Physics Reports 478 (2009) 1–69

Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

The large deviation approach to statistical mechanics

Hugo Touchette *

School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, UK

ARTICLE INFO

Article history:
Accepted 29 May 2009
Available online 6 June 2009
editor: I. Procaccia

PACS:
05.20.-y
65.40.Gr
02.50.-r
05.40.-a

ABSTRACT

The theory of large deviations is concerned with the exponential decay of probabilities of large fluctuations in random systems. These probabilities are important in many fields of study, including statistics, finance, and engineering, as they often yield valuable information about the large fluctuations of a random system around its most probable state or trajectory. In the context of equilibrium statistical mechanics, the theory of large deviations provides exponential-order estimates of probabilities that refine and generalize Einstein's theory of fluctuations. This review explores this and other connections between large deviation theory and statistical mechanics, in an effort to show that the mathematical language of statistical mechanics is the language of large deviation theory. The first part of the review presents the basics of large deviation theory, and works out many of its classical applications related to sums of random variables and Markov processes. The second part goes through many problems and results of statistical mechanics, and shows how these can be formulated and derived within the context of large deviation theory. The problems and results treated cover a wide range of physical systems, including equilibrium many-particle systems, noise-perturbed dynamics, nonequilibrium systems, as well as multifractals, disordered systems, and chaotic systems. This review also covers many fundamental aspects of statistical mechanics, such as the derivation of variational principles characterizing equilibrium and nonequilibrium states, the breaking of the Legendre transform for nonconcave entropies, and the characterization of nonequilibrium fluctuations through fluctuation relations.

Contents

* Tel.: +44 2078825555.
E-mail address: h.touchette@qmul.ac.uk.

0370-1573/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.physrep.2009.05.002

1. Introduction

The mathematical theory of large deviations initiated by Cramér [1] in the 1930s, and later developed by Donsker and Varadhan [2–5] and by Freidlin and Wentzell [6] in the 1970s, is not a theory commonly studied in physics. Yet it could be argued, without being paradoxical, that physicists have been using this theory for more than a hundred years, and are even responsible for writing down the very first large deviation result [7]. Whenever physicists calculate an entropy function or a free energy function, large deviation theory is at play. In fact, large deviation theory is almost always involved when one studies the properties of many-particle systems, be they equilibrium or nonequilibrium systems. So what are large deviations, and what is the theory that studies these deviations?

If this question were posed to a mathematician who knows about large deviation theory, he or she might reply with one of the following answers:

- A theory dealing with the exponential decay of the probabilities of large deviations in stochastic processes.
- A calculus of exponential-order measures based on the saddle-point approximation or Laplace's method.
- An extension of Cramér's Theorem related to sample means of random variables.
- An extension or refinement of the Law of Large Numbers and Central Limit Theorem.

A physicist, on the other hand, who is minimally acquainted with the concept of large deviations, would probably answer by saying that large deviation theory is

- A generalization of Einstein's fluctuation theory.
- A collection of techniques for calculating entropies and free energies.
- A rigorous expression of saddle-point approximations often used in statistical mechanics.
- A rigorous formulation of statistical mechanics.

These answers do not seem to have much in common, except for the mention of the saddle-point approximation, but they are really all fundamentally related. They differ only in the extent that they refer to two different views of the same theory: one directed at its *mathematical* applications—the other directed at its *physical* applications.

The aim of this review is to explain this point in detail, and to show, in the end, that large deviation theory and statistical mechanics have much in common. Actually, the message that runs through this review is more ambitious: we shall argue, by accumulating several correspondences between statistical mechanics and large deviation theory, that the mathematics of statistical mechanics, as a whole, is the theory of large deviations, in the same way that differential geometry, say, is the mathematics of general relativity.

At the core of all the correspondences that will be studied here is Einstein's idea that probabilities can be expressed in terms of entropy functions. The expression of this idea in large deviation theory is contained in the so-called large deviation principle, and an entropy function in this context is called a rate function. This already explains one of the answers given above: large deviation theory is a generalization of Einstein's fluctuation theory. From this first correspondence follows a string of other correspondences that can be used to build and explain, from a clear mathematical perspective, the basis of statistical mechanics. Large deviation theory explains, for example, why the entropy and free energy functions are mutually connected by a Legendre transform, and so provides an explanation of the appearance of this transform in thermodynamics. Large deviation theory also explains why equilibrium states can be calculated via the extremum principles that are the (canonical) minimum free energy principle and the (microcanonical) maximum entropy principle. In fact, large deviation theory not only justifies these principles, but also provides a prescription for generalizing them to arbitrary macrostates and arbitrary many-particle systems.

These points have already been recognized and "publicized" to some extent by a number of people, who see large deviation theory as the proper mathematical framework in which problems of statistical mechanics can be formulated and solved efficiently and, if need be, rigorously. Ellis [8] is to be credited for providing what is perhaps the most complete expression of this view, in a book that has played a major part in bringing large deviations into physics. The idea that statistical mechanics can be formulated in the language of large deviations has also been expressed in a number of review papers, including one by Oono [9], two by Ellis [7,10], and the seminal paper of Lanford [11], which is considered to be the first work on large deviations and statistical mechanics. Since these works appeared, more applications of large deviations have seen the light, so that the time seems ripe now for a new review. This especially true for the subjects of long-range interaction systems, nonconcave entropies, and nonequilibrium systems, which have all been successfully studied recently using large deviation techniques.

Our efforts in this review will go towards learning about the many applications of large deviation theory in statistical mechanics, but also, and perhaps more importantly, towards learning about large deviation theory itself. The presentation of this theory covers in fact about half of this review, and is divided into three sections. The first presents a series of simple examples that illustrate the basis of the large deviation principle (Section 2). There follows a presentation of large deviation theory proper (Section 3), and a section containing many illustrative examples of this theory (Section 4). These examples are useful, as they illustrate many important points about large deviations that one must be aware of before studying their applications.

The content of these three mathematical sections should overall be understandable by most physicists. A great deal of effort has been put into writing an account of large deviation theory which is devoid of the many mathematical details commonly found in textbooks on large deviations. These efforts have concentrated mainly on avoiding the use of measure theory and topology, and on using the level of rigor that prevails in physics for treating limits and approximations. The result is likely to upset mathematicians, but will surely please physicists who are looking for a theory with which to do calculations. Many mathematical elements that are omitted in the presentation are mentioned in the appendices, as well as in various other sections, which also point to many useful references that treat large deviations at the level of rigor demanded by mathematicians.

The physical applications of large deviations are covered in the second part of this review. The list of applications treated in the three sections that make up this part is not exhaustive, but covers most of the important applications related to equilibrium statistical mechanics (Section 5) and nonequilibrium statistical mechanics (Section 6). The correspondence between large deviation theory and Einstein's fluctuation theory is fully explained in the section dealing with equilibrium system. Other topics discussed in that section include the interpretation of the entropy as a rate function, the derivation

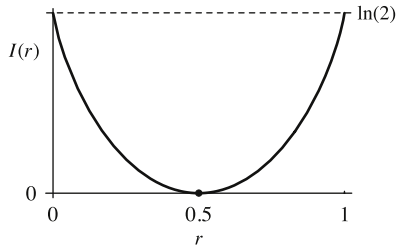

Fig. 1. Rate function $I(r)$ for Example 2.1.

of the Legendre transform connecting the entropy and the free energy, and the derivation of general variational principles that characterize equilibrium states in the microcanonical and canonical ensembles. The topics discussed in the context of nonequilibrium systems are as varied, and include the study of large deviations in stochastic differential equations (Freidlin–Wentzell theory), dynamical models of equilibrium fluctuations (Onsager–Machlup theory), fluctuation relations, and systems of interacting particles. Other applications, related to multifractals, chaotic systems, spin glasses, and quantum systems, are quickly covered in Section 7.

As a warning about the sections covering the physical applications, it should be said that this work is neither a review of statistical mechanics nor a review of large deviation theory—it is a review of the many ways in which large deviation theory can be applied in statistical mechanics. The list of applications treated in this work should be viewed, accordingly, not as a complete list of applications of large deviation theory, but as a selected list or *compendium* of representative examples that should serve as useful points of departure for studying other applications. This is especially true for the examples discussed in the section on nonequilibrium systems (Section 6). At the time of writing this review, a complete theory of nonequilibrium systems is still lacking, so it is difficult to provide a unified presentation of these systems based on large deviation theory. The aim of Section 6 is to give a broad idea of how large deviation techniques can be applied for studying nonequilibrium systems, and to convey a sense that large deviation theory is behind many results related to these systems, just as it is behind many results related to equilibrium systems. One could go further and argue, following Oono [9] and Eyink [12] among others, that large deviation theory is not only useful for studying nonequilibrium systems, but provides the proper basis for building a theory of these systems. Section 6 was written with this idea in mind.

2. Examples of large deviation results

Before we immerse ourselves into the theory of large deviations, it is useful to work out a few examples involving random sums to gain a sense of what large deviations are, and a sense of the context in which these deviations arise. The examples are purposely abstract, but are nonetheless simple. The goal in presenting them is to introduce some basic mathematical ideas and notations that will be used throughout this review. Readers who are already familiar with large deviations may skip this section, and start with Section 3 or even Section 5.

Example 2.1 (*Random Bits*). Consider a sequence $b = (b_1, b_2, \ldots, b_n)$ of n independent random bits taking the value 0 or 1 with equal probability, and define

$$R_n = \frac{1}{n}\sum_{i=1}^{n} b_i \tag{1}$$

to be the fraction of 1's contained in b. We are interested to find the probability $P(R_n = r)$ that R_n assumes one of the (rational) values $0, 1/n, 2/n, \ldots, n/n$. Since the bits are independent and unbiased, we have $P(b) = 2^{-n}$ for all $b \in \{0, 1\}^n$, so that

$$P(R_n = r) = \sum_{b:R_n(b)=r} P(b) = \frac{1}{2^n}\frac{n!}{(rn)![(1-r)n]!}. \tag{2}$$

Using Stirling's approximation, $n! \approx n^n e^{-n}$, we can extract from this result a dominant contribution having the form

$$P(R_n = r) \approx e^{-nI(r)}, \qquad I(r) = \ln 2 + r\ln r + (1-r)\ln(1-r) \tag{3}$$

for n large. The function $I(r)$ entering in the exponential is positive and convex for $r \in [0, 1]$, as shown in Fig. 1, and has a unique zero is located at $r = 1/2$.

The approximation displayed in (3) is an example of large deviation approximation. The exponential-decaying form of this approximation, combined with the expression of the decay or *rate function* $I(r)$, shows that the "unbalanced" sequences of n bits that contain more 0's than 1's, or vice versa, are unlikely to be observed as n gets large because $P(R_n)$ decays

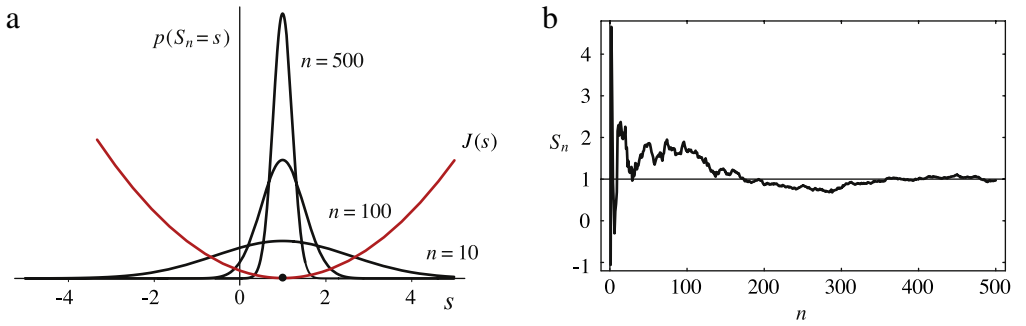

Fig. 2. Gaussian sample mean with $\mu = \sigma = 1$. (a) Probability density $p(S_n = s)$ for increasing values of n together with its corresponding rate function $J(s)$ (red line). (b) Typical realization of S_n converging to its mean.

exponentially with n for $R_n \neq 1/2$. Only the "balanced" sequences such that $R_n \approx 1/2$ have a non-negligible probability to be observed as n becomes large.

The next example discusses a different random sum for which a large deviation approximation also holds.

Example 2.2 (*Gaussian Sample Mean*). The random variable R_n, defined in the previous example as a sum of n random variables scaled by n, is called in mathematics a *sample mean*. In the present example, we consider a similar sample mean, given by

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{4}$$

and assume that the random variables X_i are independent and identically distributed (IID) according to the Gaussian probability density

$$p(X_i = x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x_i-\mu)^2/(2\sigma^2)}. \tag{5}$$

The parameters μ and σ^2 represent, as usual, the mean and variance, respectively, of the X_i's.

The probability density of S_n can be written as the integral

$$p(S_n = s) = \int_{\{x\in\mathbb{R}^n : S_n(x)=s\}} p(x)\,\mathrm{d}x = \int_{\mathbb{R}^n} \delta(S_n(x) - s)\, p(x)\,\mathrm{d}x = \langle \delta(S_n - s)\rangle\,, \tag{6}$$

where $x = (x_1, x_2, \ldots, x_n)$ is the vector of random variables, and

$$p(x) = p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n) \tag{7}$$

their product density. The solution of this integral is, of course,

$$p(S_n = s) = \sqrt{\frac{n}{2\pi\sigma^2}} \mathrm{e}^{-n(s-\mu)^2/(2\sigma^2)}, \tag{8}$$

since a sum of Gaussian random variables is also exactly Gaussian-distributed. A large deviation approximation is obtained from this exact result by neglecting the term $\sqrt{n}$, which is subdominant with respect to the decaying exponential, thereby obtaining

$$p(S_n = s) \approx \mathrm{e}^{-nJ(s)}, \qquad J(s) = \frac{(s-\mu)^2}{2\sigma^2}, \quad s \in \mathbb{R}. \tag{9}$$

The rate function $J(s)$ that we find here is similar to the rate function $I(r)$ found in the first example—it is convex and possesses a single minimum and zero; see Fig. 2(a). As was the case for $I(r)$, the minimum of $J(s)$ has also for effect that, as n grows, $p(S_n = s)$ gets more and more concentrated around the mean μ because the mean is the only point for which $J(s) = 0$, and thus for which $p(S_n = s)$ does not decay exponentially. In mathematics, this concentration property is expressed by the following limit:

$$\lim_{n\to\infty} P(S_n \in [\mu - \delta, \mu + \delta]) = 1, \tag{10}$$

where δ is any positive number. Whenever this limit holds, we say that S_n converges *in probability* to its mean, and that S_n obeys the *Law of Large Numbers*. This point will be studied in more detail in Section 3.

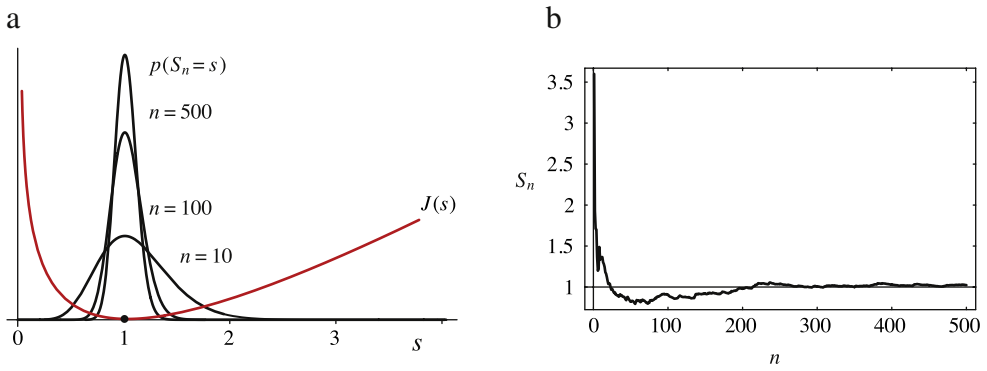

Fig. 3. Exponential sample mean with $\mu = 1$. (a) Probability density $p(S_n = s)$ for increasing values of n together with its corresponding rate function $J(s)$ (red line). (b) Typical realization of S_n converging to its mean.

In general, sums of IID random variables involving different probability distributions for the summands have different rate functions. This is illustrated next.

Example 2.3 (*Exponential Sample Mean*). Consider the sample mean S_n defined before, but now suppose that the IID random variables $X_1, X_2, \ldots, X_n$ are distributed according to the exponential distribution

$$p(X_i = x_i) = \frac{1}{\mu} e^{-x_i/\mu}, \quad x_i > 0, \mu > 0. \tag{11}$$

For this distribution, it can be shown that

$$p(S_n = s) \approx e^{-nJ(s)}, \qquad J(s) = \frac{s}{\mu} - 1 - \ln \frac{s}{\mu}, \quad s > 0. \tag{12}$$

As in the previous examples, the interpretation of the approximation above is that the decaying exponential in n is the dominant term of $p(S_n = s)$ in the limit of large values of n. Notice here that the rate function is different from the rate function of the Gaussian sample mean (Fig. 3(a)), although it is still positive, convex, and has a single minimum and zero located at $s = \mu$ that yields the most probable or *typical* value of S_n in the limit $n \to \infty$; see Fig. 3(b).

The advantage of expressing $p(S_n = s)$ in a large deviation form is that the rate function $J(s)$ gives a direct and detailed picture of the deviations or *fluctuations* of S_n around its typical value. For the Gaussian sample mean, for example, $J(s)$ is a parabola because the fluctuations of S_n around its typical value (the mean μ) are Gaussian-distributed. For the exponential sample mean, by contrast, $J(s)$ has the form of a parabola only around μ, so that only the *small* fluctuations of S_n near its typical value are Gaussian-distributed. The *large* positive fluctuations of S_n that are away from its typical value are not Gaussian; in fact, the form of $J(s)$ shows that they are exponentially-distributed because $J(s)$ is asymptotically linear as $s \to \infty$. This distinction between small and large fluctuations explains the "large" in "large deviation theory", and will be studied in more detail in the next section when discussing the Central Limit Theorem. For now, we turn to another example that shows that large deviation approximations also arise in the context of random vectors.

Example 2.4 (*Symbol Frequencies*). Let $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ be a sequence of IID random variables drawn from the set $\Lambda = \{1, 2, \ldots, q\}$ with common probability distribution $P(\omega_i = j) = \rho_j > 0$. For a given sequence ω, we denote by $L_{n,j}(\omega)$ the relative frequency with which the number or symbol $j \in \Lambda$ appears in ω, that is,

$$L_{n,j}(\omega) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\omega_i,j}, \tag{13}$$

where $\delta_{i,j}$ is the Kronecker symbol. For example, if $\Lambda = \{1, 2, 3\}$ and $\omega = (1, 3, 2, 3, 1, 1)$, then

$$L_{6,1}(\omega) = \frac{3}{6}, \qquad L_{6,2}(\omega) = \frac{1}{6}, \qquad L_{6,3}(\omega) = \frac{2}{6}. \tag{14}$$

The normalized vector[1]

$$L_n(\omega) = \big(L_{n,1}(\omega), L_{n,2}(\omega), \ldots, L_{n,q}(\omega)\big) \tag{15}$$

[1] Vectors are not written in boldface. The vector nature of a quantity should be clear from the context in which it appears.

containing all the symbol frequencies is called the *empirical vector* associated with ω [13]. It is also called the *type* of ω in information theory [14] or the *statistical distribution* of ω in physics. The name "distribution" arises because $L_n(\omega)$ has all the properties of a probability distribution, namely, $0 \leq L_{n,j}(\omega) \leq 1$ for all $j \in \Lambda$, and

$$\sum_{j\in\Lambda} L_{n,j}(\omega) = 1 \tag{16}$$

for all $\omega \in \Lambda^n$. It is important to note, however, that L_n is not a probability; it is a random vector associated with each possible sequence or *configuration* ω, and distributed according to the multinomial distribution

$$P(L_n = l) = \frac{n!}{\prod_{j=1}^{q}(nl_j)!}\prod_{j=1}^{q}\rho_j^{nl_j}. \tag{17}$$

As in Example 2.1, we can extract from this exact result a large deviation approximation by using Stirling's approximation. The result for large values of n is

$$P(L_n = l) \approx \mathrm{e}^{-nI_\rho(l)}, \qquad I_\rho(l) = \sum_{j=1}^{q} l_j \ln\frac{l_j}{\rho_j}. \tag{18}$$

The function $I_\rho(l)$ is called the *relative entropy* or *Kullback–Leibler distance* between the probability vectors l and ρ [14]. As a rate function, $I_\rho(l)$ is slightly more complicated than the rate functions encountered so far, although it shares similar properties. It can be shown, in particular, that $I_\rho(l)$ is positive and convex, and has a single minimum and zero located at $l = \rho$, that is, $l_j = \rho_j$ for all $j \in \Lambda$ (see Chap. 2 of [14]). As before, the zero of the rate function is interpreted as the most probable value of the random variable for which the large deviation result is obtained. This applies for L_n because $P(L_n = l)$ converges to 0 exponentially fast with n for all $l \neq \rho$, since $I_\rho(l) > 0$ for all $l \neq \rho$. The only value of L_n for which $P(L_n = l)$ does not converge exponentially to 0 is $l = \rho$. Hence L_n must converge to ρ in probability as $n \to \infty$.

The next and last example of this section is a simple and classical one in statistical mechanics. It is presented to show that exponential approximations similar to large deviation approximations can be defined for quantities other than probabilities, and that entropy functions are large deviation rate functions in disguise. We will return to these observations, and in particular to the association "entropy = rate function", in Section 5.

Example 2.5 (*Entropy of Non-Interacting Spins*). Consider n spins $\sigma_1, \sigma_2, \ldots, \sigma_n$ taking values in the set $\{-1, 1\}$. It is well known that the number $\Omega(m)$ of spin configurations $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ having a magnetization per spin

$$\frac{1}{n}\sum_{i=1}^{n}\sigma_i \tag{19}$$

equal to m is given by the binomial-like formula

$$\Omega(m) = \frac{n!}{[(1-m)n/2]!\,[(1+m)n/2]!}. \tag{20}$$

The similarity of this result with the one found in the example about random bits should be obvious. As in that example, we can use Stirling's approximation to obtain a large deviation approximation for $\Omega(m)$, which we write as

$$\Omega(m) \approx \mathrm{e}^{ns(m)}, \quad s(m) = -\frac{1-m}{2}\ln\frac{1-m}{2} - \frac{1+m}{2}\ln\frac{1+m}{2}, \ m \in [-1, 1]. \tag{21}$$

The function $s(m)$ is the *entropy* associated with the mean magnetization.

