inequality $<$ instead of $\leq$.*

4. *A function $g : [-\infty, \infty)$ is (strictly)* concave *if $f = -g$ is (strictly) convex,*

Exercise 87 *Show that f is convex iff its* epigraph $\mathrm{epi}(f)$ *is convex, where*

$$\mathrm{epi}(f) := \{(x,t) \in \mathscr{D}_f \times \mathbb{R} \mid f(x) \leq t\}. \tag{A.2}$$

If C is convex and $f : C \to (-\infty, \infty]$ is convex, then by (A.1) it follows that $\mathscr{D}_f$ is convex. Iterating (A.1) shows that f is convex iff for all (p_i) as in point 1 (second part), and all (x_i) in $\mathscr{D}_f$, we have

$$f\left(\sum_i p_i x_i\right) \leq \sum_i p_i f(x_i). \tag{A.3}$$

The is a special case of *Jensen's inequality*, which briefly reads $f(\langle F \rangle_P) \leq \langle f \circ F \rangle_P$, that is,

$$f\left(\int_X dP\, F\right) \leq \int_X dP\, f \circ F. \tag{A.4}$$

This applies to probability spaces (X, P), functions $F \in L^1(X,P)$ where $F(X) \subset I \subset \mathbb{R}$, and measurable convex functions $f : I \to \mathbb{R}$: If f is strictly convex we have equality iff F is constant P-a.e. Proofs are everywhere. Taking $X = \{x_1, \ldots, x_{|I|}\}$, $P(x_i) = p_i$, and $F(x) = x$ reproduces (A.3).

For differentiable functions convexity is easy to establish, here are some results:[196]

Proposition A.2 *1. If $I \subset \mathbb{R}$ is an interval, and then $f : I \to \mathbb{R}$ is C^1, then f is convex iff f' is nondecreasing on I, and strictly convex iff f' is strictly increasing on I.*

2. *In particular, if f is C^2 then f is convex iff $f''(x) \geq 0$ on I.*

3. *If $U \subset \mathbb{R}^d$ is open and convex and $f : U \to \mathbb{R}$ is C^1, then f is convex iff for all $x, y \in U$,*

$$\langle \nabla f(x) - \nabla f(g), x - y \rangle \geq 0. \tag{A.5}$$

[196] See e.g. Borwein & Vanderwerff, §2.2.1. There are deeper results than those stated here if continuous differentiability is weakened to Gâteaux or Fréchet differentiability.

4. *In particular, if f is C^2 then f is convex iff $\nabla^2 f(x) \geq 0$ for all $x \in U$ and $h \in \mathbb{R}^d$.*

Here $\nabla^2 f(x)$ is the *Hessian* at x, which in the C^2 case is simply the matrix with entries $\partial^2 f/\partial x^i \partial x^j$.

Rate functions I in the theory of large deviations are typically convex whereas their negatives, entropies $S = -I$, are concave. The so-called *relative entropy* $p \mapsto D(p\|q)$ is actually a rate function and so it is con*vex* (even jointly as a function of (p,q)); see (5.12), (7.16), and Theorem 7.3. Relative entropies actually take the value ∞. For example, for a fair coin toss with $p(1) = x$ (and hence $p(0) = 1-x$) and $q(1) = 1/2$ (and hence $q(0) = 1/2$) we may see $D(p\|q)$ as a function $x \mapsto I(x)$ on $X = C = \mathbb{R}$, given by, see also (8.11).

$$I(x) = (1-x)\log(1-x) + x\log x + \log 2 \ \ (x \in [0,1]; \tag{A.6}$$

$$I(x) = \infty \ \ (x \notin [0,1]). \tag{A.7}$$

The most natural continuity property for a convex functions is *lower semicontinuity* (lsc):

Definition A.3 *We call $f : X \to (-\infty, \infty]$* lower semicontinuous *(lsc) if for each $t \in \mathbb{R}$ the set*

$$f^{-1}((-\infty, t]) = \{x \in X \mid f(x) \leq t\} \tag{A.8}$$

is closed in X.

Equivalently, the epigraph $\mathrm{epi}(f)$ is closed in $X \times \mathbb{R}$, so that f is convex and lsc iff $\mathrm{epi}(f)$ is convex and closed. There also exists a purely topological characterization (or definition) of lower semicontinuity:

Lemma A.4 1. *A function $f : X \to (-\infty, \infty]$ is lower semicontinuous iff for each $x \in X$ and each $t < f(x)$ there is an open neighbourhood U of x such that $t < f(y)$ for all $y \in U$.*

2. *For metric spaces,*[197] *this is the case iff for any convergent sequence $x_n \to x$ one has*

$$\liminf_{n\to\infty} f(x_n) \geq f(x). \tag{A.9}$$

See Penot, Proposition 2.10. Informally, lower semicontinuity means continuity with the possible exception that if f suddenly jumps, the jump must be upwards, as in $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = 0$ if $x \leq 0$ and $f(x) = 1$ if $x > 0$. Here is an efficient way to check if an lsc function on $\mathbb{R}$ is convex:

Lemma A.5 *For some interval $C \subset \mathbb{R}$, let $f : C \to (-\infty, \infty]$ be lsc or increasing (i.e. monotone non-decreasing). Then f is convex iff*

$$f(\tfrac{1}{2}x + \tfrac{1}{2}y) \leq \tfrac{1}{2}f(x) + \tfrac{1}{2}f(y) \ \ \text{for all}\, x, y \in C. \tag{A.10}$$

Thus (A.1) only needs to be checked for $\lambda = \frac{1}{2}$. See Simon (2011), Proposition 1.4, for the lsc case. The proof iterates (A.10), which gives (A.1)for all dyadic $\lambda \in (0,1)$, upon which lsc gives it for all $\lambda \in (0,1)$. If f is increasing, assume that $C \subset \mathbb{R}^+$ (it is easy to adapt the proof to other cases). Approximate some non-dyadic $0 < \lambda < 1$ from both above and below by dyadic rationals, i.e., $d_n < \lambda < d'_n$, with $d_n \to \lambda$ and $d'_n \to \lambda$. Since $\lambda x + (1-\lambda)y) < d'_n x + (1-d_n)y$, we have

$$f(\lambda x + (1-\lambda)y)) \leq f(d'_n x + (1-d_n)y)) \leq d'_n f(x) + (1-d_n)f(y) \to \lambda f(x) + (1-\lambda)f(y). \tag{A.11}$$

Proposition A.6 *If $X = \mathbb{R}^d$ and $f : C \to (-\infty, \infty]$ is convex, then f is continuous (and even locally Lipschitz) on $\mathring{\mathscr{D}}_f$ (i.e. the interior of $\mathscr{D}_f$). This remains true if X is a possibly infinite-dimensional Banach space, under the additional assumption that f is continuous at* some *point in $\mathscr{D}_f$.*

[197] This is even true for general topological spaces if sequences are replaced by nets. Se Penot, Lemma 2.2.

See BV, Theorem 2.1.12, Proposition 4.1.4; Penot, Propositions 6.2, 6.4; Dorlas, Theorem 7.1.