As in the previous example, we can also count the number $\Omega(l)$ of spin configurations containing a relative number l_+ of $+1$ spins and a relative number l_- of -1 spins. These two relative numbers or frequencies are the components of the two-dimensional empirical vector $l = (l_+, l_-)$, for which we find

$$\Omega(l) \approx \mathrm{e}^{n\tilde{s}(l)}, \quad \tilde{s}(l) = -l_+\ln l_+ - l_-\ln l_- \tag{22}$$

for n large. The function $\tilde{s}(l)$, which plays the role of a rate function, is also called the entropy, although it is now the entropy associated with the empirical vector. Notice that since we can express m as a function of l and vice versa, $s(m)$ can be expressed in terms of $\tilde{s}(l)$ and vice versa.

3. Large deviation theory

The cornerstone of large deviation theory is the exponential approximation encountered in the previous examples. This approximation appears so frequently in problems involving many random variables, in particular those studied in statistical mechanics, that we give it a name: *the large deviation principle*. Our goal in this section is to lay down the basis of large deviation theory by first defining the large deviation principle with more care, and by then deriving a number of important consequences of this principle. In doing so, we will see that the large deviation principle is similar to the laws of thermodynamics, in that a few principles—a single one in this case—can be used to derive many far-reaching results. No attempt will be made in this section to integrate or interpret these results within the framework of statistical mechanics; this will come after Section 4.

3.1. The large deviation principle

A basic approximation or scaling law of the form $P_n \approx \mathrm{e}^{-nI}$, where P_n is some probability, n a parameter assumed to be large, and I some positive constant, is referred to as a *large deviation principle*. Such a definition is, of course, only intuitive; to make it more precise, we need to define what we mean exactly by P_n and by the approximation sign "$\approx$". This is done as follows. Let A_n be a random variable indexed by the integer n, and let $P(A_n \in B)$ be the probability that A_n takes on a value in a set B. We say that $P(A_n \in B)$ satisfies a *large deviation principle* with *rate* I_B if the limit

$$\lim_{n\to\infty} -\frac{1}{n}\ln P(A_n \in B) = I_B \tag{23}$$

exists.

The idea behind this limit should be clear. What we mean when writing $P(A_n \in B) \approx \mathrm{e}^{-nI_B}$ is that the dominant behavior of $P(A_n \in B)$ is a decaying exponential in n. Using the small-o notation, this means that

$$-\ln P(A_n \in B) = nI_B + o(n), \tag{24}$$

where I_B is some positive constant. To extract this constant, we divide both sides of the expression above by n to obtain

$$-\frac{1}{n}\ln P(A_n \in B) = I_B + o(1), \tag{25}$$

and pass to the limit $n \to \infty$, so as to get rid of the $o(1)$ contribution. The end result of these steps is the large deviation limit shown in (23). Hence, if $P(A_n \in B)$ has a dominant exponential behavior in n, then that limit should exist with $I_B \neq 0$. If the limit does not exist, then either $P(A_n \in B)$ is too singular to have a limit or else $P(A_n \in B)$ decays with n faster than e^{-na} with $a > 0$. In this case, we say that $P(A_n \in B)$ decays *super-exponentially* and set $I = \infty$. The large deviation limit may also be zero for any set B if $P(A_n \in B)$ is *sub-exponential* in n, that is, if $P(A_n \in B)$ decays with n slower than e^{-na}, $a > 0$. The cases of interest for large deviation theory are those for which the limit shown in (23) does exist with a non-trivial rate exponent, i.e., different from 0 or ∞.

All the examples studied in the previous section fall under the definition of the large deviation principle, but they are more specific in a way because they refer to particular events of the form $A_n = a$ rather than $A_n \in B$. In the case of the random bits, for example, we found that the probability $P(R_n = r)$ satisfied

$$\lim_{n\to\infty} -\frac{1}{n}\ln P(R_n = r) = I(r), \tag{26}$$

with $I(r)$ a continuous function that we called in this context a *rate function*. Similar results were obtained for the Gaussian and exponential sample means, although for these we worked with probability densities rather than probability distributions. The "density" large deviation principles that we obtained can nevertheless be translated into "probability" large deviation principles simply by exploiting the fact that

$$P(S_n \in [s, s + \mathrm{d}s]) = p(S_n = s)\,\mathrm{d}s, \tag{27}$$

where $p(S_n = s)$ is the probability density of S_n, in order to write

$$P(S_n \in [s, s + \mathrm{d}s]) \approx \mathrm{e}^{-nJ(s)}\,\mathrm{d}s. \tag{28}$$

Proceeding with $P(S_n \in [s, s+\mathrm{d}s])$, the rate function $J(s)$ is then recovered, as in the case of discrete probability distributions, by taking the large deviation limit. Thus

$$\lim_{n\to\infty} -\frac{1}{n}\ln P(S_n \in [s, s + \mathrm{d}s]) = J(s) + \lim_{n\to\infty}\frac{1}{n}\ln \mathrm{d}s = J(s), \tag{29}$$

where the last equality follows by assuming that $\mathrm{d}s$ is an arbitrary but non-zero infinitesimal element.

3.2. More on the large deviation principle

The limit defining the large deviation principle, as most limits appearing in this review, should be understood at a practical rather than rigorous level. Likewise, our definition of the large deviation principle should not be taken as a rigorous definition. In fact, it is not. In dealing with probabilities and limits, there are many mathematical subtleties that need to be taken into account (see Appendix B). Most of these subtleties will be ignored in this review, but it may be useful to mention two of them:

- The limit involved in the definition of the large deviation principle may not exist. In this case, one may still be able to find an upper bound and a lower bound on $P(A_n \in B)$ that are both exponential in n:

$$e^{-nI_B^-} \leq P(A_n \in B) \leq e^{-nI_B^+}. \tag{30}$$

  The two bounds give a precise meaning to the statement that $P(A_n \in B)$ is decaying exponentially with n, and give rise to two large deviation principles: one defined in terms of a "limit inferior" yielding I_B^-, and one defined with a "limit superior" yielding I_B^+. This approach, which is the one followed by mathematicians, is described in Appendix B. For the purposes of this review, we make the simplifying assumption that $I_B^- = I_B^+$ always holds; hence our definition of the large deviation principle involving a simple limit.
- Discrete random variables are often treated as if they become continuous in the limit $n \to \infty$. Such a "discrete to continuous" limit, or *continuum limit* as it is known in physics, was implicit in many examples of the previous section. In the first example, for instance, we noted that the proportion R_n of 1's in a random bit sequence of length n could only assume a rational value. As $n \to \infty$, the set of values of R_n becomes dense in $[0, 1]$, so it is useful in this case to picture R_n as being a continuous random variable taking values in $[0, 1]$. Likewise, in Example 2.5 we implicitly treated the mean magnetization m as a continuous variable, even though it assumes only rational values for $n < \infty$. In both examples, the large deviation approximations that we derived were continuous approximations involving continuous rate functions.

The replacement of discrete random variables by continuous random variables is justified mathematically by the notion of weak convergence. Let A_n be a discrete random variable with probability distribution $P(A_n = a)$ defined on a subset of values $a \in \mathbb{R}$, and let $\tilde{A}_n$ be a continuous random variable with probability density $p(\tilde{A}_n)$ defined on $\mathbb{R}$. To say that A_n converges weakly to $\tilde{A}_n$ means, essentially, that any sum involving A_n can be approximated, for n large, by integrals involving $\tilde{A}_n$, i.e.,

$$\sum_a f(a) P(A_n = a) \overset{n\to\infty}{\approx} \int f(a)\, p(\tilde{A}_n = a)\, \mathrm{d}a, \tag{31}$$

where f is any continuous and bounded function defined over $\mathbb{R}$. This sort of approximation is common in physics, and suggests the following replacement rule:

$$P(A_n = a) \longrightarrow p(\tilde{A}_n = a)\, \mathrm{d}a \tag{32}$$

as a formal device for taking the continuum limit of A_n. For more information on the notion of weak convergence, the reader is referred to [15,16] and Appendix B of this review.

Most of the random variables considered in this review, and indeed in large deviation theory, are either discrete random variables that weakly converge to continuous random variables or are continuous random variables right from the start. To treat these two cases with the same notation, we will try to avoid using probability densities whenever possible, to consider instead probabilities of the form $P(A_n \in [a, a + \mathrm{d}a])$. To further cut in the notations, we will also avoid using a tilde for distinguishing a discrete random variable from its continuous approximation, as done above with A_n and $\tilde{A}_n$. From now on we thus write

$$P(A_n \in [a, a + \mathrm{d}a]) \approx e^{-nI(a)} \mathrm{d}a \tag{33}$$

to mean that A_n, whether discrete or continuous, satisfies a large deviation principle. This choice of notation is convenient but arbitrary: readers who prefer probability densities may express a large deviation principle for A_n in the density form $p(A_n = a) \approx e^{-nI(a)}$ instead of the expression shown in (33). In this way, one need not bother with the infinitesimal element $\mathrm{d}a$ in the statement of the large deviation principle. In this review, we will use the probability notation shown in (33), which has to include the infinitesimal element $\mathrm{d}a$, even though this element is not exponential in n. Indeed, without the element $\mathrm{d}a$, the following expectation value would not make sense:

$$\langle f(A_n) \rangle = \int f(a)\, P(A_n \in [a, a + \mathrm{d}a]) \asymp \int f(a)\, e^{-nI(a)}\, \mathrm{d}a. \tag{34}$$

There are two final pieces of notation that need to be introduced before we go deeper into the theory of large deviations. First, we will use the more compact expression $P(A_n \in \mathrm{d}a)$ to mean $P(A_n \in [a, a + \mathrm{d}a])$. Next, we will follow Ellis [10] and use the sign "$\asymp$" instead of "$\approx$" whenever we treat large deviation principles. In the end, we thus write

$$P(A_n \in \mathrm{d}a) \asymp e^{-nI(a)}\, \mathrm{d}a \tag{35}$$

to mean that A_n satisfies a large deviation principle, in the sense of (23), with rate function $I(a)$. The sign "$\asymp$" is used to stress that, as $n \to \infty$, the dominant part of $P(A_n \in \mathrm{d}a)$ is the decaying exponential $e^{-nI(a)}$. We may also interpret the sign "$\asymp$" as

expressing an equality relationship on a logarithmic scale; that is, we may interpret $a_n \asymp b_n$ as meaning that

$$\lim_{n\to\infty} \frac{1}{n} \ln a_n = \lim_{n\to\infty} \frac{1}{n} \ln b_n. \tag{36}$$

We say in this case that a_n and b_n are equal up to first order in their exponents [14].

3.3. Calculating rate functions

The theory of large deviations can be described from a practical point of view as a collection of methods that have been developed and gathered together in one toolbox to solve two problems [13]:

- Establish that a large deviation principle exists for a given random variable.
- Derive the expression of the associated rate function.

Both of these problems can be addressed, as we have done in the examples of the previous section, by directly calculating the probability distribution of a random variable, and by deriving from this distribution a large deviation approximation using Stirling's approximation or other asymptotic formulae. In general, however, it may be difficult or even impossible to derive large deviation principles through this direct calculation path. Combinatorial methods based on Stirling's approximation cannot be used, for example, for continuous random variables, and become quite involved when dealing with sums of discrete random variables that are non-IID. For these cases, a more general calculation path is provided by a fundamental result of large deviation theory known as the Gärtner–Ellis Theorem [17,18]. What we present next is a simplified version of that theorem, which is sufficient for the applications covered in this review; for a more complete presentation, see Sec. 5 of [10] and Sec. 2.3 of [13].

3.3.1. The Gärtner–Ellis Theorem

Consider a real random variable A_n parameterized by the positive integer n, and define the *scaled cumulant generating function* of A_n by the limit

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{nkA_n} \right\rangle, \tag{37}$$

where $k \in \mathbb{R}$ and

$$\left\langle \mathrm{e}^{nkA_n} \right\rangle = \int_{\mathbb{R}} \mathrm{e}^{nka}\, P(A_n \in \mathrm{d}a). \tag{38}$$

The Gärtner–Ellis Theorem states that, if $\lambda(k)$ exists and is differentiable for all $k \in \mathbb{R}$, then A_n satisfies a large deviation principle, i.e.,

$$P(A_n \in \mathrm{d}a) \asymp \mathrm{e}^{-nI(a)}\, \mathrm{d}a, \tag{39}$$

with a rate function $I(a)$ given by

$$I(a) = \sup_{k\in\mathbb{R}} \{ka - \lambda(k)\}. \tag{40}$$

The symbol "sup" above stands for "supremum of", which for us can be taken to mean the same as "maximum of". The transform defined by the supremum is an extension of the Legendre transform referred to as the *Legendre–Fenchel transform* [19]. The Gärtner–Ellis Theorem thus states in words that, when the scaled cumulant generating function $\lambda(k)$ of A_n is differentiable, then A_n obeys a large deviation principle with a rate function $I(a)$ given by the Legendre–Fenchel transform of $\lambda(k)$.

The next sections will show how useful the Gärtner–Ellis Theorem is for calculating rate functions. It is important to know, however, that not all rate functions can be calculated with this theorem. Some examples of rate functions that cannot be calculated as the Legendre–Fenchel transform of $\lambda(k)$, even though $\lambda(k)$ exists, will be studied in Section 4.4. The argument presented next is meant to give some insight as to why $I(a)$ can be expressed as the Legendre–Fenchel transform of $\lambda(k)$ when $\lambda(k)$ is differentiable. A full understanding of this argument will also come in Section 4.4.

3.3.2. Plausibility argument for the Gärtner–Ellis Theorem

Two different derivations of the Gärtner–Ellis Theorem are given in Appendix C. To gain some insight into this theorem, we derive here the second part of this theorem, namely Eq. (40), by assuming that a large deviation principle holds for A_n, and by working out the consequences of this assumption. To start, we thus assume that

$$P(A_n \in \mathrm{d}a) \asymp \mathrm{e}^{-nI(a)}\, \mathrm{d}a, \tag{41}$$

and insert this approximation into the expectation value defined in Eq. (38) to obtain

$$\left\langle \mathrm{e}^{nkA_n} \right\rangle \asymp \int_{\mathbb{R}} \mathrm{e}^{n[ka-I(a)]}\, \mathrm{d}a. \tag{42}$$

Next, we approximate the integral by its largest integrand, which is found by locating the maximum of $ka - I(a)$. This approximation, which is known as the *saddle-point approximation* or *Laplace's approximation*,[2] is a natural approximation to consider here because the error associated with it is of the same order as the error associated with the large deviation approximation itself. Therefore, assuming that the maximum of $ka - I(a)$ exists and is unique, we write

$$\left\langle \mathrm{e}^{nkA_n} \right\rangle \asymp \exp\left(n \sup_{a\in\mathbb{R}} \{ka - I(a)\} \right) \tag{43}$$

and so

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{nkA_n} \right\rangle = \sup_{a\in\mathbb{R}} \{ka - I(a)\}. \tag{44}$$

To obtain $I(a)$ in terms of $\lambda(k)$, we then use the fact that Legendre–Fenchel transforms can be inverted when $\lambda(k)$ is everywhere differentiable (see Sec. 26 of [19]). In this case, the Legendre–Fenchel transform is *self-inverse* (we also say *involutive* or *self-dual*), so that

$$I(a) = \sup_{k\in\mathbb{R}} \{ka - \lambda(k)\}, \tag{45}$$

which is the result of Eq. (40).

This heuristic derivation illustrates two important points about large deviation theory. The first is that Legendre–Fenchel transforms appear into this theory as a natural consequence of Laplace's approximation. The second is that the Gärtner–Ellis Theorem is essentially a consequence of the large deviation principle combined with Laplace's approximation. This point is illustrated in Appendix C, and will be discussed again in the context of another important result of large deviation theory known as Varadhan's Theorem.

3.4. Cramér's Theorem

The application of the Gärtner–Ellis Theorem to a sample mean

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{46}$$

of independent and identically distributed (IID) random variables yields a classical result of probability theory known as *Cramér's Theorem* [1]. In this case, the scaled cumulant generating function has the simple form

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{k\sum_{i=1}^{n} X_i} \right\rangle = \lim_{n\to\infty} \frac{1}{n} \ln \prod_{i=1}^{n} \left\langle \mathrm{e}^{kX_i} \right\rangle = \ln \left\langle \mathrm{e}^{kX} \right\rangle, \tag{47}$$

where X is any of the summands X_i. As a result, one derives a large deviation principle for S_n simply by calculating the *cumulant generating function* $\ln\langle \mathrm{e}^{kX} \rangle$ of a single summand, and by taking the Legendre–Fenchel transform of the result. The next examples illustrate these steps. Note that the differentiability condition of the Gärtner–Ellis Theorem need not be checked for IID sample means because the *generating function* or *Laplace transform* $\langle \mathrm{e}^{kX} \rangle$ of a random variable X is always real analytic when it exists for all $k \in \mathbb{R}$ (see Theorem VII.5.1 of [8]).

Example 3.1 (*Gaussian Sample Mean Revisited*). Consider again the sample mean S_n of n Gaussian IID random variables considered in Example 2.2. For the Gaussian density of Eq. (5), $\lambda(k)$ is easily evaluated to be

$$\lambda(k) = \ln \left\langle \mathrm{e}^{kX} \right\rangle = \mu k + \frac{1}{2}\sigma^2 k^2, \quad k \in \mathbb{R}. \tag{48}$$

As expected, $\lambda(k)$ is everywhere differentiable, so that $P(S_n \in \mathrm{d}s) \asymp \mathrm{e}^{-nI(s)}\, \mathrm{d}s$ with

$$I(s) = \sup_{k} \{ks - \lambda(k)\}. \tag{49}$$

[2] The saddle-point approximation is used in connection with integrals in the complex plane, whereas Laplace's approximation or Laplace's method is used in connection with real integrals (see Chap. 6 of [20]).

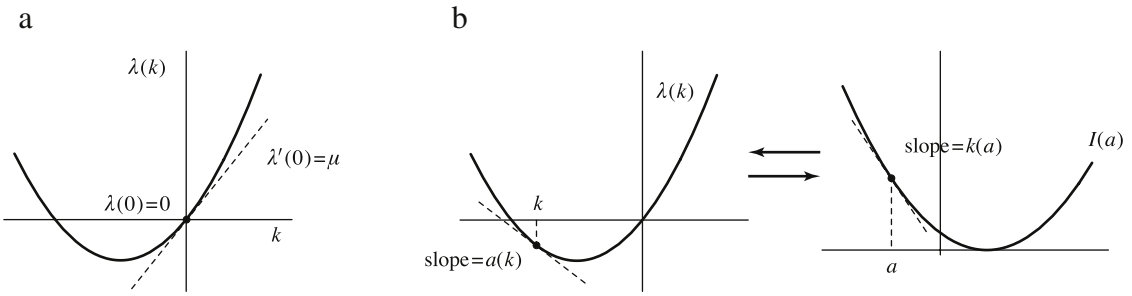

Fig. 4. (a) Properties of $\lambda(k)$ at $k = 0$. (b) Legendre duality: the slope of λ at k is the point at which the slope of I is k.

This recovers Cramér's Theorem. The supremum defining the Legendre–Fenchel transform is solved directly by ordinary calculus. The result is

$$I(s) = k(s)s - \lambda(k(s)) = \frac{(s-\mu)^2}{2\sigma^2}, \quad s \in \mathbb{R}, \tag{50}$$

where $k(s)$ is the unique maximum point of $ks - \lambda(k)$ satisfying $\lambda'(k) = s$. This recovers exactly the result of Example 2.2 knowing that $P(S_n \in \mathrm{d}s) = p(S_n = s)\mathrm{d}s$.

Example 3.2 (*Exponential Sample Mean Revisited*). The calculation of the previous example can be carried out for the exponential sample mean studied in Example 2.3. In this case, we find

$$\lambda(k) = -\ln(1-\mu k), \quad k < 1/\mu. \tag{51}$$

From Cramér's Theorem, we then obtain $P(S_n \in \mathrm{d}s) \asymp \mathrm{e}^{-nI(s)}\,\mathrm{d}s$, where

$$I(s) = \sup_k\{ks - \lambda(k)\} = k(s)s - \lambda(k(s)) = \frac{s}{\mu} - 1 - \ln\frac{s}{\mu}, \quad s > 0, \tag{52}$$

in agreement with the result announced in (12). It is interesting to note here that the singularity of $\lambda(k)$ at $1/\mu$ translates into a branch of $I(s)$ which is asymptotically linear. This branch of $I(s)$ translates, in turn, into a tail of $P(S_n \in \mathrm{d}s)$ which is asymptotically exponential. If the probability density of the IID random variables is chosen to be a double-sided rather than a single-sided exponential distribution, then both tails of $P(S_n \in \mathrm{d}s)$ become asymptotically exponential.

We will study other examples of IID sums, as well as sums involving non-IID random variables in Section 4. It should be clear at this point that the scope of the Gärtner–Ellis Theorem is not limited to IID random variables. In principle, the theorem can be applied to any random variable, provided that one can calculate the limit defining $\lambda(k)$ for that random variable, and that $\lambda(k)$ satisfies the conditions of the theorem. Examples of sample means of random variables for which $\lambda(k)$ fail to meet these conditions will be presented also in Section 4.

3.5. Properties of λ and I

We now state and prove a number of properties of scaled cumulant generating functions and rate functions in the case where the latter is obtained via the Gärtner–Ellis Theorem. The properties listed hold for an arbitrary random variable A_n under the conditions stated, not just sample means of IID random variables.

3.5.1. Properties of λ at $k = 0$

Since probability measures are normalized, $\lambda(0) = 0$. Moreover,

$$\lambda'(0) = \lim_{n\to\infty} \left.\frac{\langle A_n \mathrm{e}^{nkA_n}\rangle}{\langle \mathrm{e}^{nkA_n}\rangle}\right|_{k=0} = \lim_{n\to\infty}\langle A_n\rangle, \tag{53}$$

provided that $\lambda'(0)$ exists. For IID sample means, this reduces to $\lambda'(0) = \langle X\rangle = \mu$; see Fig. 4(a). Similarly,

$$\lambda''(0) = \lim_{n\to\infty} n\left(\langle A_n^2\rangle - \langle A_n\rangle^2\right) = \lim_{n\to\infty} n\,\mathrm{var}(A_n), \tag{54}$$

which reduces to $\lambda''(0) = \mathrm{var}(X) = \sigma^2$ for IID sample means.

3.5.2. Convexity of λ

The function $\lambda(k)$ is always convex. This comes as a general consequence of Hölder's inequality:

$$\sum_i |y_i z_i| \leq \left(\sum_i |y_i|^{1/p}\right)^p \left(\sum_i |z_i|^{1/q}\right)^q, \tag{55}$$

where $0 \leq p, q \leq 1, p + q = 1$. Applying this inequality to $\lambda(k)$ yields

$$\alpha \ln \left\langle e^{nk_1 A_n}\right\rangle + (1-\alpha) \ln \left\langle e^{nk_2 A_n}\right\rangle \geq \ln \left\langle e^{n[\alpha k_1 + (1-\alpha)k_2]A_n}\right\rangle \tag{56}$$

for $\alpha \in [0, 1]$. Hence,

$$\alpha\lambda(k_1) + (1-\alpha)\lambda(k_2) \geq \lambda\left(\alpha k_1 + (1-\alpha)k_2\right). \tag{57}$$

A particular case of this inequality, which defines a function as being convex [19], is $\lambda(k) \geq k\lambda'(0) = k\mu$; see Fig. 4(a). Note that the convexity of $\lambda(k)$ directly implies that $\lambda(k)$ is continuous in the interior of its domain, and is differentiable everywhere except possibly at a denumerable number of points [19,21].

3.5.3. Legendre transform and Legendre duality

We have seen when calculating the rate functions of the Gaussian and exponential sample means that the Legendre–Fenchel transform involved in the Gärtner–Ellis Theorem reduces to

$$I(a) = k(a)a - \lambda(k(a)), \tag{58}$$

where $k(a)$ is the unique root of $\lambda'(k) = a$. This equation plays a central role in this review: it defines, as is well known, the *Legendre transform* of $\lambda(k)$, and arises in the examples considered before because $\lambda(k)$ is everywhere differentiable, as required by the Gärtner–Ellis Theorem, and because $\lambda(k)$ is convex, as proved above. These conditions—differentiability and convexity—are the two essential conditions for which the Legendre–Fenchel transform reduces to the better known Legendre transform (see Sec. 26 of [19]).

An important property of Legendre transforms holds when $\lambda(k)$ is differentiable and is *strictly convex*, that is, convex with no linear parts. In this case, $\lambda'(k)$ is monotonically increasing, so that the function $k(a)$ satisfying $\lambda'(k(a)) = a$ can be inverted to obtain a function $a(k)$ satisfying $\lambda'(k) = a(k)$. From the equation defining the Legendre transform, we then have $I'(a(k)) = k$ and $I'(a) = k(a)$. Therefore, in this case—and this case only—the slopes of λ are one-to-one related to the slopes of I. This property, which we refer as the *duality property* of the Legendre transform, is illustrated in Fig. 4(b).

The next example shows how border points where $\lambda(k)$ diverges translate, by Legendre duality, into branches of $I(a)$ that are linear or asymptotically linear.[3] A specific random variable for which this duality behavior shows up is the sample mean of exponential random variables studied in Example 3.2. Since we can invert the roles of $\lambda(k)$ and $I(a)$ in the Legendre transform, this example can also be generalized to show that points where $I(a)$ diverges are associated with branches of $\lambda(k)$ that are linear or asymptotically linear; see Example 2.1. These sorts of diverging points and linear branches arise often in physical applications, for example, in relation to nonequilibrium fluctuations; see Section 6.3.

Example 3.3. Consider the scaled cumulant generating function $\lambda(k)$ shown in Fig. 5(a). This function has the particularity that it is defined only on a bounded (open) interval (k_l, k_h), and has diverging slopes at the boundaries, that is, $\lambda'(k) \to \infty$ as k approaches k_h from below and $\lambda'(k) \to -\infty$ as k approaches k_l from above. To determine the shape of the Legendre transform of $\lambda(k)$, which corresponds to the rate function $I(a)$ associated with $\lambda(k)$ (assume that $\lambda(k)$ is everywhere differentiable), we simply need to use Legendre duality. On the one hand, since the slope of $\lambda(k)$ diverges as k approaches k_h, the slope of $I(a)$ must approach the constant k_h as $a \to \infty$ (remember that slopes of λ are abscissas of I). On the other hand, since the slope of $\lambda(k)$ goes to $-\infty$ as k approaches k_l, the slope of $I(a)$ must approach the constant k_l as $a \to -\infty$. Overall, $I(a)$ is thus asymptotically linear; see Fig. 5(b).

Now suppose that rather than having diverging slopes at the boundaries k_l and k_h, $\lambda(k)$ has finite slopes a_l and a_h, respectively; see Fig. 5(c). What is the rate function $I(a)$ associated with this form of $\lambda(k)$? The answer, surprisingly, is that there is not one but *many* rate functions that may correspond to this $\lambda(k)$. One such rate function is the Legendre–Fenchel transform of $\lambda(k)$ shown in Fig. 5(d). This function has the particularity that it has two linear branches which arise, as before, because of the two boundary points of $\lambda(k)$. The difference here is that these branches are really linear, and not just *asymptotically* linear, because the left-derivative of $\lambda(k)$ at k_h is finite, and so is its right-derivative at k_l. To understand why these linear branches appear, one must appeal to a generalization of Legendre duality involving the concept of "supporting lines" [19]. We will not discuss this concept here; suffice it to say that the value of the left-derivative of $\lambda(k)$ at k_h corresponds to the starting point of the linear branch of $I(a)$ with slope k_h. Similarly, the right-derivative of $\lambda(k)$ at k_l corresponds to the endpoint of the linear branch of $I(a)$ with slope k_l.

The reason why the rate function shown in Fig. 5(d) is but one candidate rate function associated with the $\lambda(k)$ shown in Fig. 5(c) is explained in Section 4.4. The reason has to do, essentially, with the fact that $\lambda(k)$ is nondifferentiable at its boundaries. In large deviation theory, one says more precisely that $\lambda(k)$ is *non-steep*; see the notes at the end of this section for more information about this concept.

[3] Recall that, because $\lambda(k)$ is a convex function, it cannot have diverging points in the interior of its domain.

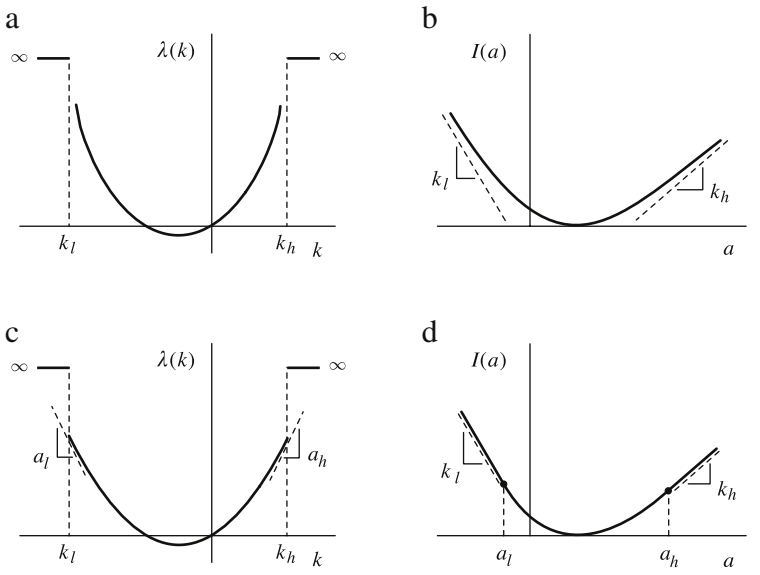

Fig. 5. (a) Scaled cumulant generating function $\lambda(k)$ defined on the open domain (k_l, k_h), with diverging slopes at the boundaries. (b) The Legendre transform $I(a)$ of $\lambda(k)$ is asymptotically linear as $|k| \to \infty$. The asymptotic slopes correspond to the boundaries of the region of convergence of $\lambda(k)$. (c) $\lambda(k)$ is defined on (k_l, k_h) as in (a) but has finite slopes at the boundaries. (d) Legendre–Fenchel transform $I(a)$ of the function $\lambda(k)$ shown in (c). The function $I(a)$ has branches that are linear rather than just asymptotically linear, with slopes corresponding to the boundaries of the region of convergence of $\lambda(k)$.

3.5.4. Varadhan's Theorem

In our heuristic derivation of the Gärtner–Ellis Theorem, we showed that if A_n satisfies a large deviation principle with rate function $I(a)$, then $\lambda(k)$ is the Legendre-fenchel transform of $I(a)$:

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \left\langle \mathrm{e}^{nkA_n} \right\rangle = \sup_a \{ka - I(a)\}. \tag{59}$$

Replacing the product kA_n by an arbitrary continuous function f of A_n yields the more general result

$$\lambda(f) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{nf(A_n)} \right\rangle = \sup_a \{f(a) - I(a)\}, \tag{60}$$

which is known as *Varadhan's Theorem* [22]. The function $\lambda(f)$ thus defined is a *functional* of f, as it is a function of the function f.

As we did for the result shown in (59), we can justify (60) as a consequence of the large deviation principle for A_n and Laplace's approximation. It is important to note, however, that Varadhan's Theorem is a consequence of Laplace's approximation only when A_n is a real random variable; for other types of random variables, such as random functions, Varadhan's Theorem still applies, and so *extends* Laplace's approximation to these random variables. Varadhan's Theorem also holds when $f(a) - I(a)$ has more than one maximum, that is, when the integral defining the expected value $\langle \mathrm{e}^{nf(A_n)} \rangle$ has more than one saddle-point. We will come back to this point in Section 4 when discussing nonconvex rate functions, and again in Section 5 when discussing nonconcave entropies.

3.5.5. Positivity of rate functions

Rate functions are always positive. This follows by noting that $\lambda(0) = 0$ and that $\lambda(k)$ can always be expressed as the Legendre–Fenchel transform of $I(a)$. Hence,

$$\lambda(0) = \sup_a \{-I(a)\} = -\inf_a I(a) = 0, \tag{61}$$

where "inf" denotes the "infimum of". A negative rate function would imply that $P(A_n \in \mathrm{d}a)$ diverges as $n \to \infty$.

3.5.6. Convexity of rate functions

Rate functions obtained from the Gärtner–Ellis Theorem are necessarily *strictly convex*, that is, they are convex and have no linear parts.[4] That Legendre–Fenchel transforms yield convex functions is easily proved from the definition of

[4] This does not mean that all rate functions are strictly convex—only that those obtained from the Gärtner–Ellis Theorem are strictly convex.

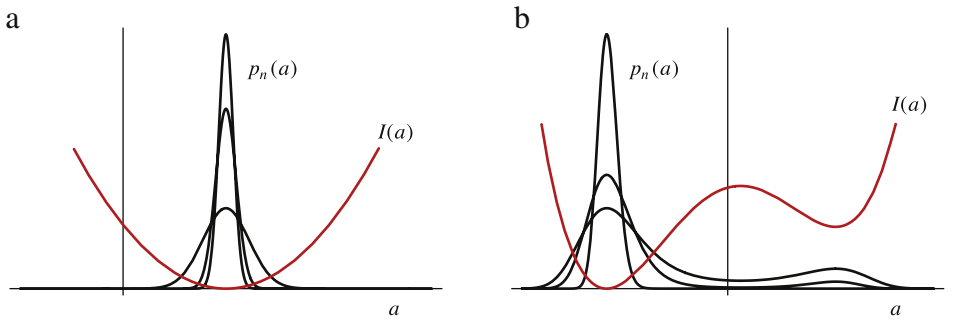

Fig. 6. (a) Example of a unimodal probability density $p_n(a)$ shown for increasing values of n (black line), and its corresponding convex rate function $I(a)$ (red line). (b) Example of a bimodal probability density $p_n(a)$ shown for increasing values of n characterized by a nonconvex rate function $I(a)$ having a local minimum in addition to its global minimum.

these transforms [21]. To prove that they yield *strictly* convex functions when $\lambda(k)$ is differentiable is another matter; see, e.g., Sec. 26 of [19]. As a special case of interest, let us assume that $\lambda(k)$ is differentiable and has no linear parts, as in our discussion of the Legendre duality property. In this case, the Legendre–Fenchel transform reduces to a Legendre transform, as noted earlier, and the equation defining the Legendre transform then implies

$$I''(a) = k'(a) = \frac{1}{\lambda''(k)}. \tag{62}$$

Since $\lambda(k)$ is convex with no linear parts ($\lambda''(k) > 0$), $I(a)$ must then also be convex with no linear parts ($I''(a) > 0$). This shows, incidentally, that the curvature of $I(a)$ is the inverse curvature of $\lambda(k)$. In the case of IID sample means, in particular,

$$I''(a = \mu) = \frac{1}{\lambda''(0)} = \frac{1}{\sigma^2}. \tag{63}$$

A similar result holds for non-IID random variables by replacing σ^2 with the general result of Eq. (54).

3.5.7. Law of large numbers

If $I(a)$ has a unique global minimum and zero a^*, then

$$a^* = \lambda'(0) = \lim_{n\to\infty} \langle A_n \rangle. \tag{64}$$

by Eq. (53). If $I(a)$ is differentiable at a^*, we further have $I'(a^*) = k(a^*) = 0$. To prove this property, simply apply the Legendre duality property:

$$I(a^*) = k(a^*)a^* - \lambda(k(a^*)) = 0 \cdot a^* - 0 = 0. \tag{65}$$

The global minimum and zero of $I(a)$ has a special property that we noticed already: it corresponds, if it is unique, to the value at which $P(A_n \in \mathrm{d}a)$ does not decay exponentially, and so around which $P(A_n \in \mathrm{d}a)$ gets more and more concentrated as $n \to \infty$; see Fig. 6(a). Because of the concentration effect, we have

$$\lim_{n\to\infty} P(A_n \in \mathrm{d}a^*) = \lim_{n\to\infty} P(A_n \in [a^*, a^* + \mathrm{d}a]) = 1, \tag{66}$$

as noted already in Eq. (10), and so we call a^* the *most probable* or *typical* value of A_n. The existence of this typical value is an expression of the Law of Large Numbers, which states in its weak form that $A_n \to a^*$ with probability 1. An important observation here is that large deviation theory extends the Law of Large Numbers by providing information as to *how fast* A_n converges in probability to its mean. To be more precise, let B be any set of values of A_n. Then

$$P(A_n \in B) = \int_B P(A_n \in \mathrm{d}a) \asymp \int_B \mathrm{e}^{-nI(a)}\,\mathrm{d}a \asymp \mathrm{e}^{-n \inf_{a\in B} I(a)} \tag{67}$$

by applying Laplace's approximation. Therefore, $P(A_n \in B) \to 0$ exponentially fast with n if $a^* \notin B$, which means that $P(A_n \in B) \to 1$ exponentially fast with n if $a^* \in B$.

In general, the existence of a Law of Large Numbers for a random variable A_n is a good sign that a large deviation principle holds for A_n. In fact, this law can often be used as a point of departure for deriving large deviation principles; see [23,24] and Appendix C. It should be emphasized, however, that $I(a)$ may have more than one global minimum, in which case the Law of Large Numbers may not hold. Rate functions may even have local minima in addition to global ones. The global minima yield typical values of A_n just as in the case of a single minimum, whereas the local minima yield what physicists would call "metastable" values of A_n at which $P(A_n \in \mathrm{d}a)$ is locally but not globally maximum; see Fig. 6(b). Physicists would also call a typical value of A_n, determined by a global minimum of $I(a)$, an "equilibrium state". We will come to this language in Section 5.

3.5.8. Gaussian fluctuations and the Central Limit Theorem

The Central Limit Theorem arises in large deviation theory when a convex rate function $I(a)$ possesses a single global minimum and zero a^*, and is twice differentiable at a^*. Approximating $I(a)$ with the first quadratic term,

$$I(a) \approx \frac{1}{2} I''(a^*)(a - a^*)^2, \tag{68}$$

then naturally leads to the Gaussian approximation

$$P(A_n \in \mathrm{d}a) \approx \mathrm{e}^{-nI''(a^*)(a-a^*)^2/2}\, \mathrm{d}a, \tag{69}$$

which can be thought of as a weak form of the Central Limit Theorem. More precise results relating the Central Limit Theorem to the large deviation principle can be found in [25,26]. We recall that for sample means of IID random variables, $I''(a^*) = 1/\lambda''(0) = 1/\sigma^2$; see Section 3.5.6.

The Gaussian approximation displayed above can be shown to be accurate for values of A_n around a^* of the order $O(n^{-1/2})$ or, equivalently, for values of nA_n around a^* of the order $O(n^{1/2})$. This explains the meaning of the name "large deviations". On the one hand, a *small* deviation of A_n is a value $A_n = a$ for which the quadratic expansion of $I(a)$ is a good approximation of $I(a)$, and for which, therefore, the Central Limit Theorem yields essentially the same information as the large deviation principle. On the other hand, a *large* deviation is a value $A_n = a$ for which $I(a)$ departs sensibly from its quadratic approximation, and for which, therefore, the Central Limit Theorem yields no useful information about the large fluctuations of A_n away from its mean. In this sense, large deviation theory can be seen as a generalization of the Central Limit Theorem characterizing the small as well as the large fluctuations of a random variable. Large deviation theory also generalizes the Central Limit Theorem whenever $I(a)$ exists but has no quadratic Taylor expansion around its minimum; see Examples 5.4 and 5.6. Note finally that having a Central Limit Theorem for A_n does not imply that $I(a)$ has a quadratic minimum. A classic counterexample is presented next.

Example 3.4 (*Sample Mean of Double-Sided Pareto Random Variables*). Let S_n be a sample mean of n IID random variables $X_1, X_2, \ldots, X_n$ distributed according to the so-called Pareto density

$$p(x) = \frac{A}{(|x| + c)^\beta}, \tag{70}$$

with $\beta > 3$, c a real, positive constant, and A a normalization constant. Since the variance of the summands is finite for $\beta > 3$, the Central Limit Theorem holds for $n^{1/2}S_n$. Yet it can be verified that the rate function of S_n is everywhere equal to zero because the probability density of S_n has power-law tails similar to those of $p(x)$ [11]. Note also that the scaled generating function $\lambda(k)$ is diverging for all $k \in \mathbb{R}$ except $k = 0$.

We will study in the next section another example of sample mean involving a power-law probability density similar to the Pareto density. This time, the power-law density will be one-sided rather than double-sided, and the rate function will be seen to be different from zero for some values of the sample mean.

3.6. Contraction principle

We have seen at this point two basic results of large deviation theory. The first is the Gärtner–Ellis Theorem, which can be used to prove that a large deviation principle exists and to calculate the associated rate function from the knowledge of $\lambda(k)$. The second result is Varadhan's Theorem, which can be used to calculate $\lambda(k)$ from the knowledge of a rate function. The last result that we now introduce is a useful calculation device, called the *contraction principle* [5], which can be used to calculate a rate function from the knowledge of another rate function.

The problem addressed by the contraction principle is the following. We have a random variable A_n satisfying a large deviation principle with rate function $I_A(a)$, and we want to find the rate function of another random variable B_n such that $B_n = h(A_n)$, where h is a continuous function. We call h a *contraction* of A_n, as this function may be many-to-one. To calculate the rate function of B_n from that of A_n, we simply use the large deviation principle for A_n and Laplace's approximation at the level of

$$P(B_n \in db) = \int_{\{a:h(a)=b\}} P(A_n \in \mathrm{d}a) \tag{71}$$

to obtain

$$P(B_n \in db) \asymp \exp\left(-n \inf_{a:h(a)=b} I_A(a)\right) \mathrm{d}a. \tag{72}$$

This shows that if a large deviation principle holds for A_n with rate function $I_A(a)$, then a large deviation principle also holds for B_n,

$$P(B_n \in db) \asymp \mathrm{e}^{-nI_B(b)}\, \mathrm{d}b, \tag{73}$$

with a rate function given by

$$I_B(b) = \inf_{a:h(a)=b} I_A(a). \tag{74}$$

This general reduction of one rate function to another is what is called the contraction principle. If h is a bijective function with inverse h^{-1}, then $I_B(b) = I_A(h^{-1}(b))$. Note also that $I_B(b) = \infty$ if there is no value a such that $h(a) = b$, i.e., if the pre-image of b is empty.

The interpretation of the contraction principle should be clear. Since probabilities in large deviation theory are measured on the exponential scale, the probability of any large fluctuation should be approximated, following Laplace's approximation, by the probability of the most probable (although improbable) event leading or giving rise to that fluctuation. We will see many applications of this idea in the next sections, including a derivation of the maximum entropy principle based on the contraction principle. The "least improbable" event underlying or leading to a large deviation—be it a "state" underlying a large deviation or a "path" leading to that deviation—is often referred to as a *dominating* or *optimal* point [27,28].

3.7. Historical notes and further reading

Large deviation theory emerged as a general theory during the 1960s and 1970s from the independent works of Donsker and Varadhan [2–5,22], and Freidlin and Wentzell [6]. Prior to that period, large deviation results were known, but there was no unified and general framework that dealt with them. Among these results, it is worth noting Cramér's Theorem [1], Chebyshev's inequality [13], Sanov's Theorem [29], which had been anticipated by Boltzmann [30] (see [7]), as well as extensions of Cramér's Theorem obtained by Lanford [11], Bahadur and Zabell [31], and by Plachky and Steinebach [32]. Sanov's Theorem was already encountered in the introductory examples of Section 2, and will be treated again in the next section. What statisticians call saddle-point approximations (see, e.g., [33–35]) are also large deviation results for the probability density of sample means; see Appendix C. For more information on the development of large deviation theory, see the historical notes found in [13,28] as well as in Sec. VII.7 of [8].

The Gärtner–Ellis Theorem is the product of a result proved by Gärtner [17], which was later generalized by Ellis [18]. The work of Ellis [18] explicitly refers to the construction of the large deviation principle currently adopted in large deviation theory (see Appendix B), which stems from the work of Varadhan [22].

As noted before, the statement of the Gärtner–Ellis Theorem given here is a simplification of that theorem. In essence, the result that we have stated and used is that of Gärtner [17]; it is less general but less technical than the result proved by Ellis [18], which can be applied to cases where $\lambda(k)$ exists and is differentiable over some limited interval (so not necessarily the whole line, as in Gärtner's result), provided that a technical condition, known as the *steepness condition*, is verified. For a statement of this condition, see Theorem 5.1 of [10] or Theorem 2.3.6 of [13]; for an illustration of it, see Examples 4.3 and 4.8 of the next section.

The statement of Varadhan's Theorem given here is also a simplification of the original and complete result proved by Varadhan [22]; see, e.g., Theorem 4.3.1 in [13] and Theorem 1.3.4 in [15]. An example of rate function for which the full conditions of Varadhan's Theorem are not satisfied is presented in Example 4.8 of the next section.

Introductions to the theory of large deviations similar to the one given in this section can be found in review papers by Oono [9], Amann and Atmanspacher [36], Ellis [7,10], Lewis and Russell [37], and Varadhan [38]. Readers who are willing to read mathematical textbooks are encouraged to consult those of Ellis [8], Deuschel and Stroock [39], Dembo and Zeitouni [13], and den Hollander [40] for a proper mathematical account of large deviation theory. The main simplifications introduced in this review concern the definition of the large deviation principle, and the fact that we do not state large deviation principles using the abstract language of topological spaces and measure theory. The precise and rigorous definition of the large deviation principle can be found in Appendix B.