Proposition A.7 *Let X be a Banach space, $C \subset X$ a convex subset, and $f : C \to (-\infty, \infty]$ convex.*

1. *Any local minimum of f is also a global minimum.*
2. *If in addition f is lsc and C is compact, then f has a local and hence global minimum.*

See BV, Theorem 1.2.2 plus the usual Weierstrass theorem.[198]

If X is a normed space (as we often assume), a (continuous) *affine function* on X is defined as

$$\ell_{(\phi,r)} : X \to \mathbb{R}; \qquad \ell_{(\phi,r)}(x) := \phi(x) + r; \qquad (\phi \in X^*, r \in X). \tag{A.12}$$

For example, if $X = \mathbb{R}^d$ we take $\phi \in \mathbb{R}^d$ with $\phi(x) = \langle \phi, x \rangle$ (the Euclidean inner product). More generally, (A.12) makes sense if X is a topological vector space and another vector space Y consists of all continuous linear functionals φ on X; also in that case one often writes $\langle \phi, x \rangle$ for $\phi(x)$.

Affine functions are clearly convex and continuous functions are evidently lsc.

Proposition A.8 *Let $f : X \to (-\infty, \infty]$.*

1. *If f is the supremum of a nonempty family of continuous affine functions on X, then either f is identically equal to ∞ or $\mathscr{D}_f \neq \emptyset$ and f is convex and lsc.*
2. *Conversely, if $\mathscr{D}_f \neq \emptyset$ and f is convex and lsc, then f is the supremum of a family (f_γ) of continuous affine functions, namely the ones that are majorized by f. That is:*

$$f = \sup\{\ell_{(\phi,r)} \mid \phi \in Y, r \in X, \ell_{(\phi,r)} \leq f\}. \tag{A.13}$$

3. *The supremum of any family (f_γ) of convex lsc functions $f_\gamma : X \to (-\infty, \infty]$ is convex lsc.*

See Penot, Theorem 6.2 and Proposition 2.11; BV, Lemma 1.2.1; Simon, Theorem 5.15. One can of course also try (A.13) for arbitrary functions $f : X \to (-\infty, \infty]$. This gives the *convex envelope*

$$\begin{aligned} f_* :&= \sup\{\ell_{(\phi,r)} \mid \phi \in Y, r \in X, \ell_{(\phi,r)} \leq f\} \\ &= \sup\{g \mid g \leq f, g \,\text{convex and lsc}, \mathscr{D}_g \neq \emptyset\} \leq f, \end{aligned} \tag{A.14}$$

which in general differs from f; but if f is convex and lsc with $\mathscr{D}_f \neq \emptyset$, we clearly have $f_* = f$, and $\mathrm{epi}(f_*)$, which in general equals the closed convex hull $\overline{\mathrm{co}}(\mathrm{epi}(f))$, coincides with $\mathrm{epi}(f)$.[199]

Definition A.9 *Let X be a normed space with dual X^* and let $f : X \to (-\infty, \infty]$ be any function.*[200] *The* Fenchel transform *(also called the* Legendre or Fenchel–Legendre transform*) of f is*

$$f^* : X^* \to (-\infty, \infty]; \qquad f^*(\phi) := \sup_{x \in X}\{\phi(x) - f(x)\}. \tag{A.15}$$

[198] Weierstrass–any continuous real-valued function on a compact space K has a maximum and a minimum–is a combination of (i): any lsc function on K has a minimum; and (ii): any usc function on K has a maximum.

[199] The *convex hull* $\mathrm{co}(S)$ of any subset S of a vector space Y is the smallest convex set in Y that contains S; this set exists and equals the intersection of all convex subsets of Y that contain S. If Y is a topological vector space, then the *closed convex hull* $\overline{\mathrm{co}}(S)$ is the closure of $\mathrm{co}(S)$.

[200] One may replace X^* by any topological vector space is separating duality with X.

One may again think of $X = \mathbb{R}^d \cong X^*$, with pairing via the inner product. Since f^* is a supremum of continuous affine functions, by Proposition A.8 it is either identically equal to ∞, or convex and lsc with $\mathscr{D}_{f^*} \neq \emptyset$. The second case therefore applies iff f majorizes *some* affine function, because Definition A.9 implies that for $\phi \in X^*$ and hence $(\phi, r) \in X^* \times \mathbb{R}$ we have

$$(\phi, r) \in \mathrm{epi}(f^*) \qquad \Leftrightarrow \qquad \ell_{(\phi,r)} \leq f, \tag{A.16}$$

where the $\leq$ is pointwise. Hence there is a pair $(\phi, r) \in X^* \times \mathbb{R}$ such that $\ell_{(\phi,r)} \leq f$ iff there exists some $\phi \in X^*$ for which $f^*(\phi) < \infty$. This leads to a nice geometric interpretation of the Fenchel transform, most easily visualized for $X = X^* = \mathbb{R}$, so that $\phi(x) = \phi x$. If $f^*(\phi) < \infty$, the definition (A.15) implies that $\phi(x) - f(x) \leq f^*(\phi)$ for all $x \in X$, and hence

$$\phi(x) - f^*(\phi) \leq f(x), \tag{A.17}$$

for all x. By definition of a supremum, $f^*(\phi)$ is therefore the smallest number $r \in \mathbb{R}$ such that $\phi(x) - r \leq f(x)$ for all x. In case that $\phi(x) \leq f(x)$ for all x, $f^*(\phi)$ is therefore the smallest number such that the graph of $x \mapsto \phi(x)$ in $X \times \mathbb{R}$ moved upward by $-f^*(\phi)$ just touches the graph (or epigraph) of f. If $\phi(x) > f(x)$ for some x, then $f^*(\phi)$ is the smallest number such that the graph of $x \mapsto \phi(x)$ moved downward by $f^*(\phi)$ just touches the graph of f. See also Dorlas, Figure 7.2.

Exercise 88 *Prove (i.e. compute) the following examples in $X = \mathbb{R}$ (we write y instead of φ):*

1. *$f(x) = e^x \rightsquigarrow f^*(y) = y\log y - y$ for $y > 0$, $f^*(y) = 0$ for $y = 0$, and $f^*(y) = \infty$ for $y < 0$.*
2. *$f(x) = -\log x$ for $x \in (0,\infty)$ and $f(x) = \infty$ otherwise $\rightsquigarrow f^*(y) = -1 - \log(-y)$ for $y \in (-\infty, 0)$ and $f^*(y) = \infty$ otherwise.*
3. *$f(x) = p^{-1}|x|^p$ with $p > 1$ on $\mathbb{R} \rightsquigarrow f^*(y) = q^{-1}|y|^q$ on $\mathbb{R}$ for the conjugate q ($p^{-1} + q^{-1} = 1$).*
4. *$f(x) = -p^{-1}x^p$ with $p < 1$ on $[0,\infty)$ and $f(x) = \infty$ otherwise $\rightsquigarrow f^*(y) = -q^{-1}(-y)^q$ for $y \in (-\infty, 0)$ and $f^*(y) = \infty$ otherwise (again for the conjugate value of q).*
5. *What is $f^*(y)$ for $f(x) = |x|$?*

Note that seemingly strange signs are chosen so as to make f (and hence f^*) convex and lsc.

Theorem A.10 *Let $f : X \to (-\infty, \infty]$ with $\mathscr{D}_f \neq \emptyset$ and also suppose that $\mathscr{D}_{f^*} \neq \emptyset$.*[201] *Then*

$$f^{**}_{|X} = f_*. \tag{A.18}$$

Consequently, $f^{**}(x) \leq f(x)$ for all $x \in X$, and if f is convex and lsc (so that $f_* = f$), then

$$f^{**}_{|X} = f. \tag{A.19}$$

Here the double transform f^{**} is by construction defined on X^{**}, which contains X as a subspace under the injection $x \hookrightarrow \hat{x}$, where $\hat{x} \in X^{**}$ is defined by $\hat{x}(\phi) = \phi(x)$. Thus we have

$$f^{**}(\hat{x}) = f^{**}_{|X}(x) = \sup_{\phi \in X^*} \{\phi(x) - f^*(\phi)\}, \tag{A.20}$$

which we simply write as $f^{**}(x)$. In the more general case discussed above, just replace X^* by Y.

[201] So this is the case iff f majorizes *some* affine function.

Proof. Since $\phi(x)+r \le f(x)$ for all x is the same as $-r \ge f^*(\phi)$, eq. (A.14) comes down to

$$f_*(x) = \sup\{\phi(x)+r \mid \phi \in Y, r \in X, -r \ge f^*(\phi)\} = \sup_{\phi\in X^*}\{\phi(x)-r \mid r \in X, r \ge f^*(\phi)\}$$
$$= \sup_{\phi\in X^*}\{\phi(x)-r \mid r = f^*(\phi)\} = \sup_{\phi\in X^*}\{\phi(x)-f^*(\phi)\} = f^{**}(x), \tag{A.21}$$

since any $r > f^*(\phi)$ can only lower $\phi(x)-r$ compared to $\phi(x)-f^*(\phi)$ and will therefore not contribute to the supremum, whereas $r < f^*(\phi)$ is excluded by the constraint $r \ge f^*(\phi)$. □

Proposition A.11 (Fenchel–Young inequality) *Let X be a normed space and let $f : X \to (-\infty,\infty]$ be convex. Then for any $x \in \mathscr{D}_f$ and $\phi \in X^*$ we have*

$$f(x)+f^*(\varphi) \ge \phi(x), \tag{A.22}$$

with equality iff $\phi \in \partial f(x)$, where the subdifferential $\partial f(x) \subset X^*$ *is defined by*

$$\partial f(x) := \{\phi \in X^* \mid \forall_{x'\in X}(\phi(x')-\varphi(x) \le f(x')-f(x))\} \ \ (f(x)<\infty); \tag{A.23}$$
$$\partial f(x) := \emptyset \ \ (f(x)=\infty). \tag{A.24}$$

Proof. The inequality is the same as (A.17). Hence equality holds iff $f(x)+f^*(\varphi) \le \phi(x)$, which by definition of f^* is the case iff for all $x' \in X$ we have $f(x)+\phi(x')-f(x') \le \phi(x)$. □

The subdifferential (defined also if f is not convex) is an important concept; it plays the role of the derivative when the latter may not exist. For example, let $f(x) = |x|$ on $\mathbb{R}$. Then

$$\partial f(x) = \{-1\} \ \ (x<0); \qquad \partial f(0) = [-1,1]; \qquad \partial f(x) = \{1\} \ \ (x>0). \tag{A.25}$$

This example generalizes as follows to normed spaces: if $f(x) = \|x\|$, then

$$\partial f(0) = \{\phi \in X^* \mid \|\phi\| \le 1\}; \quad \partial f(x) = \{\phi \in X^* \mid \|\phi\| = 1, \phi(x) = \|x\|\} \ \ (x \ne 0). \tag{A.26}$$

The subdifferential may also be (nontrivially) empty: a textbook example is the convex lsc function

$$f : \mathbb{R} \to (-\infty,\infty]; \quad f(x) = -\sqrt{1-x^2} \ \ (-1 \le x \le 1); \quad f(x) = \infty \ \ (x \notin [-1,1]). \tag{A.27}$$

Then clearly $\partial f(x) = f'(x) = \{x/\sqrt{1-x^2}\}$ for $x \in (-1,1)$ but $\partial f(x) = \emptyset$ not only for $x \notin [-1,1]$), which is true by definition, but even within its domain at $x = \pm 1$.

Proposition A.12 *A function $f : X \to (-\infty,\infty]$ on a normed space X attains a local minimum (and hence a global minimum if f is convex) iff $0 \in \partial f(x)$.*

This is clear from the definitions. If f is $C^1(\mathbb{R})$ then simply $\partial f(x) = f'(x)$. For an analogous result in higher dimension one needs the concept of a Gâteaux derivative:[202]

Definition A.13 *A function $f : X \to (-\infty,\infty]$ on a normed space X is* Gâteaux differentiable *at $x \in \mathscr{D}_f$ if the following limit exists for all $x' \in X$ and defines (the* Gâteaux derivative*) $f'(x) \in X^*$:*

$$f'(x) : x' \mapsto \lim_{t\to 0}\frac{f(x+tx')-f(x)}{t}. \tag{A.28}$$

[202] See e.g. Penot, §6.2.2., or Borwein & Vanderwerff, §4.2, etc.

Proposition A.14 *If $f : X \to (-\infty, \infty]$ is convex, as well as Gâteaux differentiable at $x \in \mathscr{D}_f$, then*

$$f'(x) \in \partial f(x). \tag{A.29}$$

If under the same assumptions x is in the interior *of $\mathscr{D}_f$, then $\partial f(x)$ contains a single element:*

$$\partial f(x) = \{f'(x)\}. \tag{A.30}$$

Conversely, if f is convex, as well as continuous *at $x \in \mathscr{D}_f$, then* (A.30) *holds iff f is Gâteaux differentiable at x.*

See Penot, Theorem 6.3 and Corollary 6.5, and Borwein & Vanderwerff, Fact 4.2.4 and Corollary 4.2.5. Here are some more properties of convex functions on $\mathbb{R}$.

Proposition A.15 *If $I \subset \mathbb{R}$ is an open interval abd $f : I \to \mathbb{R}$ is convex then f is differentiable almost everywhere on I (more precisely: the set of I where $f'(x)$ does not exist is at most countable).*

See Penot, Proposition 6.7.

Proposition A.16 (Three-slope inequality) *1. If $I \subset \mathbb{R}$ is an open interval (possibly $I = \mathbb{R}$) and $f : I \to (-\infty, \infty]$ is convex, then for all $x, y, z \in \mathscr{D}_f$ for which $x < y < z$ we have*

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x} \leq \frac{f(z) - f(y)}{z - y}. \tag{A.31}$$

2. Moreover, both the right (+) and left (-) derivatives

$$f'_{\pm}(x) = \pm \lim_{h \downarrow 0} \frac{f(x \pm h) - f(x)}{h} \tag{A.32}$$

exist and are finite at each $x \in I$ and are increasing, with $f'_-(x) \leq f'_+(x)$.