For an accessible introduction to Legendre–Fenchel transforms and convex functions, see the monograph of van Tiel [21] and Chap. VI of [8]. The definitive reference on convex analysis is the book by Rockafellar [19].

4. Mathematical applications

This section is intended to complement the previous section. We review here a number of mathematical problems for which large deviation principles can be formulated. The applications were selected to give an idea of the generality of large deviation theory, to illustrate important points about the Gärtner–Ellis Theorem, and to introduce many ideas and results that will be revisited from a more physical point of view in the next sections. We also discuss here a classification of large deviation results related from top to bottom by the contraction principle.

4.1. Sums of IID random variables

We begin our review of mathematical applications by revisiting the now familiar sample mean

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{75}$$

involving n IID random variables $X_1, X_2, \ldots, X_n$. The next three examples consider different cases of sample distributions for the X_i's, and derive the corresponding large deviation principle for S_n using the Gärtner–Ellis Theorem or, equivalently in this case, Cramér's Theorem. We start with an example closely related to the introductory example of Section 2 which was concerned with spins.

Example 4.1 (*Binary Random Variables*). Let the n random variables in S_n be such that $P(X_i = -1) = P(X_i = 1) = \frac{1}{2}$. For this distribution,

$$\lambda(k) = \ln\left\langle e^{kX_i}\right\rangle = \ln\cosh k, \quad k \in \mathbb{R}. \tag{76}$$

This function is differentiable for all $k \in \mathbb{R}$, as expected, so the rate function $I(s)$ of S_n can be calculated as the Legendre transform of $\lambda(k)$. The result is

$$I(s) = \frac{1+s}{2}\ln(1+s) + \frac{1-s}{2}\ln(1-s), \quad s \in [-1, 1]. \tag{77}$$

The minimum and zero of $I(s)$ is $s = 0$.

Surprisingly, not all sample means of IID random variables fall within the framework of Cramér's Theorem. Here is an example for which $\lambda(k)$ does not exist, and for which large deviation theory yields in fact no useful information.

Example 4.2 (*Symmetric Lévy Random Variables*). The class of *strictly stable* or *strict Lévy* random variables that are symmetric is defined by the following characteristic function:

$$\left\langle e^{i\xi X}\right\rangle = e^{-\gamma|\xi|^\alpha}, \quad \xi \in \mathbb{R}, \gamma > 0, \alpha \in (0, 2). \tag{78}$$

From this result, it is tempting to make the change of variables $i\xi = k$, often called a Wick rotation, to write $\lambda(k) = -\gamma|k|^\alpha$ for $k \in \mathbb{R}$, but the correct result for k real is actually

$$\lambda(k) = \begin{cases} 0 & \text{if } k = 0 \\ \infty & \text{otherwise.} \end{cases} \tag{79}$$

This follows because the probability density $p(x)$ corresponding to the characteristic function of (78) has power-law tails of the form $p(x) \sim x^{-1-\alpha}$ as $|x| \to \infty$, which implies that $\langle e^{kX}\rangle$ does not converge for $k \in \mathbb{R} \setminus \{0\}$, although it converges when k is purely imaginary, that is, when $k = i\xi$ with $\xi \in \mathbb{R}$.

From the point of view of large deviation theory, the divergence of $\lambda(k)$ implies that a large deviation principle cannot be formulated for sums of symmetric Lévy random variables. This is expected since the probability density of such sums is known to have power-law tails that decay slower than an exponential in n [41]. If we attempt to calculate a rate function in this case, we trivially find $I = 0$, as in Example 3.4 (see also [11]).

In some cases, Cramér's Theorem can be applied where $\lambda(k)$ is differentiable to obtain information about the deviations of a random variables for a restricted range of its values. The basis of this *local* or *pointwise* application of Cramér's Theorem has to do with Legendre duality. In the case where $\lambda(k)$ is differentiable for all $k \in \mathbb{R}$, we have seen already that the Legendre–Fenchel transform

$$I(s) = \sup_k\{ks - \lambda(k)\} \tag{80}$$

reduces to the Legendre transform

$$I(s) = k(s)s - \lambda(k(s)), \tag{81}$$

where $k(s)$ is the unique solution of $\lambda'(k) = s$. By Legendre duality, the Legendre transform can also be written as

$$I(s(k)) = ks(k) - \lambda(k), \tag{82}$$

where $s(k) = \lambda'(k)$. Thus we see that if $\lambda(k)$ is differentiable at k, then the rate function I at the point $s(k)$ can be expressed through the Legendre transform shown above. By applying this local Legendre transform to all the points k where λ is differentiable, we are then able to recover part of $I(s)$ even if $\lambda(k)$ is not everywhere differentiable. This is illustrated next with a variant of the previous example.

Example 4.3 (*Totally Skewed Lévy Random Variables*). Not all strictly stable random variables have an infinite generating function for $k \neq 0$. A particular subclass of these random variables, known as *totally skewed to the left*, is such that

$$\lambda(k) = \ln\left\langle e^{kX}\right\rangle = \begin{cases} bk^\alpha & \text{if } k \geq 0 \\ \infty & \text{otherwise,} \end{cases} \tag{83}$$

where $b > 0$ and $\alpha \in (1, 2)$ [41,42]. The probability density associated with this log-generating function is shown Fig. 7. The situation that we face here is that $\lambda(k)$ is not defined for all $k \in \mathbb{R}$. This prevents us from using Cramér's Theorem,

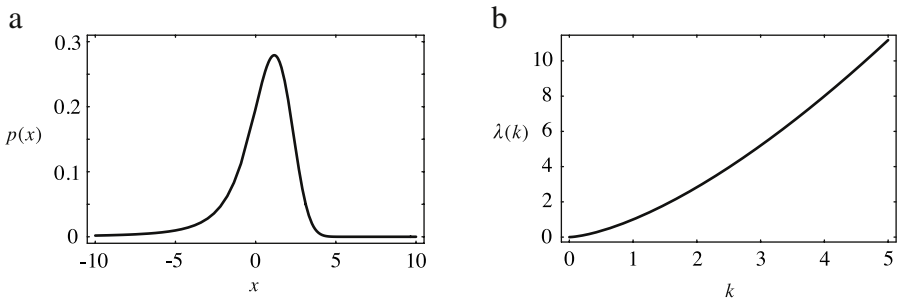

Fig. 7. (a) Probability density $p(x)$ of a Lévy random which is totally skewed to the left with $\alpha = 1.5$ and $b = 1$. The left tail of $p(x)$ decays as $|x|^{-2.5}$, while the right tail decays faster than an exponential; see [42] for more details. (b) Corresponding $\lambda(k)$ for $k \geq 0$; $\lambda(k) = \infty$ for $k < 0$.

and so from using the Legendre–Fenchel transform of Eq. (80) to obtain the full rate function $I(s)$. However, following the discussion above, we can apply Cramér's Theorem locally where $\lambda(k)$ is differentiable to obtain part of the rate function $I(s)$ through the Legendre transform of Eq. (82). Doing so leads us to obtain $I(s)$ for $s > 0$, since $\lambda'(k) > 0$ for $k > 0$ [42]. For $s \leq 0$, it can be proved that the probability density of S_n has a power-law decaying tail [43], so that $I(s) = 0$ for $s \leq 0$, as in the previous example. This part of $I(s)$ cannot be obtained from the Legendre transform of Eq. (82), but yields in any case no useful information about the precise decay of the probability density of S_n for $S_n \leq 0$.

This trick of locally applying the Legendre transform shown in Eq. (82) to the differentiable points of $\lambda(k)$ to obtain specific points of $I(s)$ works for any random variables not just sample means of IID random variables. Therefore, although the Gärtner–Ellis Theorem does not rigorously apply when $\lambda(k)$ is not everywhere differentiable, it is possible to obtain part of the rate function associated with $\lambda(k)$ simply by Legendre-transforming the differentiable points of $\lambda(k)$. In a sense, one can therefore say that *the Gärtner–Ellis Theorem holds locally where $\lambda(k)$ is differentiable*. The justification of this statement will come in Section 4.4 when we discuss nonconvex rate functions.

4.2. Sanov's Theorem

Large deviation principles can be formulated for many types of random variable, not just scalar random variables taking values in $\mathbb{R}$. One particularly important case of large deviation principles is that applying to random vectors taking values in $\mathbb{R}^d$, $d > 1$. To illustrate this case, let us revisit the problem of determining the probability distribution $P(L_n = l)$ associated with the empirical vector L_n introduced in Example 2.4. Recall that, given a sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ of n IID random variables taking values in a finite set Λ, the empirical vector $L_n(\omega)$ is the vector of empirical frequencies defined by the sample mean

$$L_{n,j}(\omega) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\omega_i,j}, \quad j \in \Lambda. \tag{84}$$

This vector has $|\Lambda|$ components, and the space of L_n, as noted earlier, is the set of probability distributions on Λ.

To find the large deviations of L_n, we consider the vector extension of the Gärtner–Ellis Theorem obtained by replacing the product kL_n in the definition of $\lambda(k)$ by the scalar product $k \cdot L_n$ involving the vector $k \in \mathbb{R}^{\Lambda}$. Thus the scaled cumulant generating function that we must now calculate is

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{nk\cdot L_n} \right\rangle, \quad k \in \mathbb{R}^{\Lambda}. \tag{85}$$

Since L_n is a sample mean of IID random variables, the expression of $\lambda(k)$ simplifies to

$$\lambda(k) = \ln \sum_{j\in\Lambda} \rho_j \mathrm{e}^{k_j}, \tag{86}$$

where $\rho_j = P(\omega_i = j), j \in \Lambda$. The expression above is necessarily analytic in k if Λ is finite. In this case, we can then use the Gärtner–Ellis Theorem to conclude that a large deviation principle holds for L_n with a rate function $I(l)$ given by

$$I(l) = \sup_k \{k \cdot l - \lambda(k)\} = k(l) \cdot l - \lambda(k(l)), \tag{87}$$

$k(l)$ being the unique root of $\nabla\lambda(k) = l$. Calculating the Legendre transform explicitly yields the rate function

$$I(l) = \sum_{j\in\Lambda} l_j \ln \frac{l_j}{\rho_j}, \tag{88}$$

which agrees with the rate function calculated by combinatorial means in Example 2.4.

The complete large deviation principle for L_n is known in large deviation theory as *Sanov's Theorem* [29]; see [7] for a discussion of Boltzmann's anticipation of this result. As already noted, $I(l)$ has a unique minimum and zero located at $l = \rho$. Moreover, as most of the rate functions encountered so far, $I(l)$ has the property that it is locally quadratic around it minimum:

$$I(l) \approx \frac{1}{2}\sum_{i,j\in\Lambda}(l_j-\rho_j)\left.\frac{\partial^2 I}{\partial l_j \partial l_i}\right|_{l=\rho}(l_i-\rho_i) = \frac{1}{2}\sum_{i\in\Lambda}\frac{(l_i-\rho_i)^2}{\rho_i}. \tag{89}$$

Extensions of Sanov's Theorem exist when Λ is infinite or even continuous. The mathematical tools needed to treat these cases are quite involved, but the essence of these extensions is easily explained at a heuristic level. For definiteness, consider the case where the IID random variables $\omega_1, \omega_2, \ldots, \omega_n$ take values in $\mathbb{R}$ according to a probability density $\rho(x)$. For this sequence, the continuous extension of the empirical vector L_n is the *empirical density*

$$L_n(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(\omega_i - x), \quad x \in \mathbb{R}, \tag{90}$$

involving Dirac's delta function δ. This is a normalized density in the sense that

$$\int_{-\infty}^{\infty} L_n(x)\,\mathrm{d}x = 1 \tag{91}$$

for all $\omega \in \mathbb{R}^n$. Since L_n is now a function (it is a random function to be more precise), the vector k used in the discrete version of Sanov's Theorem must be replaced by a function $k(x)$, so that

$$k \cdot L_n = \int_{-\infty}^{\infty} k(x)L_n(x)\,\mathrm{d}x. \tag{92}$$

Similarly, the analog of $\lambda(k)$ found in Eq. (86) is now a functional of $k(x)$ having the form

$$\lambda(k) = \ln\int_{-\infty}^{\infty}\rho(x)\,\mathrm{e}^{k(x)}\,\mathrm{d}x = \ln\left\langle \mathrm{e}^{k(X)}\right\rangle_\rho. \tag{93}$$

To apply the Gärtner–Ellis Theorem to this functional, we note that $\lambda(k)$ is differentiable in the sense of functional derivatives:

$$\frac{\delta\lambda(k)}{\delta k(y)} = \frac{\rho(y)\mathrm{e}^{k(y)}}{\left\langle \mathrm{e}^{k(X)}\right\rangle_\rho}. \tag{94}$$

By analogy with the discrete case, L_n must then satisfy a large deviation principle with a rate function $I(\mu)$ equal to the (functional) Legendre transform of $\lambda(k)$. The result of that transform, as should be expected, is the continuous version of the relative entropy:

$$I(\mu) = \int_{-\infty}^{\infty}\mathrm{d}x\,\mu(x)\ln\frac{\mu(x)}{\rho(x)}. \tag{95}$$

To complete this result, it must be added that $I(\mu) = \infty$ if μ has a larger support than ρ, that is mathematically, if μ is not continuous relative to ρ. This makes sense: the realizations of L_n cannot have a support larger than that of ρ.

4.3. Markov processes

Sample means of IID random variables constitute the simplest example of stochastic processes for which large deviation principles can be derived. The natural application to consider next concerns the class of Markov processes. Large deviation results have been formulated for this class of processes mainly by Donsker and Varadhan [2–5], who established through their work much of the basis of large deviation theory as we know it today. Our treatment of these processes will follow the path of the Gärtner–Ellis Theorem, and will be presented, for simplicity, for finite Markov chains. The case of continuous-time Markov processes will be discussed in Section 6 when dealing with nonequilibrium systems. Some subtleties of infinite-state Markov chains will also be discussed in Section 6.

The study of Markov chains is similar to the study of IID sample means, in that we consider a sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ of n random variables taking values in some finite set Λ, and study the sample mean

$$S_n = \frac{1}{n}\sum_{i=1}^{n} f(\omega_i) \tag{96}$$

involving an arbitrary function $f : \Lambda \to \mathbb{R}^d$, $d \geq 1$. The difference with the IID case, apart from the added function f, is that we now assume that the ω_i's form a *Markov chain* defined by

$$P(\omega) = P(\omega_1, \omega_2, \ldots, \omega_n) = \rho(\omega_1) \prod_{i=2}^{n} \pi(\omega_i|\omega_{i-1}). \tag{97}$$

In this expression, $\rho(\omega_1)$ denotes the probability distribution of the initial state ω_1, while $\pi(\omega_i|\omega_{i-1})$ is the conditional probability of ω_i given ω_{i-1}. We consider here the case where π is a fixed function of ω_i and ω_{i-1}, in which case the Markov chain is said to to be *homogeneous*. The sample mean S_n thus defined on the Markov chain ω is often referred to as a *Markov additive process* [13,28,44,45].

To derive a large deviation principle for S_n, we proceed as before to calculate $\lambda(k)$. The generating function of this random variable can be written as

$$\begin{aligned} \left\langle \mathrm{e}^{nk\cdot S_n} \right\rangle &= \sum_{\omega_1,\omega_2,\ldots,\omega_n} \rho(\omega_1)\mathrm{e}^{k\cdot f(\omega_1)}\pi(\omega_2|\omega_1)\mathrm{e}^{k\cdot f(\omega_2)}\cdots\pi(\omega_n|\omega_{n-1})\mathrm{e}^{k\cdot f(\omega_n)} \\ &= \sum_{\omega_1,\omega_2,\ldots,\omega_n} \pi_k(\omega_n|\omega_{n-1})\cdots\pi_k(\omega_2|\omega_1)\rho_k(\omega_1), \end{aligned} \tag{98}$$

by defining $\rho_k(\omega_1) = \rho(\omega_1)\mathrm{e}^{k\cdot f(\omega_1)}$ and $\pi_k(\omega_i|\omega_{i-1}) = \pi(\omega_i|\omega_{i-1})\mathrm{e}^{k\cdot f(\omega_i)}$. We recognize in the second equation a sequence of matrix products involving the vector of values $\rho_k(\omega_1)$ and the *transition matrix* $\pi_k(\omega_i|\omega_{i-1})$. To be more explicit, let us denote by ρ_k the vector of probabilities $\rho_k(\omega_1 = i)$, that is, $(\rho_k)_i = \rho_k(\omega_1 = i)$, and let Π_k denote the matrix formed by the elements of $\pi_k(\omega_i|\omega_{i-1})$, that is, $(\Pi_k)_{ji} = \pi_k(j|i)$. In terms of ρ_k and Π_k, we then write

$$\left\langle \mathrm{e}^{nk\cdot S_n} \right\rangle = \sum_{j\in\Lambda} \left(\Pi_k^{n-1}\rho_k\right)_j, \tag{99}$$

The function $\lambda(k)$ is extracted from this expression by determining the asymptotic behavior of the product $\Pi_k^{n-1}\rho_k$ using the Perron–Frobenius theory of positive matrices. Depending on the form of Π, one of three cases arises:

Case A: Π is ergodic (irreducible and aperiodic), and has therefore a unique stationary probability distribution ρ^* such that $\Pi\rho^* = \rho^*$. In this case, Π_k has a unique *principal* or *dominant* eigenvalue $\zeta(\Pi_k)$ from which it follows that $\left\langle \mathrm{e}^{nk\cdot S_n} \right\rangle \asymp \zeta(\Pi_k)^n$, and thus that $\lambda(k) = \ln\zeta(\Pi_k)$. Given that Π is assumed to be finite, $\zeta(\Pi_k)$ must be analytic in k. From the Gärtner–Ellis Theorem, we therefore conclude that S_n satisfies a large deviation principle with rate function

$$I(s) = \sup_k\{k\cdot s - \ln\zeta(\Pi_k)\}. \tag{100}$$

Case B: Π is not irreducible, which means that it has two or more stationary distributions (broken ergodicity). In this case, $\lambda(k)$ exists but depends generally on the initial distribution $\rho(\omega_1)$. Furthermore, $\lambda(k)$ may be nondifferentiable, in which case the Gärtner–Ellis Theorem does not apply. This arises, for example, when two of more eigenvalues of Π_k compete to be the dominant eigenvalue for different initial distribution $\rho(\omega_1)$ and different k.

Case C: Π has no stationary distributions (e.g., Π is periodic). In this case, no large deviation principle can generally be found for S_n. In fact, in this case, the Law of Large Numbers does not even hold in general.

The next two examples study Markov chains falling in Case A. The first example is a variation of Example 2.1 on random bits, whereas the second generalizes Sanov's Theorem to Markov chains. For examples of Markov chains falling in Case B, see [46,47].

Example 4.4 (*Balanced Markov Bits [9]*). Consider again the bit sequence $b = (b_1, b_2, \ldots, b_n)$ of Example 2.1, but now assume that the bits have the Markov dependence shown in Fig. 8(a) with $\alpha \in (0, 1)$. The symmetric and irreducible transition matrix associated with this Markov chain is

$$\Pi = \begin{pmatrix} \pi(0|0) & \pi(0|1) \\ \pi(1|0) & \pi(1|1) \end{pmatrix} = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix}. \tag{101}$$

The largest eigenvalue of

$$\Pi_k = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha\mathrm{e}^k & (1-\alpha)\mathrm{e}^k \end{pmatrix} \tag{102}$$

can be calculated explicitly to obtain $\lambda(k)$. The result is shown in Fig. 8 for various values of α. Also shown in this figure is the corresponding rate function $I(r)$ obtained by calculating the Legendre transform of $\lambda(k)$. The rate function clearly differs from the rate function found for independent bits. In fact, to second order in $\varepsilon = 1/2 - \alpha$ we have

$$I(r) \approx I_0(r) + 2(1-2r)^2\varepsilon + (2 - 32r^2 + 64r^3 - 32r^4)\varepsilon^2, \tag{103}$$

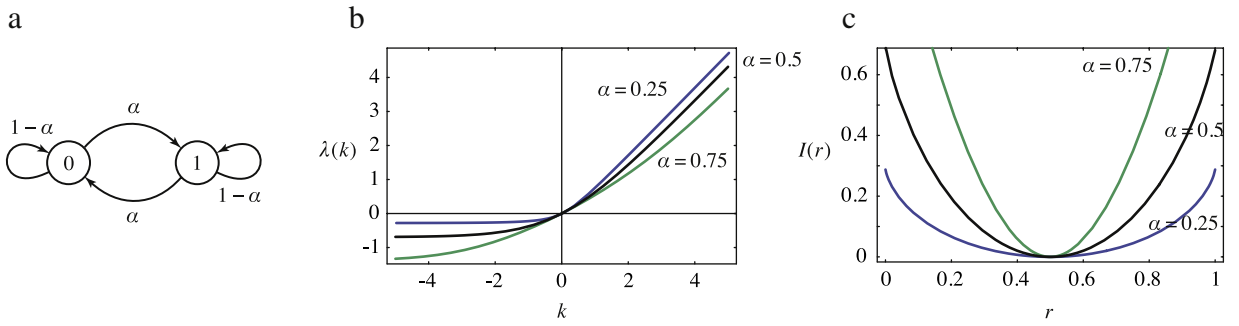

Fig. 8. (a) Transition probabilities between the two states of the symmetric binary Markov chain of Example 4.4. (b) Corresponding scaled cumulant generating function $\lambda(k)$ and (c) rate function $I(r)$ for $\alpha = 0.25, 0.5$, and 0.75.

where $I_0(r)$ is the rate function of the independent bits obtained here for $\alpha = 1/2$ or, equivalently, for $\varepsilon = 0$; see Eq. (3). Note that the zero of $I(s)$ does not change with α because the stationary distribution of the Markov chain is uniform for all $\alpha \in (0, 1)$, which means that the most probable sequences are the balanced sequences such that $R_n = 1/2$. What changes with α is the propensity of generating repeated strings of 0's or 1's in a given bit sequence. For $\alpha < 1/2$, a bit is more likely to be followed by the same bit, while for $\alpha > 1/2$, a bit is more likely to be followed by its opposite. The effect of this correlation, as can be seen from Fig. 8(c), is that empirical frequencies of 1's close to 0 or 1 are exponentially more probable for $\alpha < 1/2$ than for $\alpha > 1/2$.

Oono [9] discusses an interesting variant of the example above having absorbing states and a corresponding linear rate function. General quadratic approximations of rate functions of Markov chains are also discussed in that paper.