3. For any $\lambda \in [f'_-(x), f'_+(x)]$ and all $y \in I$ we have

$$f(y) \geq f(x) + \lambda(y - x). \tag{A.33}$$

See BV, Fact 2.1.1, Theorem 2.1.2, and Corollary 2.1.3. No. 3 obviously follows from 1 and 2.

Corollary A.17 *If $f : \mathbb{R} \to (-\infty, \infty]$ is convex, lsc, and increasing,*[203] *and we know that*

$$p(t) = \sup_{x \in \mathbb{R}} \{xt - f(x)\} \tag{A.34}$$

only for all $t \geq 0$, for some given (necessarily lsc convex) function p, then

$$f(x) = \sup_{t \geq 0} \{xt - p(t)\}. \tag{A.35}$$

Proof. We have

$$F(x) := \sup_{t \geq 0} \{xt - p(t)\} = \sup_{t \geq 0} \inf_{y \in \mathbb{R}} \{(x - y)t + f(y)\}. \tag{A.36}$$

Then $F(x) \leq f(x)$ by taking $y = x$. Conversely, since f is increasing we have $\lambda \geq 0$ in (A.33), which gives $F(x) \geq f(x)$. □

The following result (more specifically its corollary) forms the ultimate explanation of the equality between the two different expressions (8.3) and (8.5) for the rate function in Cramér's theorem, and more generally is a very abstract form of maximum entropy principles.[204]

[203] By "increasing" we always mean *monotone non-decreasing*, and likewise for "decreasing".

[204] Good sources are Borwein & Vanderwerff, §4.4.2, Borwein & Zhu, §4.4, and Penot, §6.3. Borwein c.s. assume that X and Y are Banach spaces, but completeness is not used in the proof and indeed Penot, Theorem 6.18, just assumes normed spaces like we do. Dohmatob (undated) also provides a nice (but unfortunately unfinished) introduction.

Theorem A.18 (Fenchel–Rockafellar duality) *Let X and Y be normed spaces, and let*

$$f : X \to (-\infty, \infty]; \qquad g : Y \to (-\infty, \infty]; \qquad T : X \to Y \tag{A.37}$$

be two convex lsc functions, and a bounded linear operator, respectively. Define

$$p := \inf_{x \in X} \{f(x) + g(Tx)\}; \tag{A.38}$$

$$d := \sup_{\varphi \in Y^*} \{-g^*(-\varphi) - f^*(T^*\varphi)\} \tag{A.39}$$

to be the so-called primal *and* dual *optimization problems. Then*

$$p \geq d, \tag{A.40}$$

and if either $T(\mathscr{D}_f)$ contains a point of continuity of g, or $\operatorname{int}(\mathscr{D}_g - T(\mathscr{D}_f))$ contains 0, then

$$p = d. \tag{A.41}$$

Before proving this, let us derive a crucial consequence for the theory of entropy:

Corollary A.19 *For f and T as in Theorem A.18, and any $y \in Y$, we have*

$$\inf_{x \in X} \{f(x) \mid Tx = y\} \geq \sup_{\varphi \in Y^*} \{\varphi(y) - f^*(T^*\varphi)\}, \tag{A.42}$$

with equality under the additional assumption $y \in \operatorname{int}(T(\mathscr{D}_f))$.

Proof of Corollary A.19. Take $g = \iota_{\{y\}}$ in Theorem A.18, where, for any subset $S \subset Y$ we define

$$\iota_S : Y \to (-\infty, \infty]; \qquad \iota_S(y) = 0 \ \ (y \in S); \qquad \iota_S(y) = \infty \ \ (y \notin S). \tag{A.43}$$

Its Fenchel transform is easy to compute, viz. $1_S^*(\varphi) = \sup_{y \in S}\{\varphi(y)\}$, so for $S = \{y\}$ we obtain

$$1^*_{\{y\}} : Y^* \to (-\infty, \infty]; \qquad 1^*_{\{y\}}(\varphi) = \varphi(y). \tag{A.44}$$

The term $g(Tx)$ in (A.38) then obviously becomes the constraint $Tx = y$ on the left-hand side of (A.42), whereas the term $-g^*(-\varphi)$ in (A.39) comes down to $\varphi(y)$ on the right-hand side of (A.42).

Finally, since obviously $\mathscr{D}_{\iota_{\{y\}}} = \{y\}$, the condition $0 \in \operatorname{int}(\mathscr{D}_g - T(\mathscr{D}_f))$ in Theorem A.18 becomes the assumption $y \in \operatorname{int}(T(\mathscr{D}_f))$ in the second claim of Corollary A.19. □

Proof of Theorem A.18. For (A.40), we note that $g^{**}(x) \leq g(x)$ (see Theorem A.10) and hence:

$$\begin{aligned}
p \geq \inf_{x \in X}\{f(x) + g^{**}(Tx)\} &= \inf_{x \in X} \sup_{\phi \in Y^*} \{f(x) + \phi(Tx) - g^*(\phi)\} \\
&\geq \sup_{\phi \in Y^*} \inf_{x \in X} \{f(x) + \phi(Tx) - g^*(\phi)\} \\
&= \sup_{\phi \in Y^*} \inf_{x \in X} \{f(x) - T^*\phi(x) - g^*(-\phi)\} \\
&= \sup_{\phi \in Y^*} \inf_{x \in X} \{-(T^*\phi(x) - f(x) + g^*(-\phi))\} \\
&= \sup_{\phi \in Y^*} \{-g^*(-\phi) - \sup_{x \in X}\{(T^*\phi(x) - f(x)\}\} \\
&= \sup_{\varphi \in Y^*} \{-g^*(-\varphi) - f^*(T^*\varphi)\} = d.
\end{aligned} \tag{A.45}$$

For the equality (A.41), we first note that if $p = -\infty$, then $p \geq d$ forces $p = d$ and we are ready. We may therefore assume that $p \neq -\infty$, in which case we introduce an auxiliary function

$$h : Y \to [-\infty, \infty]; \qquad h(y) := \inf_{x \in X} \{f(x) + g(Tx + y)\}. \tag{A.46}$$

The additional assumptions stated above (A.41) are used to show that $\partial h(0) \neq \emptyset$; we omit this highly technical part of the proof.[205] Granting this, there exists $-\phi \in \partial h(0)$ and hence, by definition of the subdifferential, $h(0) \leq h(y) + \phi(y)$ for all $y \in Y$. Hence for all $y \in Y$ we have:

$$\begin{aligned}
p &\leq \inf_{x \in X} \{f(x) + g(Tx + y) + \phi(y)\} \\
&\Rightarrow \forall_{x \in X, y \in Y}\ p \leq f(x) + g(Tx + y) + \phi(y) \\
&\overset{y \rightsquigarrow y - Tx}{\Longrightarrow} \forall_{x \in X, y \in Y}\ p \leq f(x) - T^*\phi(x) - (-\phi(y) - g(y)) \\
&\overset{\inf_y}{\Longrightarrow} \forall_{x \in X}\, p \leq -(T^*\phi(x) - f(x)) - g^*(-\varphi) \\
&\overset{\inf_x}{\Longrightarrow} p \leq -f^*(T^*\phi) - g^*(-\varphi) \\
&\overset{\sup_\phi}{\Longrightarrow} p \leq d,
\end{aligned} \tag{A.47}$$

where we repeatedly used "$-\inf = \sup -$" and (A.15). □

Definition A.20 *The* **(extreme) boundary** $\partial_e K$ *of a convex set* K *consists of all points* $v \in K$ *satisfying the following condition:*

if $v = tw + (1-t)x$ *for certain* $w, x \in K$ *and* $t \in (0,1)$*, then* $v = w = x$.