Example 4.5 (*Sanov's Theorem for Markov Chains*). The extension of Sanov's Theorem to irreducible Markov chains can be derived from the general Legendre–Fenchel transform shown in (100) by choosing $f(\omega_i) = \delta_{\omega_i,j}$, $j \in \Lambda$, in which case $\pi_k(j|i) = \pi(j|i)e^{kj}$. Ellis shows in [18] (see Theorem III.1) that the supremum over all vectors $k \in \mathbb{R}^\Lambda$ involved in that transform can be simplified to the following supremum:

$$I(l) = \sup_{u>0} \sum_{j\in\Lambda} l_j \ln \frac{u_j}{(\Pi u)_j} = \sup_{u>0} \left\langle \ln \frac{u}{\Pi u} \right\rangle_l , \tag{104}$$

which involves only the strictly positive vectors u in $\mathbb{R}^\Lambda$. This result was first obtained by Donsker and Varadhan [2]; for a proof of it, see Sec. 3.1.2 of [13] or Sec. V.B of [28]. The minimum and zero of this rate function is the stationary distribution ρ^* of Π.

The expression of the rate function for the empirical vector is obviously more complicated for Markov chains because of the correlations introduced between the random variables $\omega_1, \omega_2, \ldots, \omega_n$. Since these random variables "interact" only between pairs, it might be expected that a rate function similar in structure to the relative entropy is obtained if we replace the single-site empirical vector by a double-site empirical vector or *empirical matrix*, that is, if we look at the frequencies of occurrences of pair values in a Markov chain. Mathematically, this pair empirical matrix should be defined as

$$Q_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\omega_i,x} \delta_{\omega_{i+1},y}, \quad x, y \in \Lambda \tag{105}$$

by requiring that $\omega_{n+1} = \omega_1$ because $Q_n(x, y)$ has then the nice property that

$$\sum_{x\in\Lambda} Q_n(x, y) = L_n(y), \quad \text{and} \quad \sum_{y\in\Lambda} Q_n(x, y) = L_n(x), \tag{106}$$

where L_n is the usual empirical vector of the random sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$. In this case, Q_n is said to be *balanced* or to have *shift-invariant* marginals.

The rate function of Q_n can be derived in many different ways. One which is particularly elegant focuses on the sequence $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_n)$ which is built from the contiguous pairs $\zeta_i = (\omega_i, \omega_{i+1})$ appearing in the sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$. The empirical vector of ζ is the pair empirical distribution of ω, and the probability distribution of ζ factorizes in a way that partially mimics the IID case. Combining these two observations in Sanov's Theorem, it then follows that Q_n satisfies a large deviation principle with rate function

$$I_3(q) = \sum_{(x,y)\in\Lambda^2} q(x, y) \ln \frac{q(x, y)}{\pi(y|x)l(x)}, \tag{107}$$

where $l(x)$ is the marginal of $q(x, y)$. The complete derivation of this large deviation result can be found in Sec. 3.1.3 of [13]. Note that the zero of $I_3(q)$ is reached when $q(x, y)/l(x) = \pi(y|x)$, in which case $l(x) = \rho^*(x)$, where ρ^* is again the unique stationary distribution of Π.

4.4. Nonconvex rate functions

Since Legendre–Fenchel transforms yield functions that are necessarily convex (see Section 3.5.6), one obvious limitation of the Gärtner–Ellis Theorem is that it cannot be used to calculate *nonconvex* rate functions and, in particular, rate functions that have two or more local or global minima. The breakdown of this theorem for this class of rate functions is related to the differentiability condition on $\lambda(k)$. This is illustrated and explained next using a combination of examples and results about convex functions.

Example 4.6 (*Multi-atomic Distribution [10]*). A nonconvex rate function is easily constructed by considering a continuous random variable having a Dirac-like probability density supported on two or more points. The rate function associated with $p(Y_n = y) = \frac{1}{2}\delta(y \pm 1)$, for example, is

$$I(y) = \begin{cases} 0 & \text{if } y = \pm 1 \\ \infty & \text{otherwise,} \end{cases} \tag{108}$$

and is obviously nonconvex as it has two minima corresponding to its two non-singular values (a convex function always has only one minimum). Therefore, it cannot be expressed as the Legendre–Fenchel transform of the scaled cumulant generating function $\lambda(k)$ of Y_n. To be sure, calculate $\lambda(k)$:

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \frac{\mathrm{e}^{-nk} + \mathrm{e}^{nk}}{2} = |k| \tag{109}$$

and its Legendre–Fenchel transform:

$$I^{**}(y) = \sup_k \{ky - \lambda(k)\} = \begin{cases} 0 & \text{if } y \in [-1, 1] \\ \infty & \text{otherwise.} \end{cases} \tag{110}$$

The result does indeed differ from $I(y)$; in fact, $I(y) \neq I^{**}(y)$ for $y \in (-1, 1)$.

The Gärtner–Ellis Theorem is obviously not applicable here because $\lambda(k)$ is not differentiable at $k = 0$. However, as in the example of the skewed Lévy random variables (Example 4.3), we could apply the Legendre transform of Eq. (82) locally where $\lambda(k)$ is differentiable to obtain some part of $I(y)$. In this case, we obtain only two points of this function, namely, $I(-1) = 0$ and $I(1) = 0$, since $\lambda'(k) = -1$ for $k < 0$ and $\lambda'(k) = 1$ for $k > 0$.

The previous example raises a number of important questions related to the Gärtner–Ellis Theorem and the way rate functions are calculated. The most obvious has to do with the differentiability of $\lambda(k)$: Is there a general connection between the differentiability of this function and the convexity of rate functions? Indeed, why is $\lambda(k)$ required to be differentiable in the Gärtner–Ellis Theorem? Moreover, what is the result of the Legendre–Fenchel transform of $\lambda(k)$ in general? To answer these questions, we list and discuss next four results of convex analysis that characterize the Legendre–Fenchel transform. All of these results can be found in [19] (see also [21] and Chap. VI of [8]).

Result 1: The Legendre–Fenchel transform of I yields λ whether I is convex or not.

This result follows essentially because $\lambda(k)$ is always convex. In convex analysis, the Legendre–Fenchel transform of I is denoted by I^*. Thus $I^* = \lambda$ for all λ, in accordance with Varadhan's Theorem.

Result 2: If I is nonconvex, then the Legendre–Fenchel transform of λ, denoted by λ^*, does not yield I; rather, it yields the *convex envelope* of I.

This result is illustrated in Fig. 9. The convex envelope is usually denoted by I^{**}, since it is given by the double Legendre–Fenchel transform of I, and is such that $I^{**} \leq I$. With this notation, we then have $\lambda^* = I^{**} \neq I$ if I is nonconvex, and $I = \lambda^* = I^{**}$ if I is convex. Accordingly, when a rate function I is convex, it can be calculated as the Legendre–Fenchel transform of λ.

Result 3: The convex envelope I^{**} of I has the same Legendre–Fenchel transform as I, that is, $(I^{**})^* = I^* = \lambda$; see Fig. 9. In general, functions having the same convex envelope have the same Legendre–Fenchel transform.

This property explains why nonconvex rate functions cannot be obtained from λ. Put simply, the Legendre–Fenchel transform is a many-to-one transformation for the class of nonconvex functions. We also say that the Legendre–Fenchel transform is *non-self-dual* or *non-involutive* for nonconvex functions.

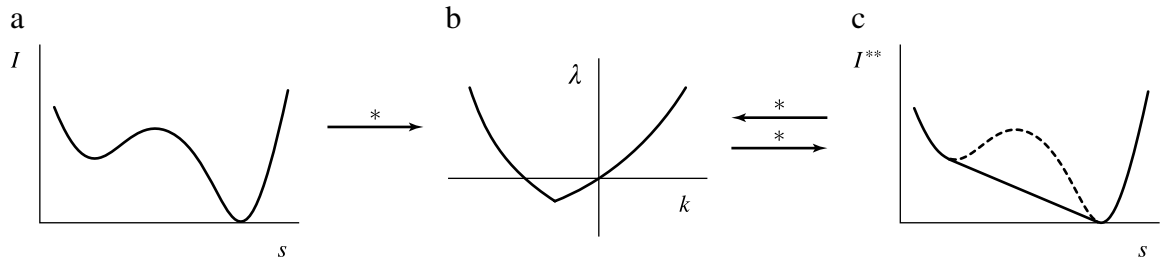

Fig. 9. Legendre–Fenchel transforms connecting (a) a nonconvex rate function $I(s)$, (b) its associated scaled cumulant generating function $\lambda(k)$, and (c) the convex envelope $I^{**}(s)$ of $I(s)$. The arrows illustrate the relations $I^* = \lambda$, $\lambda^* = I^{**}$ and $(I^{**})^* = \lambda$.

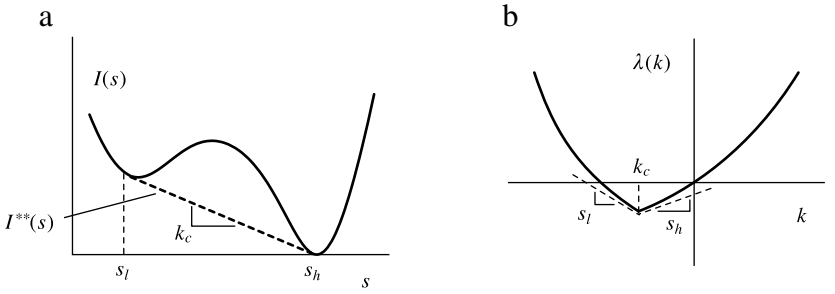

Fig. 10. (a) Nonconvex rate function I and its convex envelope I^{**}. (b) Associated scaled cumulant generating function $\lambda(k)$ having a nondifferentiable point at k_c.

Result 4: λ is nondifferentiable if I is nonconvex. To be more precise, suppose that $I(s)$ differs from its convex envelope $I^{**}(s)$ over some open interval (s_l, s_h), as in Fig. 10(a). Then its Legendre–Fenchel transform $I^* = \lambda$ is nondifferentiable at some value k_c corresponding to the slope of $I^{**}(s)$ over the interval (s_l, s_h); see Fig. 10(b). Moreover, the left- and right-derivatives of λ at k_c equal s_l and s_h, respectively. The same results hold when $I(s)$ is linear (we also say *affine*) over (s_l, s_h).

The condition of differentiability of $\lambda(k)$ entering in the Gärtner–Ellis Theorem can be understood from these results as follows. From the Results 2 and 3, we have that a rate function I can be obtained as the Legendre–Fenchel transform of λ only if I is convex. This leaves us with two possibilities: either I is strictly convex, that is, it is convex with no linear parts, or else I is convex but has one or more linear parts. The second possibility leads to a nondifferentiable $\lambda(k)$, as is the case for a nonconvex I according to the Results 3 and 4, so these two cases cannot be distinguished from the point of view of λ; see Fig. 9. Hence, the only possibility for which the sole knowledge of λ enables us to write $I = \lambda^*$ is when I is strictly convex. In this case, λ is differentiable by the Result 4, as required by the Gärtner–Ellis Theorem.

This reasoning shows, incidentally, that the differentiability of $\lambda(k)$ is a *sufficient but not a necessary* condition for having $I = \lambda^*$. Simply consider the case where I is convex but has one or more linear parts. Then $I = \lambda^*$, since I is convex, but $\lambda = I^*$ is nondifferentiable by the Result 4. The problem with rate functions having linear parts, as pointed out, is that they cannot be distinguished from nonconvex rate functions if we only know λ. That is, without any a priori knowledge of I, we know for sure that $I = \lambda^*$ only when $\lambda(k)$ is differentiable. The next example puts these observations into practice.

Example 4.7 (*Mixed Gaussian Sum*). This example is due to Ioffe [48]. Consider the sum

$$S_n = Y + \frac{1}{n}\sum_{i=1}^{n} X_i \tag{111}$$

where the X_i's are IID random variables distributed according to the normal distribution with unit mean and unit variance, and where Y is a discrete random variable, taken to be independent of the X_i's and such that $P(Y = -1) = P(Y = 1) = \frac{1}{2}$. To find the rate function of S_n, we use our knowledge of the Gaussian sample mean (see Examples 2.2 and 3.1) to write

$$P(S_n \in \mathrm{d}s | Y = \pm 1) \asymp \mathrm{e}^{-nI_{\pm}(s)}\,\mathrm{d}s, \tag{112}$$

where $I_{\pm}(s) = (s \mp 1)^2/2$. As a result,

$$P(S_n \in \mathrm{d}s) = \sum_{y=\pm 1} P(S_n \in \mathrm{d}s | Y = y)P(Y = y) \asymp \mathrm{e}^{-nI_-(s)}\,\mathrm{d}s + \mathrm{e}^{-nI_+(s)}\,\mathrm{d}s \asymp \mathrm{e}^{-nI(s)}\,\mathrm{d}s, \tag{113}$$

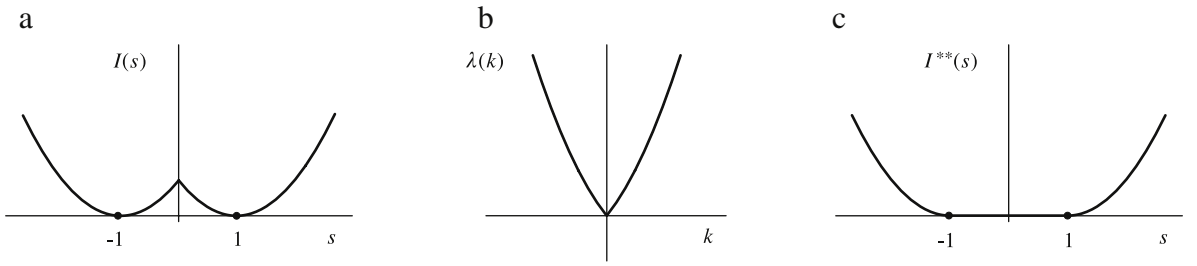

Fig. 11. (a) Nonconvex rate function $I(s)$ for Example 4.7. (b) Corresponding $\lambda(k)$. (c) Convex envelope $I^{**}(s)$ of $I(s)$.

where

$$I(s) = \min\{I_-(s), I_+(s)\} = \begin{cases} I_-(s) & \text{if } s < 0 \\ I_+(s) & \text{if } s \geq 0. \end{cases} \tag{114}$$

This rate function is nonconvex, as seen in Fig. 11(a).

We can verify that the Legendre–Fenchel transform of $\lambda(k)$ yields the convex envelope of $I(s)$ rather than $I(s)$ itself. Following the previous example, we find here

$$\lambda(k) = \frac{1}{2}k^2 + |k|, \quad k \in \mathbb{R}. \tag{115}$$

Notice that $\lambda(k)$ is nondifferentiable at $k = 0$; see Fig. 11(b). Taking the Legendre–Fenchel transform yields

$$I^{**}(s) = \sup_k\{ks - \lambda(k)\} = \begin{cases} I_-(s) & \text{if } s < -1 \\ 0 & \text{if } s \in [-1, 1] \\ I_+(s) & \text{if } s > 1. \end{cases} \tag{116}$$

Fig. 11(c) shows that $I^{**}(s)$ is the convex envelope of $I(s)$, which differs from $I(s)$ for $s \in (-1, 1)$. The part of $I(s)$ that can be obtained by applying the local Legendre transform of Eq. (82) to the differentiable branches of $\lambda(k)$ is the part of $I(s)$ that coincides with its convex envelope. We leave it to the reader in the end to show that $(I^{**})^* = I^* = \lambda$. The calculation of these Legendre–Fenchel transforms involves an interplay of local and global maximizers which accounts for the nondifferentiable point of $\lambda(k)$.

The last example of this section is there to show that boundary points of $\lambda(k)$ can also be thought of as nondifferentiable points for the purpose of the Gärtner–Ellis Theorem.

Example 4.8 (*Non-Steep* λ). Consider again the sum S_n shown in (111), but let $Y = Z/n$, where Z is an exponentially-distributed random variable with unit mean, that is, $p(Z = z) = \mathrm{e}^{-z}$, $z \geq 0$. The calculation of $\lambda(k)$ for S_n yields

$$\lambda(k) = \frac{k^2}{2} + \lim_{n\to\infty} \frac{1}{n} \ln\left\langle \mathrm{e}^{kZ}\right\rangle = \begin{cases} k^2/2 & \text{if } k < 1 \\ \infty & \text{if } k \geq 1. \end{cases} \tag{117}$$

Applying the Legendre transform of Eq. (82) to the differentiable branch of this function leads to $I(s) = s^2/2$ for $s < 1$, since $\lambda'(k) < 1$ for $k \in (-\infty, 1)$; see Fig. 12(a). As in the previous examples, the Legendre transform of $\lambda(k)$ yields here only part of $I(s)$ because the image of $\lambda'(k)$ does not cover the whole range of S_n. In the present example, we say that $\lambda(k)$ is *non-steep* because its derivative is upper bounded.

To obtain the full rate function of S_n, we can follow the previous example by noting that, conditionally on $Z = z$, S_n must be Gaussian with mean z/n and unit variance. Therefore,

$$p(S_n = s) = \int_0^\infty p(S_n = s|Z = z)\, p(Z = z)\, \mathrm{d}z \asymp \int_0^\infty \mathrm{e}^{-n(s-z/n)^2/2}\, \mathrm{e}^{-z}\, \mathrm{d}z. \tag{118}$$

A large deviation principle is extracted from the last integral by performing the integral exactly or by using Laplace's approximation. In both cases, we obtain $p(S_n = s) \asymp \mathrm{e}^{-nI(s)}$, where

$$I(s) = \begin{cases} s - 1/2 & \text{if } s > 1 \\ s^2/2 & \text{if } s \leq 1. \end{cases} \tag{119}$$

As seen in Fig. 12(b), $I(s)$ is not strictly convex, since it is linear for $s > 1$; hence the fact that this part cannot be obtained from the Gärtner–Ellis Theorem. The nondifferentiable point of $\lambda(k)$ associated with this part of $I(s)$ is non-trivial: it is the boundary point of $\lambda(k)$ located at $k = 1$. That nondifferentiable point would also arise if $I(s)$ were nonconvex for $s > 1$ instead of being linear; see Fig. 12(c).

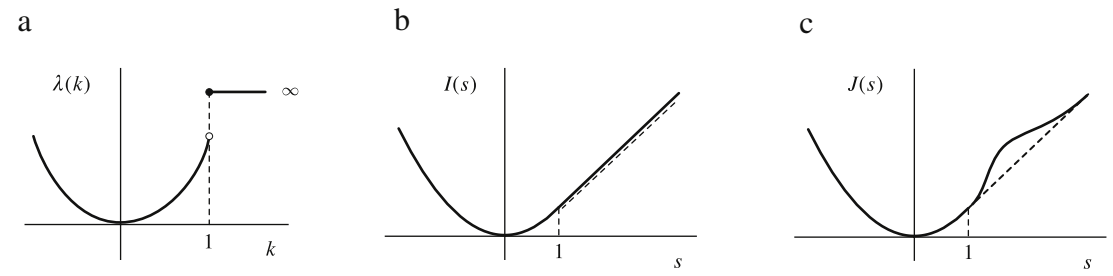

Fig. 12. (a) $\lambda(k)$ for Example 4.8. (b) Rate function $I(s)$ for that example. (c) Nonconvex rate function $J(s)$ having the same Legendre–Fenchel transform as $I(s)$.

There is an extra mathematical subtlety related to the boundary point of $\lambda(k)$, namely, that the Legendre–Fenchel transform of $I(s)$ does not yield $\lambda(k)$ at $k = 1$. This may seem to contradict Varadhan's Theorem, but there is in fact no contradiction here. Simply, there is a technical condition associated with this theorem which we have not mentioned (see, e.g., Theorem 5.1 of [10] or Theorem 4.3.1 of [13]), and which happens to be violated in the present example. Mathematically, the problem arises because Legendre–Fenchel transforms yield functions that are *lower semi-continuous* in addition to being convex [19,21]. Here $\lambda(k)$ is convex but not lower semi-continuous, since its domain is open; hence the fact that $\lambda \neq I$ at $k = 1$. In practice, boundary points of $\lambda(k)$ appear to be the only points for which we may have $\lambda \neq I^*$, and thus for which Varadhan's Theorem must be applied with care.

Other examples of nonconvex rate functions related to non-irreducible Markov chains having more that one stationary distributions are discussed by Dinwoodie [46,47]. These examples relate to the Case B of Markov chains mentioned before. In the end, it should be kept in mind that nonconvex rate functions pose no limitations for large deviation theory; they pose only a limitation for the Gärtner–Ellis Theorem because of the way that theorem relies on Legendre–Fenchel transforms. We will see in the next section other methods that can be used to calculate rate functions, be they convex or not.

4.5. Self-processes

There are many random variables, apart from sample means, that can be studied from the point of view of large deviation theory. One which is often studied in information theory and in nonequilibrium statistical mechanics is

$$A_n(\omega) = -\frac{1}{n}\ln P_n(\omega), \tag{120}$$

where $P_n(\omega)$ denotes as usual the probability distribution of the sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$. We call A_n a *self-process*, since it is a transformed version of $P_n(\omega)$. Therefore, studying the large deviations of A_n with respect to P_n is, in a way, the same as studying the large deviations of P_n with respect to itself.

The rate function of A_n can be calculated using the Gärtner–Ellis Theorem once the nature of ω is specified. The random variables $\omega_1, \omega_2, \ldots, \omega_n$ can form, for example, a Markov chain or can be IID, in which case A_n reduces to a simple IID sample mean. Note in this case the form of $\lambda(k)$:

$$\lambda(k) = \lim_{n\to\infty}\frac{1}{n}\ln\left\langle \mathrm{e}^{-k\ln P_n}\right\rangle = \lim_{n\to\infty}\frac{1}{n}\ln\sum_{\omega\in\Lambda^n}P_n(\omega)^{1-k} = \ln\sum_{j\in\Lambda}P(\omega_i = j)^{1-k}. \tag{121}$$

Other types of sequences of random variables or stochastic processes can be dealt with by calculating $\lambda(k)$ from its definition; see in particular [49,50] for the treatment of continuous-time Markov processes.