Elements $v \in \partial_e K$ *of the boundary are called* **extreme points** *of* K.

Here the simplest case is the *simplex* Δ_n, defined for all $n \geq 1$ by

$$\Delta_n = \{x \in \mathbb{R}^{n+1} \mid x_i \geq 0, \sum_i x_i = 1\}. \tag{A.48}$$

This is the set $\mathrm{Prob}(X_{n+1})$ of all probability distributions on a set X_{n+1} of cardinality $n+1$, and

$$\partial_e \Delta_n = \{\vec{e}_1, \ldots . \vec{e}_{n+1}\}, \tag{A.49}$$

where $(\vec{e}_1, \ldots . \vec{e}_{n+1})$ is the standard basis of $\mathbb{R}^{n+1}$ (i.e., $\vec{e}_1 = (1, 0, \ldots, 0)$, etc.). Beware of surprises: for $K = \Delta_n$ the *extreme* boundary $\partial_e K$ therefore consists of the *vertices* of K, whereas its *faces* form the *geometric* boundary. More generally:

Proposition A.21 *Any finite set* X *is isomorphic to the boundary* $\partial_e \mathrm{Prob}(X)$ *through* $x \mapsto P_x$.

Exercise 89 *Prove this.*

The simplest example is $X = \{0, 1\}$, so that $\mathrm{Prob}(X) \cong [0, 1]$ by mapping the distribution $p \in \mathrm{Prob}(X)$ to $p(1)$. Since one may directly verify that $\partial_e [0, 1] = \{0, 1\}$, under the above isomorphism one therefore has $\partial_e \mathrm{Prob}(X) \cong \{0, 1\}$. Analogously, $\partial_e (0, 1) = \emptyset$, so that the boundary of a convex set may apparently be empty. Hence we see that one remarkable ingredient of Proposition A.21 lies in the claim that the convex set $\mathrm{Prob}(X)$ actually *has* a (nonempty) boundary! This is the case

[205] See Lemma 4.3.1 in Borwein & Zhu. It would be helpful if this step could be simplified.

because it is compact, see below. And if $\partial_e K$ is nonempty, the problem arises whether some point $v \in K$ of a compact convex set K may be written as a convex sum (or, more generally, an integral) of extreme points of K, and if so, to what extent this ***extremal decomposition***

$$v = \sum_{i \in I} t_i v_i,\ t_i \geq 0, \sum_i t_i = 1,\ v_i \in \partial_e K, \tag{A.50}$$

which for simplicity has been assumed to be a finite sum here, is unique. Without proof, we state a general result of finite-dimensional convexity theory, called ***Caratheodory's Theorem***:

Theorem A.22 *If K is a nonempty compact convex subset of $\mathbb{R}^n$, then $\partial_e K \neq \emptyset$, and each point of K is a convex sum of at most $n+1$ points in $\partial_e K$.*

If $K = \Delta_n$, then this sum generically has $n+1$ points and is unique. Probabilistically:

Proposition A.23 *If X is finite, then any probability measure $P \in \mathrm{Prob}(X)$ may be written in a unique way as a* finite *mixture of extreme probability measures, viz.*

$$P = \sum_{x \in X} t_x P_x. \tag{A.51}$$

Proof Take $t_x = P(\{x\})$. To see that this decomposition is unique, use Proposition A.21, i.e. $\partial_e \mathrm{Prob}(X) \cong X$, in (A.50) to force $I = X$ and apply both sides of (A.51) to δ_x. □

Here are some results for the general case, assuming that X is compact.

Theorem A.24 *Let X be a compact Hausdorff space. Then we have a homeomorhism*

$$X \cong \partial_e \mathrm{Prob}(X); \qquad\qquad x \mapsto \delta_x. \tag{A.52}$$

Here δ_x is the (probability) measure defined by $\delta_x(B) = 1$ iff $x \in B$ and $\delta(_x(B) = 0$ iff $x \notin B$. If measures are seen as functionals on $C(X)$ via integration, that is,

$$\varphi(f) = \int_X d\mu\, f,\ f \in C(X), \tag{A.53}$$

the measure δ_x corresponds to the functional $f \mapsto f(x)$.

The ***convex hull*** $\mathrm{co}(X)$ of a subset X of a vector space is defined as the set of all convex sums $tx + (1-t)y$, where $t \in (0,1)$ and $x, y \in X$; this is the smallest convex set containing X.

Theorem A.25 (Krein-Milman) *Let V be a real normed vector space with dual V^*, and let K be a convex subset of V^* that is compact in the w^*-topology. Then $\partial_e K \neq \emptyset$, and each point of K lies in the w^*-closure of the convex hull of $\partial_e K$. In other words,*

$$K = (\mathrm{co}(\partial_e K))^-. \tag{A.54}$$

References

[1] Aczél, J., Daróczy, Z. (1975). *On Measures of Information and Their Characterizations* (Academic Press).

[2] Austin, T. (2017). *Math 254A: Entropy and Ergodic Theory*. https://www.math.ucla.edu/~tim/entropycourse.html.

[3] Beck, C., & Schlögl, F. (1993). *Thermodynamics of Chaotic Systems: An Introduction* (Cambridge University Press).

[4] Bodineau, T., Gallagher, I., Saint-Raymond, L., Simonella, S. (2020). Statistical dynamics of a hard sphere gas: fluctuating Boltzmann equation and large deviations. https://arxiv.org/pdf/2008.10403.pdf.

[5] Boltzmann, L. (1872). Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. *Wiener Berichte* 66, 275–370. Reprinted in Boltzmann, L., *Wissenschaftliche Abhandlungen, Vol. I*, paper 23, ed. F. Hasenöhrl (Chelsea, 1969). English translation: S. Brush, *The Kinetic Theory of Gases: An Anthology of Classic Papers with Historical Commentary*, pp. 262–349 (Imperial College Press, 2003).

[6] Boltzmann, L. (1877). Über die Beziehung dem zweiten Haubtsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. *Wiener Berichte* 76, 373–435. Reprinted in Boltzmann, L., *Wissenschaftliche Abhandlungen, Vol. II*, paper 39, ed. F. Hasenöhrl (Chelsea, 1969). English translation: *Entropy* 17, 1971–2009 (2015). https://www.mdpi.com/1099-4300/17/4/1971.

[7] Boltzmann, L. (1896). *Vorlesungen über Gastheorie. I. Theil* (Verlag von Johann Ambrosius Barth). https://www.deutschestextarchiv.de/book/show/boltzmann_gastheorie01_1896.

[8] Boltzmann, L. (1898). *Vorlesungen über Gastheorie. I. Theil* (Verlag von Johann Ambrosius Barth). https://www.deutschestextarchiv.de/book/view/boltzmann_gastheorie02_1898?p=7.

[9] Borwein, J.M., Vanderwerff, J.D. (2010). *Convex Functions: Constructions, Characterizations and Counterexamples* (Cambridge University Press).

[10] Borwein, J.M., Zhu, Q.J. (2005). *Techniques of Variational Analysis* (Springer).

[11] Brémaud, P. (2020). *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues. Second Edition* (Springer).

[12] Bricmont, L. (2022). *Making Sense of Statistical Mechanics* (Springer).

[13] Brush, S.G. (1974). The development of the kinetic theory of gases: VIII. Randomness and irreversibility. *Archive for History of Exact Sciences* 12, 1–88. https://www.jstor.org/stable/41133384.

[14] Brush, S.G. (1976). *The Kind of Motion We Call Heat* (North-Holland).

[15] Brush, S.G. (2003). *The Kinetic Theory of Gases: An Anthology of Classic Papers with Historical Commentary* (Imperial College Press).

[16] Calude, C.S. (2002). *Information and Randomness: An Algorithmic Perspective*, 2nd Edition (Springer).

[17] Calude, C. S., Staiger, L. (2018). Liouville, computable, Borel normal and Martin-Löf random numbers. *Theory of Computing Systems* 62, 1573–1585.

[18] Carnot, S. (1824). *Réflexions sur la Puissance Motrice du feu* (Bachelier). http://www.numdam.org/article/ASENS_1872_2_1__393_0.pdf. Translation: *Reflections on the Motive Power of Fire, and Other Papers on the Second Law of Thermodynamics by E. Clapeyron and R. Clausius*, ed. Mendoza, E., pp. 1–59 (Dover, 1960). *Chaos and Coarse Graining in Statistical Mechanics* (Cambridge University Press).

[19] Carcignani, C. (1998). *Ludwig Boltzmann: The Man Who Trusted Atoms* (Oxford University Press).

[20] Cerf, R., Petit, P. (2011). A short proof of Cramér's theorem in $\mathbb{R}$. *The American Mathematical Monthly* 118, 925–931. https://doi.org/10.4169/amer.math.monthly.118.10.925,

[21] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics* 23, 493–507. https://www.jstor.org/stable/2236576.

[22] Clausius, R. (1865). Über verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie: vorgetragen in der naturforsch. Gesellschaft den 24. April 1865. *(Poggendorff's) Annalen der Physik und Chemie* Band CXXV, 353–400. https://doi.org/10.1002/andp.18652010702. Reprinted in Clausius, R. (1867). *Abhandlungen über die mechanische Wärmetheorie. Band 2* (Friedrich Vieweg und Sohn). https://www.digitale-sammlungen.de/en/view/bsb10133538?page=1.

[23] Collet, P., Eckmann,, J.-P. (2006). *Concepts and Results in Chaotic Dynamics: A Short Course* (Springer).

[24] Coppersmith, J. (2015). *Energy: The Subtle Concept. Second Edition* (Oxford University Press).

[25] Cover, T.M., Thomas, J.A. (2006). *Elements of Information Theory. Second Edition* (Wiley).

[26] Cramér, H. (1938). Sur un nouveau théoreme-limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles* 736, 5–23. Reprinted in *Harald Cramér, Collected Works II*, ed. Martin-Löf, A., pp. 895–913 (Springer, 1994). https://doi.org/10.1007/978-3-642-40607-2_8.

[27] Dajani, K., Kalle, C. (2021). *A First Course in Ergodic Theory* (CRC Press).

[28] Darrigol. O. (2018). *Atoms, Mechanics, and Probability: Ludwig Boltzmann's Statistico-Mechanical Writings – An Exegesis* (Oxford University Press).

[29] Dasgupta, A. (2011). Mathematical Foundations of Randomness, *Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics*, eds. Bandyopadhyay, P.S., Forster, M.R., pp. 641–710 (North-Holland/Elsevier).

[30] Dembo, A., Zeitouni, A. (1998). *Large Deviations: Techniques and Applications. Second Edition* (Springer).

[31] den Hollander, F. (2000). *Large Deviations* (AMS).

[32] Denker, M., Grillenberger, C., Sigmund, K. (1976). *Ergodic Theory on Compact Spaces* (Springer).

[33] Dohmatob, E. (undated). Fenchel–Rockafellar duality theorem, one ring to rule'em all! - Part 1 https://dohmatob.github.io/research/2019/10/31/duality.html

[34] Dorlas, T.C. (2021). *Statistical Mechanics: Fundamentals and Model Solutions, Second Edition* (CRC).

[35] Dudley, R.M. (1989). *Real Analysis and Probability* (Wadsworth & Brooks/Cole).

[36] Downey, R., Hirschfeldt, D.R. (2010). *Algorithmic Randomness and Complexity* (Springer).

[37] Eagle, A. (2019). Chance versus Randomness. *The Stanford Encyclopedia of Philosophy (Spring 2019 Edition)*, Zalta, E.N. (ed.). https://plato.stanford.edu/archives/spr2019/entries/chance-randomness/.

[38] Ehrenfest, P., Ehrenfest (Afanassjewa), T. (1907a). Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. *Physikalische Zeitschrift* 8: 311–314.

[39] Ehrenfest, P., Ehrenfest (Afanassjewa), T. (1907b). *Begriffliche Grundlagen der statistischen Auffassung in der Mechanik* (Vieweg und Teubner Verlag). English translation: *The Conceptual Foundations of the Statistical Approach in Mechanics* (Dover,1959). https://archive.org/details/conceptualfounda0000ehre.

[40] Einstein, A. (1909). Zum gegenwärtigen Stand des Strahlungsproblem. *Physikalische Zeitschrift* 10, 185–193. Reprinted in *The Collected Papers of Albert Einstein, Vol. 2*, eds. Stachel, J., et al., Doc. 56, pp. 542–550 (Princeton University Press, 1990). https://einsteinpapers.press.princeton.edu/vol2-doc/577. *English Translation Supplement*, pp. 357–375. https://einsteinpapers.press.princeton.edu/vol2-trans/371.

[41] Einstein, A. (1910). Theorie der Opaleszenz von homogenen Flüssigkeiten und Flüssigkeitsgemisch in der Nähe des kritischen Zustandes. *Annalen der Physik 33*, 1275–1298. Reprinted in *The Collected Papers of Albert Einstein, Vol. 3*, eds. M.J. Klein et al., Doc. 9, pp. 286–310. https://einsteinpapers.press.princeton.edu/vol3-doc/325. *English Translation Supplement*, pp. 231–249.