An important point to note here is that, if the rate function $I(a)$ of A_n has a unique global minimum and zero, as is the case when $I(a)$ is convex, then

$$\lim_{n\to\infty}\langle A_n\rangle = a^* \tag{122}$$

and

$$\lim_{n\to\infty} A_n = a^* \tag{123}$$

with probability 1. The first limit involving $\langle A_n\rangle$ implies, on the one hand, that the *mean Boltzmann–Gibbs–Shannon entropy*

$$H_n = -\frac{1}{n}\sum_{\omega\in\Lambda^n}P_n(\omega)\ln P_n(\omega) \tag{124}$$

converges to the constant a^*, which is called the *entropy rate* [14] or *Kolmogorov–Sinai entropy* [51,52]. On the other hand, the limit (123) involving A_n alone implies

$$-\frac{1}{n}\ln P_n(\omega_1,\omega_2,\ldots,\omega_n)\to a^* \tag{125}$$

with probability 1 as $n\to\infty$. The latter limit is known as the *Asymptotic Equipartition Theorem* or the *Shannon–McMillan–Breiman Theorem* [14]. What this result says concretely is that most of the probability $P_n(\omega)$ is concentrated on sequences in Λ^n such that $P_n(\omega)\asymp \mathrm{e}^{-na^*}$. The set $\mathcal{T}_n$ containing these sequences is commonly called the *typical set* of Λ^n [14]. Thus $P(\mathcal{T}_n)\to 1$ and $P_n(\omega)\asymp \mathrm{e}^{-na^*}$ for all $\omega\in\mathcal{T}_n$ in the limit $n\to\infty$, which implies that $\mathcal{T}_n$ must contain about e^{na^*} typical sequences. These results are fundamental in information theory; for a more detailed discussion of this point, see Chap. X of [28], Sec. 3.6 of [13] or Chap. 12 of [14]. For an application of the self-process in the context of nonequilibrium systems, see Example 6.10.

Example 4.9 (*Entropy Rate of a Markov Source*). For an ergodic Markov chain with transition matrix $\pi(j|i)$, the entropy rate is

$$a^*=\lambda'(0)=-\sum_{i,j\in\Lambda}\pi(j|i)\rho^*(i)\ln\pi(j|i), \tag{126}$$

ρ^* being as usual the unique stationary distribution of the Markov chain.

4.6. Level 1, 2 and 3 of large deviations

It is customary since the work of Donsker and Varadhan to define three levels of large deviation results referred to as the Level-1, 2 and 3 of large deviations [8]. *Level-1* is the level of sample means, whereas *Level-2* is the level of Sanov's Theorem, that is, the level of the large deviations of the empirical vector L_n. The reason for ordering the large deviations of the empirical vector above those of sample means is that the latter can be derived from the former using the contraction principle. To see this, let $\omega=(\omega_1,\omega_2,\ldots,\omega_n)\in\Lambda^n$ be a sequence of random variables, which are not necessarily independent, and let

$$S_n=\frac{1}{n}\sum_{i=1}^{n}f(\omega_i). \tag{127}$$

In terms of the empirical vector L_n of ω, S_n can always be written as

$$S_n=\int_\Lambda f(x)L_n(x)\,\mathrm{d}x=f\cdot L_n. \tag{128}$$

Thus, given the rate function $I_2(\mu)$ of L_n, we can use the contraction formula (74) with $h(\mu)=f\cdot\mu$ to express the rate function $I_1(s)$ of S_n as

$$I_1(s)=\inf_{\mu:h(\mu)=s}I_2(\mu). \tag{129}$$

The contraction function $h(\mu)$ is often written as $\mu(f)$ or $\langle f\rangle_\mu$ to emphasize that it is an average of f taken with respect to the random density μ. The empirical vectors solving the constrained minimization problem of Eq. (129) have an interesting probabilistic interpretation: they are, in the limit $n\to\infty$, the most probable vectors L_n such that $h(L_n)=s$. Consequently, these vectors must maximize the conditional probability

$$P(L_n\in d\mu|h(L_n)\in \mathrm{d}s)=\frac{P(L_n\in d\mu,h(L_n)\in \mathrm{d}s)}{P(h(L_n)\in \mathrm{d}s)}, \tag{130}$$

which implies that they must also globally minimize the rate function

$$I_2^s(\mu)=\begin{cases}I_2(\mu)-I_1(s) & \text{if } h(\mu)=s\\ \infty & \text{otherwise.}\end{cases} \tag{131}$$

Consequently, $I_1(s)=\inf_\mu I_2^s(\mu)$.

Example 4.10 (*Sample Means via Sanov's Theorem*). The constrained minimization arising from the contraction of Level-2 to Level-1 can be solved explicitly for IID sample means. For this case, $I_2(\mu)$ is the relative entropy and the contraction formula (129) is referred to as the *minimum relative entropy principle* [9]. To solve the constrained minimization, we use Lagrange's multipliers method and search for the unconstrained critical points of

$$F_\alpha(\mu)=\alpha h(\mu)-I_2(\mu),\quad \alpha\in\mathbb{R}. \tag{132}$$

Since $I_2(\mu)$ is strictly convex and $h(\mu) = \langle f \rangle_\mu$ is a linear and differentiable functional of μ, $F_\alpha(\mu)$ has a unique maximum μ_α for all $\alpha \in \mathbb{R}$ satisfying $\delta F_\alpha(\mu_\alpha) = 0$. Given the expression of the relative entropy, the expression of μ_α is found to be

$$\mu_\alpha(x) = \frac{\rho(x)\mathrm{e}^{\alpha f(x)}}{W(\alpha)}, \qquad W(\alpha) = \int_\Lambda \rho(x)\mathrm{e}^{\alpha f(x)}\,\mathrm{d}x = \left\langle \mathrm{e}^{\alpha f(X)} \right\rangle_\rho \tag{133}$$

with the value of α implicitly determined by the constraint $h(\mu_\alpha) = \langle f \rangle_{\mu_\alpha} = s$ or, equivalently, $W'(\alpha)/W(\alpha) = s$. The expression for μ_α makes sense, obviously, provided that $W(\alpha) < \infty$. If this is the case, then $I_1(s) = I_2(\mu_\alpha)$. Equivalently,

$$I_1(s) = \alpha s - \ln W(\alpha), \tag{134}$$

since

$$F_\alpha(\mu_\alpha) = \alpha h(\mu_\alpha) - I_2(\mu_\alpha) = \alpha s - I_2(\mu_\alpha) = \ln W(\alpha). \tag{135}$$

We recognize in Eq. (134) the result of Cramér's Theorem.

Example 4.11 (*Symmetric Lévy Random Variable Revisited*). The contraction of Level-2 to Level-1 does not work for sample means of symmetric Lévy random variables because $W(\alpha) = \infty$ for all $\alpha \neq 0$. This case is discussed by Lanford [11], who proves the result reached in Example 4.2, namely, $I(s) = 0$.

The *Level-3* of large deviations, from which Level-2 is obtained by contraction, is the level of the pair empirical distribution $Q_n(x, y)$, which is commonly completed by including all the m-tuple empirical distributions defined on $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$, $2 \leq m \leq n$. In defining m-tuple empirical distributions, we require, as we did for the pair empirical distribution, that $\omega_{n+1} = \omega_1$, $\omega_{n+2} = \omega_2$, and so forth until $\omega_{n+m} = \omega_m$, so as to guarantee that all the $(m-1)$-tuple distributions obtained by contraction of an m-tuple distribution are the same. The n-tuple distribution of an n-tuple sequence is the ultimate empirical distribution that can be defined. In the limit $n \to \infty$, such a distribution becomes an infinite joint empirical distribution called the *empirical process* [5]. The construction of this abstract process is explained in Sec. 6.5.3 of [13] or Chap. IX of [8]. We will limit ourselves here to noting that, for sequences of IID random variables, the empirical process possesses a convex rate function, and that the zero of this rate function is the infinite product of ρ, the common probability distribution of the IID random variables.

We close this section by noting an alternative characterization of the Level-2 rate function of Markov chains, derived by contracting the large deviations of the pair empirical matrix.

Example 4.12 (*Sanov's Theorem for Markov Chains Revisited*). Let L_n and Q_n denote, respectively, the empirical vector and empirical matrix of an irreducible Markov chain. The contraction $h(Q_n) = L_n$ that takes Q_n to L_n is the usual "tracing-out" operation expressed in Eq. (106). Given the rate function $I_3(q)$ of Q_n found in Eq. (107), we then have

$$I_2(\mu) = \inf_{q:h(q)=\mu} I_3(q) \tag{136}$$

for the rate function of L_n.

5. Large deviations in equilibrium statistical mechanics

The previous sections introduced all the large deviation results that will now be applied to study the properties of physical systems composed of many particles. We start in this section with the equilibrium properties of many-particle systems described at a probabilistic level by statistical–mechanical ensembles, such as the microcanonical or canonical ensembles. The use of large deviation techniques for studying these systems has its roots in the work of Ruelle [53], Lanford [11], and especially Ellis [7,8,10]. Of these sources, Ellis [8] is the first that explicitly referred to the mathematical theory of large deviations, as developed by Donsker and Varadhan [2–5], among others. The many links that exist between large deviations and equilibrium statistical mechanics have also been discussed by Lewis, Pfister, and Sullivan [54–59], as well as by Oono [9]. A basic overview of some of these links can be found in [36].

The material presented in this section borrows from all these sources. By defining statistical–mechanical ensembles in a way that is explicitly focused on large deviation theory, we aim to show here that the study of equilibrium states and their fluctuations in a given ensemble can be reduced to the study of properly-defined rate functions. In the process, many connections between equilibrium statistical mechanics and large deviation theory will be established and discussed. We will see, in particular, that entropy functions are special rate functions, and that variational principles, such as the maximum entropy principle or the minimum free energy principle, follow from the contraction principle. The last observation is especially useful because it provides us with a clear explanation of why variational principles arise in equilibrium statistical mechanics. It also provides us with a systematic method or *scheme* for deriving such variational principles in general.

5.1. Basic principles

The following list of common definitions and postulates, inspired from [8,11,53], establishes the basis of equilibrium statistical mechanics on which we will work:

- We consider a collection of n particles (atoms, spins, molecules, etc.) that interact with one another through some forces or potentials.[5]
- The *collective* or *joint* state of the n particles is denoted by a sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ of n variables, with ω_i denoting the state of the ith particle.
- A sequence ω is called a *microstate*, as it gives a complete description of the n-particle system at the microscopic level. The set or space Λ_n of all microstates is the n-fold product Λ^n of the one-particle state space Λ.
- The physical interactions or dependencies between the n particles are determined by a *Hamiltonian* or *energy function* $H_n(\omega)$. Given $H_n(\omega)$, we define the *mean energy* or *energy per particle* by $h_n(\omega) = H_n(\omega)/n$.
- The microstate ω of the n-particle system is modeled abstractly as a random variable, which is distributed according to a reference or *prior* probability measure $P(\mathrm{d}\omega)$ on Λ_n. The form of $P(\mathrm{d}\omega)$ is determined by physical considerations. For most models, Liouville's Theorem dictates that $P(\mathrm{d}\omega)$ be the uniform measure $P(\mathrm{d}\omega) = \mathrm{d}\omega/|\Lambda_n|$, where $|\Lambda_n| = |\Lambda|^n$ is the *volume* of Λ_n. Since $|\Lambda_n|$ is a constant, one can work equivalently with the unnormalized (Lebesgue) measure $P(\mathrm{d}\omega) = \mathrm{d}\omega$.
- The probabilistic description of the n-particle system is completed by specifying the external conditions or constraints under which that system is prepared or studied. The specification of these conditions is tantamount to selecting a given *statistical–mechanical ensemble*, which corresponds mathematically to a probability distribution on Λ_n involving the constraints and the prior distribution $P(\mathrm{d}\omega)$.
- The interactions between the particles give rise to a *macroscopic* or *thermodynamic* behavior of the whole system that can be described by having recourse to a few macroscopic or "coarse-grained" variables called *macrostates*. Mathematically, a macrostate is just a function $M_n(\omega)$ of the microstates.
- The thermodynamic behavior of the whole system is characterized by one or more *equilibrium states*, defined as the most probable values of a set of macrostates in a chosen ensemble.
- When calculating equilibrium states, the limit $n \to \infty$ is assumed to obtain states that are representative of macroscopic systems. This limit is called the *thermodynamic limit*, and entails in many cases the continuum limit.

The mathematical basis for the notion of thermodynamic behavior is the Law of Large Numbers [11]. The idea, in a nutshell, is that the outcomes of a macrostate, say $M_n(\omega)$, involving n particles should concentrate in probability around certain stable or equilibrium values (macroscopic determinism) despite the fact that the particles' state is modeled by a random variable ω (microscopic chaos). Large deviation theory enters this picture by noting that, in many cases, the outcomes of M_n are ruled by a large deviation principle, and that, in these cases, the concentration of M_n around equilibrium values is "exponentially effective" in the limit $n \to \infty$, as the probability of observing a departure from these equilibrium values is exponentially small with the number n of particles. Consequently, all that is needed to describe the state of a large many-particle system at the macroscopic level is to know the equilibrium values of M_n which correspond to the global minima of the rate function governing the fluctuations of M_n.

These considerations summarize the application of large deviation techniques in equilibrium statistical mechanics. What remains to be done at this point is to show how the probabilities of microstates and macrostates are to be constructed depending on the nature of the many-particle system studied, and to show how rate functions are extracted from these probabilities. For simplicity, we will review here only two types of many-particle systems, namely, closed systems at constant energy, and open systems exchanging energy with a heat bath at constant temperature. The first type of system is modeled, as is well known, by the *microcanonical ensemble*, whereas the second type is modeled by the *canonical ensemble*. The treatment of other ensembles follows the treatment of these two ensembles.

5.2. Large deviations of the mean energy

Before exploring the large deviations of general macrostates, it is useful to study the large deviations of the mean energy $h_n(\omega)$ with respect to the prior distribution $P(\mathrm{d}\omega)$ defined on Λ_n. The rate function turns out in this case to be the microcanonical entropy function up to an additive constant, whereas the scaled cumulant generating function of h_n turns out to be the canonical free energy function, again up to a constant. These associations and their consequences for thermodynamics are explained next.

5.2.1. Entropy as a rate function

Using the notation developed in the previous sections, we write the probability distribution of h_n with respect to the prior $P(\mathrm{d}\omega)$ on Λ_n as

$$P(h_n \in \mathrm{d}u) = \int_{\{\omega\in\Lambda_n : h_n(\omega)\in \mathrm{d}u\}} P(\mathrm{d}\omega), \tag{137}$$

[5] In keeping with the notations of the previous sections, we use n to denote the number of particles rather that the more common N used in physics.

where $\mathrm{d}u = [u, u+\mathrm{d}u]$ is an infinitesimal interval of mean energy values. For the uniform prior measure $P(\mathrm{d}\omega) = \mathrm{d}\omega/|\Lambda|^n$, $P(h_n \in \mathrm{d}u)$ is proportional to the volume

$$\Omega(h_n \in \mathrm{d}u) = \int_{\{\omega\in\Lambda_n : h_n(\omega)\in \mathrm{d}u\}} \mathrm{d}\omega \tag{138}$$

of microstates ω such that $h_n(\omega) \in \mathrm{d}u$. Therefore, if $P(h_n \in \mathrm{d}u)$ scales exponentially with n, then so must $\Omega(h_n \in \mathrm{d}u)$. Defining the rate function $I(u)$ of $P(h_n \in \mathrm{d}u)$ by the usual limit

$$I(u) = \lim_{n\to\infty} -\frac{1}{n} \ln P(h_n \in \mathrm{d}u), \tag{139}$$

we must then have

$$I(u) = \ln|\Lambda| - s(u), \tag{140}$$

where

$$s(u) = \lim_{n\to\infty} \frac{1}{n} \ln \Omega(h_n \in \mathrm{d}u) \tag{141}$$

is the *microcanonical entropy* or *entropy density*. Eq. (140) proves our first claim, namely, that the rate function $I(u)$, if it exists, is the negative of the entropy $s(u)$ up to the additive constant $\ln|\Lambda|$.

In the following, we shall absorb the constant $\ln|\Lambda|$ by re-defining the entropy using the limit

$$s(u) = \lim_{n\to\infty} \frac{1}{n} \ln P(h_n \in \mathrm{d}u) \tag{142}$$

rather than the limit displayed in Eq. (141). This re-definition simply amounts to replacing the Lebesgue measure $\mathrm{d}\omega$ in the integral of $\Omega(h_n \in u)$ by the uniform prior measure $P(\mathrm{d}\omega)$, in which case $I(u) = -s(u)$. This minor re-definition of the entropy complies with the definition used in works on large deviations and statistical mechanics (see, e.g., [8–10,59,60]). It brings, for one thing, the notion of entropy closer to large deviation theory, and allows one to use prior measures that are not uniform. Note that throughout this review, we also use $k_B = 1$.

5.2.2. Free energy as a scaled cumulant generating function

The proportionality of $P(h_n \in \mathrm{d}u)$ and $\Omega(h_n \in \mathrm{d}u)$ noted above implies that

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{nkh_n} \right\rangle \tag{143}$$

satisfies

$$\lambda(k) = -\varphi(\beta)|_{\beta=-k} - \ln|\Lambda|. \tag{144}$$

where

$$\varphi(\beta) = \lim_{n\to\infty} -\frac{1}{n} \ln Z_n(\beta), \tag{145}$$

and

$$Z_n(\beta) = \int_{\Lambda_n} \mathrm{e}^{-n\beta h_n(\omega)}\, \mathrm{d}\omega = \int_{\Lambda_n} \mathrm{e}^{-\beta H_n(\omega)}\, \mathrm{d}\omega. \tag{146}$$

The latter function is the well-known n-particle *partition function* associated with H_n, which means, therefore, that $\varphi(\beta)$ is the *canonical free energy function*. With these associations, we see, as announced, that $\lambda(k)$ is the free energy function of the canonical ensemble up to a constant and a change of variable ($\beta = -k$). As we did for the entropy, we shall absorb the constant $\ln|\Lambda|$ in $\varphi(\beta)$ by re-defining this function as

$$\varphi(\beta) = \lim_{n\to\infty} -\frac{1}{n} \ln \int_{\Lambda_n} \mathrm{e}^{-n\beta h_n(\omega)}\, P(\mathrm{d}\omega) = \lim_{n\to\infty} -\frac{1}{n} \ln \left\langle \mathrm{e}^{-n\beta h_n(\omega)} \right\rangle \tag{147}$$

using $P(\mathrm{d}\omega)$ instead of $\mathrm{d}\omega$ as the measure entering in the integral of the partition function. We this new definition, we then have $\lambda(k) = -\varphi(\beta)|_{\beta=-k}$. This form of free energy function will be used from now on, since it also complies with the form used in works on large deviations and statistical mechanics.

It should be noted for correctness that what is commonly referred to as the free energy in thermodynamics is not the function $\varphi(\beta)$ but the function $f(\beta) = \varphi(\beta)/\beta$. Here we use $\varphi(\beta)$ as the free energy because this function has the convenient property of always being concave in β. The function $f(\beta)$, by contrast, is concave or convex (negative concave) depending on the sign of β. In textbooks of statistical mechanics, $\varphi(\beta)$ is sometimes called the *Massieu potential* [61].

5.2.3. Legendre transforms in thermodynamics

The relationships that we have established between $I(u)$ and $s(u)$, on the one hand, and $\lambda(k)$ and $\varphi(\beta)$, on the other, are important because they imply that the Gärtner–Ellis Theorem and Varadhan's Theorem can be applied to express $s(u)$ as the Legendre–Fenchel transform of $\varphi(\beta)$, and vice versa. By transposing Varadhan's Theorem at the level of $s(u)$ and $\varphi(\beta)$, we indeed obtain

$$\varphi(\beta) = \inf_u\{\beta u - s(u)\}, \tag{148}$$

whereas for the Gärtner–Ellis Theorem, we obtain

$$s(u) = \inf_\beta\{\beta u - \varphi(\beta)\}, \tag{149}$$

provided that $\varphi(\beta)$ exists and is differentiable. In both expressions, "inf" stands as before for the "infimum of". The reason why the Legendre–Fenchel transform is now expressed with an "inf" instead of a "sup" is, of course, because the entropy is defined as the negative of a rate function; see Eq. (140).

The two Legendre–Fenchel transforms shown above are fundamental in statistical mechanics. The first one shown in Eq. (148) provides a precise formulation of the basic thermodynamic principle that states that the free energy is the Legendre transform of the entropy. Varadhan's Theorem refines this result by establishing that $\varphi(\beta)$ is, in general, the *Legendre–Fenchel* transform of $s(u)$, not simply the *Legendre* transform, and that this Legendre–Fenchel transform is valid for essentially any $s(u)$. Legendre–Fenchel transforms rather than Legendre transforms must be used in particular when $s(u)$ is not concave.

The second Legendre–Fenchel transform shown in Eq. (149) is the converse of the first, expressing the entropy as the Legendre–Fenchel transform of the free energy. This result is also well known in thermodynamics, but in a form that usually also involves the Legendre transform rather than the Legendre–Fenchel transform, and without reference to any conditions about the validity of that transform. These conditions are the conditions of the Gärtner–Ellis Theorem; they are important, and will be studied in detail later when discussing nonconcave entropies.

For convenience, we will often refer thereafter to the two Legendre–Fenchel transforms shown above using the "star" notation introduced in Section 4.4. With this notation, Eq. (148) is expressed as $\varphi = s^*$, while Eq. (149) is expressed as $s = \varphi^*$.

5.3. Microcanonical ensemble

We now come to the problem that we set ourselves to solve in this section: we consider an n-particle system represented by a Hamiltonian function $H_n(\omega)$, and attempt to derive a large deviation principle for a macrostate $M_n(\omega)$ of that system. We consider first the case of a closed system constrained to have a fixed energy $H_n(\omega) = U$. Other constraints can also be included (see, e.g., [62]). The statistical–mechanical ensemble that models the stationary properties of such a system is, as is well known, the microcanonical ensemble, which we define mathematically next.

5.3.1. Definition of the ensemble

The microcanonical ensemble is based on the assumption that all the microstates $\omega \in \Lambda_n$ such that $H_n(\omega) = U$ or, equivalently, such that $h_n(\omega) = U/n = u$ are equally probable (equiprobability postulate). Therefore, what we call the microcanonical ensemble at the level of microstates is the conditional probability measure

$$P^u(\mathrm{d}\omega) = P(\mathrm{d}\omega|h_n \in \mathrm{d}u) = \begin{cases} \dfrac{P(\mathrm{d}\omega)}{P(h_n \in \mathrm{d}u)} & \text{if } h_n(\omega) \in \mathrm{d}u \\ 0 & \text{otherwise,} \end{cases} \tag{150}$$

which assigns a non-zero and constant probability only to those microstates having a mean energy lying in the interval $\mathrm{d}u$. The probability $P(h_n \in \mathrm{d}u)$ was introduced earlier, and is there to make $P^u(\mathrm{d}\omega)$ a normalized measure:

$$\int_{\Lambda_n} P^u(\mathrm{d}\omega) = \frac{1}{P(h_n \in \mathrm{d}u)} \int_{\{\omega\in\Lambda_n : h_n(\omega)\in\mathrm{d}u\}} P(\mathrm{d}\omega) = 1. \tag{151}$$

The extension of the microcanonical measure $P^u(\mathrm{d}\omega)$ to macrostates follows the standard rules of probability theory. Given a macrostate $M_n(\omega)$, we define $P^u(M_n \in \mathrm{d}m)$ to be the conditional or constrained probability measure given by

$$P^u(M_n \in \mathrm{d}m) = P(M_n \in \mathrm{d}m|h_n \in \mathrm{d}u) = \frac{P(h_n \in \mathrm{d}u, M_n \in \mathrm{d}m)}{P(h_n \in \mathrm{d}u)}, \tag{152}$$

where

$$P(h_n \in \mathrm{d}u, M_n \in \mathrm{d}m) = \int_{\{\omega\in\Lambda_n : h_n(\omega)\in\mathrm{d}u, M_n(\omega)\in\mathrm{d}m\}} P(\mathrm{d}\omega) \tag{153}$$

is the joint probability of h_n and M_n. It is this probability measure that we have to use to find the most probable values of M_n given that the system represented by the Hamiltonian H_n has a fixed energy $H_n = U$ or, equivalently, a fixed mean energy

$h_n = U/n = u$. The latter expression of the energy constraint involving h_n is generally preferred over the former involving H_n, since we are interested in finding the most probable values of M_n in the large-n or thermodynamic limit. Thermodynamic limits involving a different rescaling of the energy are also conceivable, depending on the form of the Hamiltonian.