[42] Elkana, Y. (1974). *The Discovery of the Conservation of Energy* (Harvard University Press).

[43] Ellis, R.S. (1985). *Entropy, Large Deviations, and Statistical Mechanics* (Springer).

[44] Ellis, R.S. (1995). An overview of the theory of large deviations and applications to statistical mechanics. *Scandinavian Actuarial Journal* 1, 97–142.

[45] Ellis, R.S. (1999). The theory of large deviations: From Boltzmann's 1877 calculation to equilibrium macrostates in 2D turbulence. *Physica D* 133, 106–136.

[46] Emch, G.G., Liu, C. (2013). *The Logic of Thermostatistical Physics* (Springer).

[47] Faddeev, D.K. (1956). On the concept of entropy of a finite probabilistic scheme. *Uspekhi Matematicheskikh Nauk* 11, 227–231. https://arrowtheory.com/pub/notes/025-faddeev-entropy.html.

[48] Franklin, J.Y. , Porter, C.P., eds. (2020). *Algorithmic Randomness: Progress and Prospects* (Cambridge University Press).

[49] Friedli, S., Velenik, Y. (2018). *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction* (Cambridge University Press). https://www.unige.ch/math/folks/velenik/smbook/Statistical_Mechanics_of_Lattice_Systems.pdf.

[50] Gaudenzi, R. (2019). Entropy? *Exercices de Style*. *Entropy* 21: 742. https://doi.org/10.3390/e21080742.

[51] Gelfert, K., Kwietniak, D. (2018). The (Poulsen) simplex of invariant measures. *Ergodic Theory and Dynamical Systems* 38, 1745–1767. https://doi.org/doi:10.1017/etds.2016.97.

[52] Georgii, H.-O. (1993). Large deviations and maximum entropy principle for interacting random fields on $\mathbb{Z}^d$. *The Annals of Probability*, 1845–1875. https://www.jstor.org/stable/2244702.

[53] Georgii, H.-O. (2003). Probabilistic aspects of entropy. *Entropy*, eds. Greven, A., Keller, G., Warnecke, G., pp. 37–54 (Princeton University Press).

[54] Georgii, H.-O. (2011). *Gibbs Measures and Phase Transitions. Second Edition* (De Gruyter).

[55] Gibbs, J.W. (1902). *Elementary Principles of Statistical Mechanics* (Scribner). https://www.gutenberg.org/ebooks/50992.

[56] Gleick, J. (2011). *The Information: A History, a Theory, a Flood* (Pantheon).

[57] Grünwald, P.D., Vitányi, P.M.B. (2003). Kolmogorov complexity and Information theory. With an interpretation in terms of questions and answers. *Journal of Logic, Language and Information* 12, 497–529. https://arxiv.org/abs/cs/0410002.

[58] Grünwald, P.D., Vitányi, P.M.B. (2008). Algorithmic information theory. *Handbook of the Philosophy of Information*, eds. Adriaans, P., van Benthem, J., pp. 281–320 (Elsevier). https://arxiv.org/abs/0809.2754.

[59] Guizzo, E.M. (2003). *The Essential Message: Claude Shannon and the Making of Information Theory* (M.Sc Thesis, MIT). https://dspace.mit.edu/bitstream/handle/1721.1/39429/54526133-MIT.pdf?sequence=2.

[60] Halmos, P. (1958). Von Neumann on measure and ergodic theory. *Bulletin of the American Mathematical Society* 64, 86–94. https://doi.org/10.1090/S0002-9904-1958-10203-7.

[61] Harman, P.M. (1982). *Energy, Force, and Matter: The Conceptual Development of Nineteenth-Century Physics* (Cambridge University Press).

[62] Hoyer, U. (1980). Von Boltzmann zu Planck. *Archive for History of Exact Sciences* 47–86. https://www.jstor.org/stable/41133587.

[63] Jauch, J.M., Baron, J.G. (1972). Entropy, information and Szilard's paradox. *Helvetica Physica Acta* 45, 220–232. https://access.archive-ouverte.unige.ch/access/metadata/2461b11c-d1cc-459c-ae84-2f0c16e34b93/download.

[64] Jona-Lasinio, G. (2015). Large deviations and the Boltzmann entropy formula. *Brazilian Journal of Probability and Statistics* 29, 494–501. https://www.jstor.org/stable/26358989.

[65] Kamae, T., Keane, M. (1997). A simple proof of the ratio ergodic theorem. *Osaka Journal of Mathematics* 34, 653–657.

[66] Keane, M. (2005). The essence of the law of large numbers. *Algorithms, Fractals, and Dynamics*, Takahashi, Y. (ed.), pp. 125–129 (Springer). https://doi.org/10.1007/978-1-4613-0321-3_11. https://www.isibang.ac.in/~athreya/Teaching/c12/lln3.pdf.

[67] Keller, G. (1998). *Equilibrium States in Ergodic Theory* Cambridge University Press).

[68] Khinchin, A.I. (1957). *Mathematical Foundations of Information Theory* (Dover). Translated by R. A. Silverman and M. D. Friedman from Uspekhi Matematicheskikh Nauk, 7, 3–20 (1953) and 9, 17–75 (1956).

[69] Kifer, Y. (1990). Large deviations in dynamical systems and stochastic processes. *Transactions of the American Mathematical Society* 321, 505–524. https://doi.org/10.1090/S0002-9947-1990-1025756-7.

[70] Khoshnevisan, D. (2006). Normal numbers are normal. https://www.claymath.org/library/annual_report/ar2006/06report_normalnumbers.pdf.

[71] Klein, M.J. et al. (1994). *The Collected Papers of Albert Einstein, Volume 3: The Swiss Years: Writings, 1909-1911* (Princeton University Press, 1990). https://einsteinpapers.press.princeton.edu/vol3-doc/.

[72] Klenke, A. (2020). *Probability Theory: A Comprehensive Course. Third Edition* (Springer).

[73] Klir, G.J. (2006). *Uncertainty and Information: Foundations of Generalized Information Theory* (Wiley–Interscience).

[74] Kolmogorov, A.N. (1958). New metric invariant of transitive dynamical systems and endomorphisms of Lebesgue spaces. *Doklady of Russian Academy of Sciences* 119, 861–864.

[75] Kullback, S., Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79–86. https://www.jstor.org/stable/pdf/2236703.

[76] Kullback, S. (1959). *Information Theory and Statistics* (Wiley).

[77] Landsman, K. (2020). Randomness? What randomness?, *Foundations of Physics* 50, 61–104.

[78] Landsman, K. (2023).Typical = Random. *Axioms* 12:727.

[79] Lanford, O.E. (1973). Entropy and equilibrium states in classical statistical mechanics. *Lecture Notes in Physics* 20, 1-113.

[80] Lesne, A. (2014). Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Mathematical Structures in Computer Science* 24:e240311 (63 pages).

[81] Levine, R.D., Tribus, M. (1979). *The Maximum Entropy Formalism (A Conference Held at the Massachusetts Institute of Technology on May 2–4, 1978)* (MIT Press).

[82] Lindenstrauss, J., Olsen, G., Sternfeld, Y. (1978). The Poulsen simplex. *Annales de l'institut Fourier* 28, 91–114. https://doi.org/10.5802/aif.682.

[83] Lindenstrauss, E. (2001). Pointwise theorems for amenable groups. *Inventiones Mathematicae* 146, 259–295. https://doi.org/10.1007/s002220100162.

[84] Li, M., Vitányi, P.M.B. (2008). *An Introduction to Kolmogorov Complexity and Its Applications. Third Edition* (Springer).

[85] Mackey, G. W. (1974). Ergodic theory and its significance for statistical mechanics and probability theory. *Advances in Mathematics* 12, 178–268. https://doi.org/10.1016/S0001-8708(74)80003-4.

[86] McKean, H. (2014). *Probability: The Classical Limit Theorems* (Cambridge University Press).

[87] Moore, C.C. (2015). Ergodic theorem, ergodic theory, and statistical mechanics. *PNAS* 112, 1907–1911. https://doi.org/10.1073/pnas.1421798112.

[88] Müller, I. (2007). *A History of Thermodynamics: The Doctrine of Energy and Entropy* (Springer).

[89] Myrvold, W. (2021). *Beyond Chance and Credence: A Theory of Hybrid Probabilities* (Oxford University Press).

[90] Norton, J.D. (2022). How analogy helped create the new science of thermodynamics. *Synthese* 200:269. https://doi.org/10.1007/s11229-022-03708-9.

[91] Olivieri, E., Vares, M. (2004). *Large Deviations and Metastability* (Elsevier).

[92] Penot, J.-P. (2016). *Analysis: From Concepts to Applications* (Springer).

[93] Pfister, C E. (2002). Thermodynamical aspects of classical lattice systems. *In and Out of Equilibrium: Probability with a Physics Flavor*, ed. Sidoravicius, V., pp. 393–472 (Birkhäuser).

[94] Pitts, J.B. (2021). Conservation of energy: Missing features in its nature and justification and why they matter. *Foundations of Science* 26, 559–584. https://doi.org/10.1007/s10699-020-09657-1.

[95] Planck, M. (1906). *Vorlesungen über die Theorie der Wärmestrahlung* (J.A. Barth).

[96] Porter, C.P. (2012). *Mathematical and Philosophical Perspectives on Algorithmic Randomness*. PhD Thesis, University of Notre Dame. https://www.cpporter.com/wp-content/uploads/2013/08/PorterDissertation.pdf.

[97] Raatikainen, P. (1998). On interpreting Chaitin's incompleteness theorem, *Journal of Philosophical Logic* 27, 569–586.

[98] Rassoul-Agha, F., Seppäläinen, T. (2015). *A Course on Large Deviations with an Introduction to Gibbs Measures* (AMS).

[99] Roberts, B.W. (2022). *Reversing the Arrow of Time* (Cambridge University Press).

[100] Ruelle, D. (2004). *Thermodynamic Formalism: The Mathematical Structure of Equilibrium Statistical Mechanics. Second Edition* (Cambridge University Press).

[101] Saslow, W. M. (2020). A history of thermodynamics: The missing manual. *Entropy* 22:77.

[102] Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal* 27, 379–423, 623–656.

[103] Shields, P.C. (1996). *The Ergodic Theory of Discrete Sample Paths* (AMS).

[104] Simon, B. (1993). *The Statistical Mechanics of Lattice Gases, Volume I* (Princeton University Press).

[105] Simon, B. (2011). *Convexity: An Analytic Viewpoint* (CUP).

[106] Sinai, Ya.G. (1959). On the notion of entropy of a dynamical system. *Doklady of Russian Academy of Sciences* 124, 768–771.

[107] Sklar, L. (1993). *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics* (Cambridge University Press).

[108] Smith, C. (1998). *The Science of Energy: A Cultural History of Energy Physics in Victorian Britain* (The Athlone Press).

[109] Stachel, J. et al. (1990). *The Collected Papers of Albert Einstein, Volume 2: The Swiss Years: Writings, 1900-1909* (Princeton University Press, 1990). https://einsteinpapers.press.princeton.edu/vol2-doc/.

[110] Tao, T. (2007). *254A: Topics in Ergodic Theory*. https://terrytao.wordpress.com/2007/12/14/254a-topics-in-ergodic-theory/.

[111] Touchette, H. (2009). The large deviation approach to statistical mechanics. *Physics Reports* 478, 1–69.

[112] Towsner, H. (2020). Algorithmic randomness in ergodic theory. Franklin & Porter, pp. 40–57.

[113] Truesdell, C. (1980). *The Tragicomical History of Thermodynamics 1822–1854* (Springer).

[114] Uffink, J. (2001). Bluff your way in the second law of thermodynamics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32, 305–394. https://doi.org/10.1016/S1355-2198(01)00016-8.

[115] Uffink, J. (2006). Insuperable difficulties: Einstein's statistical road to molecular physics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 37 36–70. https://doi.org/10.1016/j.shpsb.2005.07.004.

[116] Uffink, J. (2007). Compendium of the foundations of classical statistical physics. *Handbook of the Philosophy of Science. Vol. 2: Philosophy of Physics, Part B*, eds. Butterfield, J., Earman, J., pp. 923–1074 (North-Holland).

[117] Uffink, J. (2022). Boltzmann's work in statistical physics. *The Stanford Encyclopedia of Philosophy (Summer 2022 Edition)*, ed. Zalta, E.N. https://plato.stanford.edu/archives/sum2022/entries/statphys-Boltzmann/.

[118] Van Enter, A.C., Fernández, R., Sokal, A.D. (1993). Regularity properties and pathologies of position-space renormalization-group transformations: Scope and limitations of Gibbsian theory. *Journal of Statistical Physics* 72, 879–1167.

[119] Viana, M. (1997). *Stochastic Dynamics of Deterministic Systems (Vol. 21)*. Rio de Janeiro: IMPA. https://impa.br/wp-content/uploads/2017/04/21_CBM_97_03.pdf.

[120] Viana, M., Oliveira, K. (2016). *Foundations of Ergodic Theory* (Cambridge University Press).

[121] Lambalgen, M. van (1987). *Random Sequences*. PhD Thesis, University of Amsterdam, https://www.academia.edu/23899015/RANDOM_SEQUENCES.

[122] von Plato, J. (1994). *Creating Modern Probability* (Cambridge University Press).

[123] Walters, P. (1982). *An Introduction to Ergodic Theory* (Springer).

[124] Weinberger, P. (2013). The discovery of thermodynamics. *Philosophical Magazine* 93, 2576–2612. https://doi.org/10.1080/14786435.2013.784402.

[125] Wegener, F.D.A. (2009). *A True Proteus: A History of Energy Conservation in German Science and Culture, 1874–1914*. PhD Thesis, Utrecht University. https://dspace.library.uu.nl/handle/1874/36626.