5.3.2. Microcanonical large deviations

The theory of large deviations enters in the description of the microcanonical ensemble as a basic tool for finding the values of M_n that maximize the microcanonical probability measure $P^u(M_n \in \mathrm{d}m)$. From our knowledge of sample means, we should expect at this point to be able to prove a large deviation principle for $P^u(M_n \in \mathrm{d}m)$, and to find the most probable values of M_n by locating the global minima of the corresponding rate function. As shown next, this is possible if a large deviation principle holds for the *unconstrained* measure $P(M_n \in \mathrm{d}m)$, and if there exists a contraction of M_n to h_n. In this case, the equilibrium values of M_n that globally minimize the rate function of the *constrained* measure $P^u(M_n \in \mathrm{d}m)$ can be calculated as the global minima of the rate function of the *unconstrained* measure $P(M_n \in \mathrm{d}m)$ subject to the constraint $h_n(\omega) = u$ [7,10].

To prove this result, consider a macrostate $M_n(\omega)$, and suppose that a large deviation principle holds for this macrostate with respect to the unconstrained prior measure $P(\mathrm{d}\omega)$, that is,

$$P(M_n \in \mathrm{d}m) = \int_{\{\omega\in\Lambda_n : M_n(\omega)\in \mathrm{d}m\}} P(\mathrm{d}\omega) \asymp \mathrm{e}^{n\tilde{s}(m)}\,\mathrm{d}m. \tag{154}$$

The rate function of this large deviation principle is written without the usual minus sign to conform with the notation used in physics. The negative "rate function" $\tilde{s}(m)$ is called the *macrostate entropy* of M_n, since it effectively corresponds to the entropy of M_n defined with the volume measure $\Omega(M_n \in \mathrm{d}m)$ up to the constant $\ln|\Lambda|$.

Suppose now that the mean energy or energy per particle $h_n(\omega)$ can be rewritten as a function of the macrostate $M_n(\omega)$. That is to say, suppose that there exists a bounded, continuous function $\tilde{h}(m)$ of M_n, called the *energy representation function*, such that $h_n(\omega) = \tilde{h}(M_n(\omega))$ for all $\omega \in \Lambda_n$ or, more generally, such that

$$|h_n(\omega) - \tilde{h}(M_n(\omega))| \to 0 \tag{155}$$

uniformly over all $\omega \in \Lambda_n$ as $n \to \infty$. Given that this function exists, it is readily seen that the most probable values m of $M_n(\omega)$ with respect to $P^u(M_n \in \mathrm{d}m)$ are those that maximize the macrostate entropy $\tilde{s}(m)$ subject to the constraint $\tilde{h}(m) = u$. To be sure, construct the explicit large deviation principle for $P^u(M_n \in \mathrm{d}m)$. Assuming that $P(M_n \in \mathrm{d}m)$ satisfies the large deviation principle shown in (154), it follows by contraction that $P(h_n \in \mathrm{d}u)$ also satisfies a large deviation principle which we write, as before, as

$$P(h_n \in \mathrm{d}u) \asymp \mathrm{e}^{ns(u)}\,\mathrm{d}u. \tag{156}$$

Combining these large deviations in the expression of $P^u(M_n \in \mathrm{d}m)$ shown in Eq. (152), we then obtain

$$P^u(M_n \in \mathrm{d}m) \asymp \mathrm{e}^{-nI^u(m)}\,\mathrm{d}m, \tag{157}$$

where

$$I^u(m) = \begin{cases} s(u) - \tilde{s}(m) & \text{if } \tilde{h}(m) = u \\ \infty & \text{otherwise.} \end{cases} \tag{158}$$

The rate function $I^u(m)$ is similar to the rate function $I_2^s(\mu)$ discussed in connection with the contraction of the level-2 large deviations to the level-1 large deviations; see Eq. (131). The main point to observe here is that the global minimizers of $I^u(m)$, which correspond to the equilibrium values of M_n in the microcanonical ensemble with $h_n = u$, are the global maximizers of the macrostate entropy $\tilde{s}(m)$ subject to the constraint $\tilde{h}(m) = u$. Denoting by $\mathcal{E}^u$ the set of all such *equilibrium values* or *equilibrium states*, we then write

$$\mathcal{E}^u = \{m : I^u(m) = 0\} = \{m : m \text{ globally maximizes } \tilde{s}(m) \text{ with } \tilde{h}(m) = u\}. \tag{159}$$

The class of macrostates for which $\mathcal{E}^u$ can be calculated using the formula above depends on the model studied and, more precisely, on the form of its Hamiltonian. This point will be discussed in more detail later.

It useful to know that $\mathcal{E}^u$ can be calculated, at least in theory, without the macrostate entropy $\tilde{s}(m)$ and the energy representation function $\tilde{h}(m)$. If we can prove, for instance, that $P(h_n \in \mathrm{d}u, M_n(\omega) \in \mathrm{d}m)$ satisfies a joint large deviation principle of the form

$$P(h_n \in \mathrm{d}u, M_n(\omega) \in \mathrm{d}m) \asymp \mathrm{e}^{n\tilde{s}(u,m)}\,\mathrm{d}u\,\mathrm{d}m, \tag{160}$$

then we obtain

$$P^u(M_n \in \mathrm{d}m) \asymp \mathrm{e}^{-nj^u(m)}\,\mathrm{d}m, \tag{161}$$

where $J^u(m) = s(u) - \tilde{s}(u, m)$. In this case,

$$\mathcal{E}^u = \{m : J^u(m) = 0\} = \{m : m \text{ globally maximizes } \tilde{s}(u, m)\}. \tag{162}$$

One may also attempt to obtain $I^u(m)$ directly using the Gärtner–Ellis Theorem. This method, however, is of limited use, since the calculation of the scaled cumulant generating function of M_n in the microcanonical ensemble involves a constrained integral on Λ_n which can be evaluated explicitly only for certain combinations of macrostates and Hamiltonians (e.g., non-interacting particles). The Gärtner–Ellis Theorem is also limited in that it cannot be used, as we have seen, to calculate nonconvex rate functions, which implies that it cannot be used to calculate nonconcave entropy functions. In this case, the representation of $\mathcal{E}^u$ based on $\tilde{s}(m)$ and $\tilde{h}(m)$ or $\tilde{s}(u, m)$ alone should be used. The former representation based on $\tilde{s}(m)$ and $\tilde{h}(m)$ is generally more practical.

5.3.3. Einstein's fluctuation theory and the maximum entropy principle

The microcanonical large deviation principles displayed in (157) and (161) provide a precise formulation of Einstein's theory of microcanonical fluctuations [63]. They embody the main result of that theory, which is that probabilities in the microcanonical ensemble can be expressed in terms of entropies. But they also refine that result, in that

- They provide a precise expression of the exponential scaling of $P^u(M_n \in \mathrm{d}m)$ with n, which need not be satisfied by all macrostates. That exponential scaling is somewhat hidden in Einstein's theory in the implicit assumption that the entropy is an extensive quantity.[6]
- They lead us to identify the equilibrium values of M_n not just as those values m maximizing $P^u(M_n \in \mathrm{d}m)$, but, more precisely, as the zeros of $I^u(m)$ or $J^u(m)$, thereby bringing the study of equilibrium states in direct contact with the Law of Large Numbers.[7]
- They suggest a procedure—a scheme—for deriving general *maximum entropy principles* that can be used to find the equilibrium values of M_n, and to calculate $s(u)$ in terms of a maximization involving a macrostate entropy.

The last point simply follows by examining the two representations of the set $\mathcal{E}^u$ of microcanonical equilibrium states, defined in Eqs. (159) and (162). The representation of Eq. (159), which involves the rate function $I^u(m)$, implies on the one hand that

$$s(u) = \sup_{m:\tilde{h}(m)=u} \tilde{s}(m). \tag{163}$$

On the other hand, the representation of Eq. (162), which involves $J^u(m)$ instead of $I^u(m)$, implies that

$$s(u) = \sup_m \tilde{s}(u, m). \tag{164}$$

These variational formulae can also be derived from the contraction principle. In the first formula, the contraction is the energy representation function, whereas in the second, the contraction is the map $(u, m) \to u$. Each formula provides, in the end, a general *maximum entropy principle* that can be used to calculate the microcanonical entropy $s(u)$ from the knowledge of a macrostate entropy. The well-known maximum entropy principle of Jaynes [64,65] is a particular application of these formulae, obtained by considering the Level-2 large deviations of systems of independent particles. This is explained in the next example.

Example 5.1 (*Jaynes's Maximum Entropy Principle*). Consider a system of n particles with individual energies $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, $\varepsilon_i \in \Lambda$. Assuming that the particles do not interact with each other, we write the mean energy h_n of the n particles as

$$h_n = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i. \tag{165}$$

An obvious choice of macrostate for this model, apart from the mean energy itself, is the empirical vector or *one-particle energy distribution*

$$L_n(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n} \delta(\varepsilon_i - \varepsilon) \tag{166}$$

[6] Of course, one can always put a probability P in the form $P = \mathrm{e}^S$ simply by defining $S = \ln P$. The non-trivial result stated in Einstein's theory, and more precisely in large deviation theory, is that $S = \ln P$ is extensive with the number of particles, which means that P decays exponentially fast with the number of particles. This statement is the essence of the large deviation principle.

[7] This shows, incidentally, that thermodynamics and statistical mechanics are not really concerned about *average* values so much as about *most probable* values. Equilibrium states are, first and foremost, most probable states.

which counts the relative number of particles having an energy ε. The energy representation for this choice of macrostate is

$$\tilde{h}(L_n) = \int_\Lambda \varepsilon L_n(\varepsilon)\, \mathrm{d}\varepsilon. \tag{167}$$

From Sanov's Theorem, the large deviations of L_n are ruled by the relative entropy $I_\rho(\mu)$. For the uniform prior $\rho = |\Lambda|^{-1}$, $I_\rho(\mu)$ is related to the Boltzmann–Gibbs–Shannon entropy

$$\tilde{s}(\mu) = -\int_\Lambda \mathrm{d}\varepsilon\, \mu(\varepsilon) \ln \mu(\varepsilon) \tag{168}$$

through $I_\rho(\mu) = -\tilde{s}(\mu) + \ln|\Lambda|$. Therefore,

$$s(u) = \sup_{l:\tilde{h}(\mu)=u} \tilde{s}(\mu) - \ln|\Lambda| \tag{169}$$

by the general maximum entropy principle derived in (163). For independent particles, the microcanonical entropy $s(u)$ is thus obtained by maximizing the Boltzmann–Gibbs–Shannon entropy $\tilde{s}(\mu)$ subject to the energy constraint $\tilde{h}(\mu) = u$. It is this version of the maximum entropy principle that we refer to as Jaynes's maximum entropy principle [64,65].

A variational problem similar to the one displayed above was solved in Example 4.10 when treating the contraction of Level-2 to Level-1. Its explicit solution, re-written in a more thermodynamic form, is

$$\mu_\beta(\varepsilon) = \frac{\mathrm{e}^{-\beta\varepsilon}}{Z(\beta)}, \quad Z(\beta) = \int_\Lambda \mathrm{e}^{-\beta\varepsilon}\, \mathrm{d}\varepsilon \tag{170}$$

with β implicitly determined by $\tilde{h}(\mu_\beta) = u$ or, equivalently, by $Z'(\beta)/Z(\beta) = u$. Similarly as in Example 4.10, we therefore obtain

$$s(u) = \tilde{s}(\mu_\beta) - \ln|\Lambda| = \beta u - \varphi(\beta), \tag{171}$$

which is nothing but Cramér's Theorem written in terms of $s(u)$ and $\varphi(\beta)$.[8] Since the linear form of $\tilde{h}(L_n)$ is directly related to the additive form of h_n, the explicit expression of μ_β does not carry over to the case where there is some interaction between the particles. Thus, strictly speaking, Jaynes's maximum entropy principle is only applicable to non-interacting particles.[9]

5.3.4. Treatment of particular models

The microcanonical equilibrium properties of systems of non-interacting particles can always be treated, as in the previous example, at the Level-2 of large deviations using Sanov's Theorem, or directly at the Level-1 using Cramér's Theorem (see, e.g., [66]). These two levels of large deviations can also be used in general to study the equilibrium properties of mean-field models of particles involving an all-to-all coupling between particles. Examples of such models, for which the general maximum entropy principles mentioned before have been applied successfully, include the mean-field versions of the Curie–Weiss model [8,10,67] and its parent model, the Potts model [10,68–70], the Blume–Emery–Griffiths model [71–73], the mean-field Hamiltonian model [74], as well as mean-field versions of the spherical model [75,76], and the ϕ^4 model [77–79]. In all of these models, the energy representation function is either a nonlinear function of the empirical vector (Level-2) or a function of properly-chosen Level-1 macrostates, commonly referred to as *mean fields* or *order parameters*. This is illustrated in the next two examples.

Example 5.2 (*Mean-Field Potts Model*). The mean-field Potts model with q states is defined by the Hamiltonian

$$H_n(\omega) = -\frac{1}{2n} \sum_{i,j} \delta_{\omega_i,\omega_j}, \tag{172}$$

where $\omega_i \in \Lambda = \{1, 2, \ldots, q\}$. The factor n in front of the sum is there to make the energy an extensive variable or, equivalently, to make the mean energy an intensive variable. In terms of the empirical vector

$$L_n(\omega) = (L_{n,1}(\omega), L_{n,2}(\omega), \ldots, L_{n,q}(\omega)), \qquad L_{n,j}(\omega) = \frac{1}{n} \sum_{i=1}^{n} \delta_{i,j} \tag{173}$$

[8] Cramér's Theorem appears here because the mean energy h_n of non-interacting particles is a sample mean of IID random variables under the uniform prior measure $P(\mathrm{d}\omega)$.

[9] The form of the entropy function also depends on the observable considered. To be more precise, one should therefore say that Jaynes's maximum entropy principle applies only to the *one-particle distribution* of *non-interacting* particle systems. Other maximum principles that are applicable to other observables and other systems can be derived, following this section, by obtaining an explicit large deviation expression for the probability of a given observable, defined in the context of a given system and ensemble. In the end, it is probability that defines the form of a maximum entropy principle, not the choice of an arbitrary entropy function or some *ad hoc* information-based argument.

we obviously have $h_n(\omega) = \tilde{h}(L_n(\omega))$, where

$$\tilde{h}(\mu) = -\frac{1}{2}\mu \cdot \mu = -\frac{1}{2}\sum_{j=1}^{q} \mu_j^2. \tag{174}$$

The reader is referred to [70] for the complete calculation of $s(u)$ based on this energy representation function, and for the calculation of the equilibrium values of L_n in the microcanonical ensemble. Note that the macrostate entropy that needs to be maximized here is the Boltzmann–Gibbs–Shannon entropy, or relative entropy, as in the case of non-interacting particles; however, now the energy representation function is a nonlinear function of the empirical vector.

Example 5.3 (*Mean-Field ϕ^4 Model [77–79]*). The mean-field ϕ^4 model is defined by the Hamiltonian

$$H_n = \sum_{i=1}^{n}\left(\frac{p_i^2}{2} - \frac{q_i^2}{2} + \frac{q_i^4}{4}\right) - \frac{1}{4n}\sum_{i,j=1}^{n} q_i q_j, \tag{175}$$

where $p_i, q_i \in \mathbb{R}$. We can re-write the mean energy of this model using the following energy representation function:

$$\tilde{h}(k, v, m) = k + v - \frac{m^2}{4}, \tag{176}$$

where

$$k = \frac{1}{2n}\sum_{i=1}^{n} p_i^2, \qquad v = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_i^4}{4} - \frac{q_i^2}{2}\right), \quad m = \frac{1}{n}\sum_{i=1}^{n} q_i \tag{177}$$

are, respectively, the mean kinetic energy, the mean potential energy, and the mean magnetization of the model. The entropy functions of these macrostates can be derived using the Gärtner–Ellis Theorem, since they are all strictly convex. The entropy $\tilde{s}(m)$ of m, for example, is the magnetization entropy found in Example 2.5, which is also the negative of the rate function calculated for the binary sample mean of Example 4.1. The calculation of the two other macrostate entropies $\tilde{s}(k)$ and $\tilde{s}(v)$ is reported in [79]. In the end, we obtain $s(u)$ by solving

$$s(u) = \sup_{(k,v,m):\tilde{h}(k,v,m)=u} \tilde{s}(k) + \tilde{s}(v) + \tilde{s}(m). \tag{178}$$

The details of this calculation can be found in [79].

When going beyond non-interacting and mean-field systems, two different classes of systems must be distinguished: those involving *long-range* interactions, such as systems of gravitating particles, and those involving *short-range* interactions, such as the nearest-neighbor Ising model. From the point of view of the formalism developed here, long-range systems (see [80] for a definition of long-range interactions) are similar to mean-field systems, in that their mean energy often admits a representation function involving Level-1 or Level-2 macrostates [81]. This is the case, for example, for systems of gravitating particles and plasmas, which can be investigated in the mean-field or Vlasov limit using the empirical distribution (see [80,82,83] for recent reviews). Some statistical models of two-dimensional (2D) turbulence can also be treated with an energy representation function involving a relatively simple macrostate (see, e.g., [62]). A particularity of these models is that the prior distribution $P(\mathrm{d}\omega)$ is not always chosen to be the uniform measure [84]. Another model worth mentioning, finally, is the so-called α-Ising model in one dimension, which admits the local magnetization function as a mean-field [74]. Other models of long-range systems are discussed in [82,83].

Systems involving short-range interactions are much more complicated to study due to the fact that their large deviation analysis must be based on the Level-3 empirical process mentioned in Section 4. The empirical process can be used in principle to study non-interacting and mean-field models, but this is never done in practice, as there are simpler macrostates to work with. The problem with short-range models is that the empirical process is, in general, the *only* macrostate admitting an energy representation function. This is the case, for example, for the nearest-neighbor Ising model, which has been studied extensively in one and two dimensions from the point of view of the empirical process (see, e.g., [8,10,60,85]). We summarize in the next example the equilibrium properties of the mean magnetization of the 2D version of this model, obtained by contracting the large deviations of the empirical process down to the mean energy and mean magnetization. The main sources for this example are Ellis [10], Pfister [85], and Kastner [86]. For a discussion of the Ising model in one dimension, see [8,87].

Example 5.4 (*2D Ising Model*). Consider the 2D nearest-neighbor Ising model, defined by the usual Hamiltonian

$$H_n = -\frac{1}{2}\sum_{\langle i,j\rangle} \sigma_i\sigma_j, \quad \sigma_i \in \{-1, 1\}, \tag{179}$$

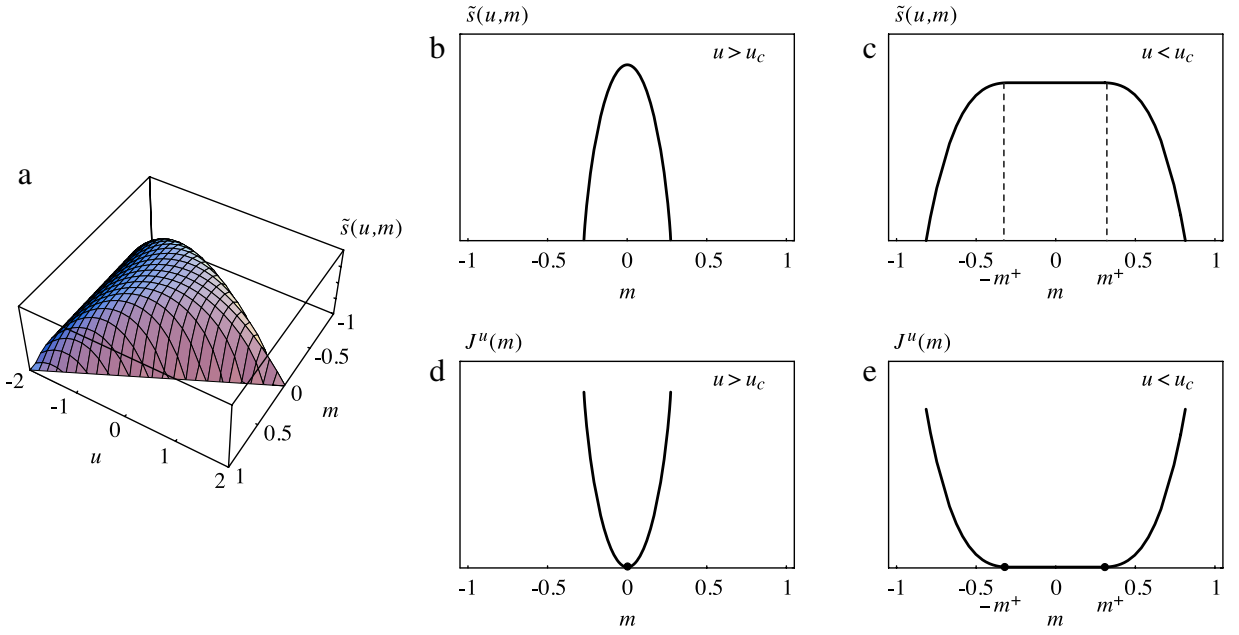

Fig. 13. (a) Sketch of the joint entropy $\tilde{s}(u, m)$ of the 2D Ising model (after [86]). (b) Projection of $\tilde{s}(u, m)$ as a function of m for $u < u_c$. (c) Projection of $\tilde{s}(u, m)$ as a function of m for $u > u_c$. (d)–(e) Microcanonical rate function $J^u(m)$ above and below the critical mean energy u_c.

where $\langle i, j\rangle$ denotes first-neighbor sites on the finite 2D square lattice containing n spins. The entropy of this model is sketched, following [86], in Fig. 13 as a function of the mean energy u and mean magnetization m. There are many properties of this entropy worth noting from the point of large deviations:

- $\tilde{s}(u, m)$ is strictly concave in m for $u \in [u_c, 2]$, where $u_c = -\sqrt{2}$, with a maximum located at $m = 0$; see Fig. 13(b).
- $\tilde{s}(u, m)$ is concave in m for $u \in [-2, u_c)$, but not strictly concave. In fact, for this range of mean energies, $\tilde{s}(u, m)$ is constant in m in the interval $[-m^+(u), m^+(u)]$; see Fig. 13(c). The boundary point $m^+(u)$ is called the *spontaneous magnetization*, and is such that $m^+(u_c) = 0$ and $m^+(u) \to 1$ as $u \to -2$.
- By contraction, $s(u) = \tilde{s}(u, 0)$, since $m = 0$ is always a maximum of $\tilde{s}(u, m)$ for all $u \in [-2, 2]$. Although not plotted, $s(u)$ is known to be concave and differentiable, which means that it can be calculated in principle as the Legendre transform of the canonical free energy calculated by Onsager [88].
- The rate function $J^u(m) = s(u) - \tilde{s}(u, m)$ has a unique minimum and zero for $[u_c, 2]$, corresponding to the unique equilibrium value of the mean magnetization in the microcanonical ensemble for all $u \in [u_c, 2]$; see Fig. 13(d).
- For $u \in [-2, u_c)$, $J^u(m)$ is zero for all m in the interval $[-m^+(u), m^+(u)]$, which is called the *phase transition interval* or *phase coexistence region* [8,10]; see Fig. 13(e).
- In the coexistence region, $P^u(M_n \in \mathrm{d}m)$ decays as $\mathrm{e}^{-a\sqrt{n}}$, where a is some positive constant, instead of the anticipated (bulk) decay e^{-bn}, $b > 0$. This "slower" large deviation principle describes a surface effect related to a change of phase, and depends on boundary conditions imposed on the model (see, e.g., [85,89]).

The fact that $J^u(m)$ is zero over the whole coexistence region when $u < u_c$ means that the large deviations of the mean magnetization are undetermined in that region [10,11]. In particular, we cannot conclude from the shape of $J^u(m)$ that there is a whole interval of equilibrium values for the mean magnetization. The actual equilibrium values are determined by the refined large deviation principle for $P^u(M_n \in \mathrm{d}m)$ mentioned in the last point above. The same remark applies when studying the 2D Ising model in the canonical ensemble as a function of the temperature; see Example 5.6.

5.4. Canonical ensemble

The canonical ensemble differs from the microcanonical ensemble in the way microstates are weighted. In the microcanonical ensemble, the control parameter is the energy U or the mean energy $u = U/n$, and the microstates ω are taken to be distributed according to the constrained prior distribution $P^u(\mathrm{d}\omega)$, which assigns the same probabilistic weight to all the microstates with the same energy U or mean energy u. In the canonical ensemble, the control parameter is the *temperature* T or, equivalently, the *inverse temperature* $\beta = (k_B T)^{-1}$, and the relevant probability measure that one considers on Λ_n is the so-called *canonical* or *Gibbs measure*

$$P_\beta(\mathrm{d}\omega) = \frac{\mathrm{e}^{-\beta H_n(\omega)}}{Z_n(\beta)} P(\mathrm{d}\omega). \tag{180}$$

In this expression, $Z_n(\beta)$ is the n-particle partition function defined earlier in Eq. (146). We will not discuss the physical interpretation of $P_\beta(\mathrm{d}\omega)$, apart from mentioning that it arises as the probability distribution of a sample system with Hamiltonian $H_n(\omega)$ placed in thermal contact with a heat bath at inverse temperature β (see, e.g., Sec. 28 of [90]). It should be mentioned also that, although the expression of $P_\beta(\mathrm{d}\omega)$ involves the exponential function, that expression is not a large deviation principle—it is just the definition of a measure on Λ_n.

The derivation of a large deviation principle for a general macrostate $M_n(\omega)$ in the canonical ensemble proceeds similarly as in the microcanonical ensemble. On the one hand, if an energy representation $\tilde{h}(m)$ and a macrostate entropy $\tilde{s}(m)$ exist for M_n, it is relatively easy to show that

$$P_\beta(M_n \in \mathrm{d}m) = \int_{\{\omega\in\Lambda_n : M_n(\omega)\in \mathrm{d}m\}} P_\beta(\mathrm{d}\omega) \asymp \mathrm{e}^{-nI_\beta(m)}\,\mathrm{d}m \tag{181}$$

where

$$I_\beta(m) = \beta\tilde{h}(m) - \tilde{s}(m) - \varphi(\beta) \tag{182}$$

and $\varphi(\beta)$ is the free energy defined in Eq. (147) [62]. On the other hand, if one knows that a joint large deviation holds for $P(h_n \in \mathrm{d}u, M_n \in \mathrm{d}m)$ with macrostate entropy $\tilde{s}(u, m)$, then

$$P_\beta(M_n \in \mathrm{d}m) \asymp \mathrm{e}^{-nJ_\beta(m)}\,\mathrm{d}m, \tag{183}$$

where

$$J_\beta(m) = \inf_u\{\beta u - \tilde{s}(u, m)\} - \varphi(\beta). \tag{184}$$

These two large deviation principles generalize Einstein's theory of fluctuations to the canonical ensemble. The rate functions $I_\beta(m)$ and $J_\beta(m)$ that we obtain in this ensemble are the macrostate free energies that form the basis of the Ginzburg–Landau theory of phase transitions [90].

As in the microcanonical ensemble, the global minima of $I_\beta(m)$ or $J_\beta(m)$ define the most probable or *equilibrium* values of the macrostate M_n which now appear in the canonical ensemble with inverse temperature β. The set $\mathcal{E}_\beta$ containing these canonical equilibrium values is thus defined as

$$\mathcal{E}_\beta = \{m : I_\beta(m) = 0\} \quad \text{or } \mathcal{E}_\beta = \{m : J_\beta(m) = 0\}, \tag{185}$$

depending on the rate function (I_β or J_β, respectively) used for analyzing the canonical large deviations of M_n. Equivalently, we have

$$\mathcal{E}_\beta = \{m : m \text{ is a global minimum of } I_\beta(m)\} \tag{186}$$

or

$$\mathcal{E}_\beta = \{m : m \text{ is a global minimum of } J_\beta(m)\}, \tag{187}$$

The canonical analog of the maximum entropy principle that we obtain from these definitions of $\mathcal{E}_\beta$ is called the *minimum free energy principle*, and is expressed as

$$\varphi(\beta) = \inf_m\{\beta\tilde{h}(m) - \tilde{s}(m)\} \tag{188}$$

or

$$\varphi(\beta) = \inf_m \inf_u\{\beta u - \tilde{s}(u, m)\}, \tag{189}$$

depending again on the rate function (I_β or J_β, respectively) used. These formulae play the same role as the maximum entropy principle, in that they enable us to obtain the thermodynamic free energy $\varphi(\beta)$ as the solution of a variational problem involving a function that we call the macrostate free energy. The solutions of this variational principle are the canonical equilibrium values of M_n.

There is a connection between the maximum entropy of the microcanonical ensemble that can be made here. Since the two infima in Eq. (189) can be interchanged, we can used the maximum entropy principle of (164) to write

$$\varphi(\beta) = \inf_u \inf_m\{\beta u - \tilde{s}(u, m)\} = \inf_u \left\{\beta u - \sup_m \tilde{s}(u, m)\right\} = \inf_u\{\beta u - s(u)\}. \tag{190}$$

We recover, therefore, the basic Legendre–Fenchel transform found in Eq. (148). The formulae (188) and (189) can also be derived by recasting the integral defining $Z_n(\beta)$ as an integral over M_n, and by applying Laplace's Method to the latter integral [72]. In the case where $\tilde{h}(m)$ and $\tilde{s}(m)$ exist, for example,

$$Z_n(\beta) \asymp \int \mathrm{e}^{-n\beta\tilde{h}(m)}\,P(M_n \in \mathrm{d}m) \asymp \int \mathrm{e}^{-n\{\beta\tilde{h}(m)-\tilde{s}(m)\}}\,\mathrm{d}m \asymp \mathrm{e}^{-n\inf\{\beta\tilde{h}(m)-\tilde{s}(m)\}}. \tag{191}$$

The class of macrostates for which large deviation principles can be derived in the canonical ensemble is exactly the same as in the microcanonical ensemble, since the rate functions of these two ensembles are built from the same energy

representation function and macrostate entropies. Hence, if a large deviation principle holds for some macrostate M_n in the microcanonical ensemble, then a large deviation principle also holds for M_n in the canonical ensemble, and vice versa. This is not to say that the two ensembles yield the same sets of equilibrium states. There are models for which the microcanonical equilibrium set $\mathcal{E}^u$ and the canonical equilibrium set $\mathcal{E}_\beta$ are *equivalent*, in the sense that they can be put in a one-to-one correspondence. But there are also models for which the two sets are not equivalent. This problem of ensemble equivalence is the subject of the next subsection.

To close our discussion of canonical large deviations, we discuss next two examples of canonical large deviations: one involving the mean energy h_n, the other the mean magnetization of the 2D Ising model. The first example is important for understanding the content of the next section. For a discussion of other large deviation results derived in the canonical ensemble, the reader is referred to [7,8,10,74].

Example 5.5 (*Equilibrium Mean Energy*). The probability distribution of the mean energy h_n in the canonical ensemble is

$$P_\beta(h_n \in \mathrm{d}u) = \int_{\{\omega\in\Lambda_n : h_n(\omega)\in \mathrm{d}u\}} P_\beta(\mathrm{d}\omega) = \frac{\mathrm{e}^{-n\beta u}}{Z_n(\beta)} P(h_n \in \mathrm{d}u). \tag{192}$$

Assuming that the microcanonical entropy $s(u)$ exists, that is, assuming that $P(h_n \in \mathrm{d}u) \asymp \mathrm{e}^{ns(u)}\,\mathrm{d}u$, then

$$P_\beta(h_n \in \mathrm{d}u) \asymp \mathrm{e}^{-nI_\beta(u)}\,\mathrm{d}u, \qquad I_\beta(u) = \beta u - s(u) - \varphi(\beta). \tag{193}$$

The mean energy u_β realized at equilibrium in the canonical ensemble at inverse temperature β is determined from this large deviation principle by requiring that $I_\beta(u_\beta) = 0$. Thus,

$$\varphi(\beta) = \inf_u\{\beta u - s(u)\} = \beta u_\beta - s(u_\beta). \tag{194}$$

By Legendre duality, u_β must be such that $\varphi'(\beta) = u_\beta$ if $\varphi(\beta)$ is differentiable. If $s(u)$ is differentiable, then we also have $s'(u_\beta) = \beta$, thereby recovering the standard thermodynamic definition of the inverse temperature. Observe, however, that $I_\beta(u)$ may have many critical points satisfying $s'(u) = \beta$; the global minimizer u_β is only one of them.

Example 5.6 (*2D Ising Model*). The rate function $J_\beta(m)$ associated with the mean magnetization of the 2D Ising model in the canonical ensemble is sketched in Fig. 14. This rate function is directly obtained from the macrostate entropy $\tilde{s}(u, m)$ discussed in Example 5.4 via Eq. (184). The properties of $J_\beta(m)$ are similar to its microcanonical counterpart $J^u(m)$. In particular,

- $J_\beta(m)$ is a symmetric function of m and is convex for all $\beta \in \mathbb{R}$. One difference with $J^u(m)$ is that $J_\beta(m)$ is finite for all $m \in (-1, 1)$.
- For $\beta \le \beta_c$, $J_\beta(m)$ is *strictly* convex, which implies that it has a unique global minimum located at $m = 0$; see Fig. 14(a). In other words, for $\beta \ge \beta_c$, $m_\beta = 0$ is the unique equilibrium mean magnetization.
- For $\beta > \beta_c$, $J_\beta(m)$ is convex but achieves, as in the case of $J^u(m)$, its zero on a whole interval of mean magnetizations denoted by $[-m^+(\beta), m^+(\beta)]$; see Fig. 14(b). This canonical phase transition interval is such that $m^+(\beta) \to 0$ when $\beta \to \beta_c$ and $m^+(\beta) \to 1$ when $\beta \to \infty$.
- As in the microcanonical case, $P(M_n \in \mathrm{d}m) \asymp \mathrm{e}^{-b\sqrt{n}}$ with $b > 0$ inside the phase transition interval. This "slower" large deviation principle is also a surface effect.
- The phase transition in the canonical ensemble is second-order, as in the microcanonical ensemble, with critical inverse temperature $\beta_c = s'(u_c)$. The critical exponent associated with the phase transition is non-trivial, as is well known, because $J_\beta(m)$ has no Taylor expansion around $m = 0$ above the critical β_c.

The 2D Ising model is interesting from the point of view of large deviation theory because it shows that the interactions or correlations between the components of a system (here the spins) can change the scaling of a large deviation principle, and can lead to a breakdown of the Central Limit Theorem and the Law of Large Numbers. The breakdown of the Central Limit Theorem is related here to the fact that $J_\beta(m)$ is not locally quadratic around $m = 0$ when $\beta > \beta_c$. The breakdown of the Law of Large Numbers, on the other hand, is related to the fact that $J_\beta(m)$ does not have a unique concentration point (i.e., global minimum or equilibrium state) for $\beta > \beta_c$.

5.5. Equivalence of ensembles

We have seen before that not all rate functions I can be obtained as the Legendre–Fenchel transform of a scaled generating cumulant function λ; only those that are convex are such that $I = \lambda^*$. When applied to entropy functions, this observation directly implies that *entropy functions that are nonconcave cannot be calculated as the Legendre–Fenchel transform of free energies*. Consider, for instance, the entropy $s(u)$ as a function of the mean energy. Then we have $s = \varphi^*$ if s is concave in u, but $s \ne \varphi^*$ if s is nonconcave in u; see Fig. 15. In the first case, namely when $s(u)$ is concave, we say that the microcanonical and canonical ensembles are *equivalent at the thermodynamic level* because $s(u)$ and $\varphi(\beta)$ are then one-to-one related by

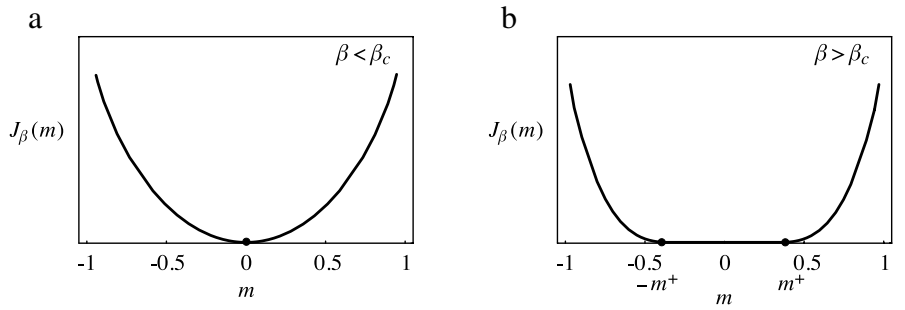

Fig. 14. Canonical rate function $J_\beta(m)$ for (a) $\beta < \beta_c$ and (b) $\beta > \beta_c$.

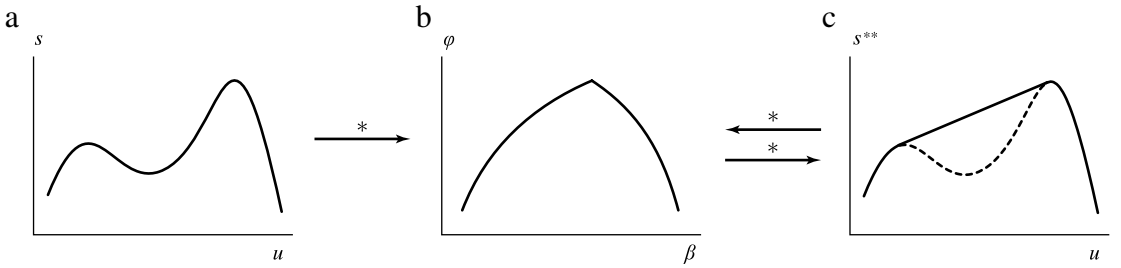

Fig. 15. Thermodynamic nonequivalence of the microcanonical and canonical ensembles as related to the nonconcavity of the entropy (see also Fig. 9). The Legendre–Fenchel transform of $\varphi(\beta)$ yields the concave envelope $s^{**}(u)$ of $s(u)$ rather than $s(u)$ itself. The Legendre–Fenchel transform of both $s(u)$ and $s^{**}(u)$ yield $\varphi(\beta)$.

Legendre–Fenchel transform. In the second case, namely when $s(u)$ is nonconcave, we say that the two ensembles are *nonequivalent at the thermodynamic level*, since part of $s(u)$ cannot be obtained from $\varphi(\beta)$ [62]. What is obtained by taking the Legendre–Fenchel transform of $\varphi(\beta)$ is the *concave envelope* $s^{**}(u)$ rather than $s(u)$ itself; see Fig. 15. Recall that $\varphi = s^*$ always holds, as noted after Eq. (148), so that the nonequivalence of the microcanonical and canonical ensembles only goes in one direction: the free energy can always be obtained as the Legendre–Fenchel transform of the entropy, but the entropy can be obtained as the Legendre–Fenchel transform of the free energy only when the entropy is concave.

These results and definitions are simple applications of the results that we have discussed before in the context of nonconvex rate functions. Further applications arise from what we know about Legendre–Fenchel transforms. In particular, the result relating the nonconvexity or affinity of rate functions with the nondifferentiability of the scaled cumulant generating function implies at the level of $s(u)$ and $\varphi(\beta)$ that, if $s(u)$ is nonconcave or is affine, then $\varphi(\beta)$ is nondifferentiable. Physically, this means that a first-order phase transition in the canonical ensemble can arise in two ways from the point of view of the microcanonical ensemble: either $s(u)$ is nonconcave or $s(u)$ is affine [91]. The *latent heat* Δu of the phase transition corresponds in both cases to the length of the affine portion of the concave envelope $s^{**}(u)$ of $s(u)$. Indeed, if $s^{**}(u)$ if affine over some open interval (u_l, u_h), then $\varphi(\beta)$ is nondifferentiable at a critical inverse temperature β_c corresponding to the slope of $s^{**}(u)$ over (u_l, u_h); see Figs. 15 and 16. Moreover, $\varphi'(\beta_c + 0) = u_l$ and $\varphi'(\beta_c - 0) = u_h$, so that

$$\Delta u = \varphi'(\beta_c - 0) - \varphi'(\beta_c + 0) = u_h - u_l. \tag{195}$$

The Maxwell or equal-area construction [92,93] used in physics to calculate the latent heat is nothing but the construction of the concave envelope $s^{**}(u)$ [72]; see Fig. 16(c).

These relationships between entropies, free energies, and phase transitions lead us to one last "physical" reformulation of a mathematical result that we have discussed before, namely, the Gärtner–Ellis Theorem. Put simply: *If there is no first-order phase transition in the canonical ensemble, then the microcanonical entropy is the Legendre transform of the canonical free energy*. This follows by noting that if there is no phase transition at the level of the free energy or only a second-order phase transition, then the free energy is once-differentiable, which implies by the Gärtner–Ellis Theorem that the entropy can be calculated as the Legendre transform of the free energy. Of course, a concave yet affine entropy, such as the entropy $\tilde{s}(u, m)$ of the 2D Ising model, can also be calculated as the Legendre(–Fenchel) transform of the free energy, even though the latter has a nondifferentiable point. But, as in the case of nonconvex rate functions, it is impossible to distinguish from the sole knowledge of the free energy an affine entropy from a nonconcave entropy.

Examples of models with nonconcave entropies include the mean-field Blume–Emery–Griffiths model [71–73], the mean-field Potts model [70,94], some models of plasmas [95] and 2D turbulence [84,96] mentioned before, as well as models of gravitational systems (see [97] for a recent review). The latter systems were historically the first to be discovered as having nonconcave entropies or, equivalently, as having negative heat capacities due to the long-range nature of the gravitational

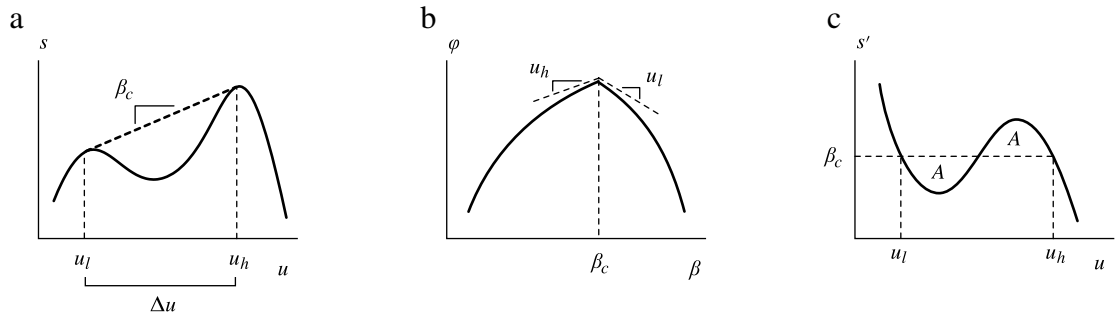

Fig. 16. (a) Nonconcave entropy $s(u)$ and its concave envelope $s^{**}(u)$. (b) Associated free energy $\varphi(\beta)$ characterized by a nondifferentiable point. (c) Maxwell's construction: The two areas A defined by the intersection of $s'(u)$ and $s^{**\prime}(u)$ are equal.

force (see [80,98]). In general, the long-range nature of the interaction in a many-particle system is a necessary, but not sufficient, condition for having nonconcave entropies.

In some cases, the entropy may be concave as a function of u alone but nonconcave as a function of some other macrostate. The mean-field ϕ^4 model, for example, has a concave $s(u)$ but a nonconcave macrostate entropy $\tilde{s}(u, m)$ involving the mean energy u and mean magnetization m [77–79]. Thus, although $s(u)$ for this model can be calculated in the spirit of the Gärtner–Ellis Theorem as the Legendre–Fenchel transform of $\varphi(\beta)$, $\tilde{s}(u, m)$ cannot be obtained as the Legendre–Fenchel transform of a free energy function because that Legendre–Fenchel transform yields a concave function. The nonconcave $s(u, m)$ can be obtained, however, by other methods. One suggested by large deviation theory is to obtain $s(u, m)$ by contraction of another macrostate entropy that can hopefully be calculated using the Gärtner–Ellis Theorem; see Example 5.3.

The next example discusses the calculation of $\tilde{s}(u, m)$ in the case where this function is concave.

Example 5.7 (*Concave Entropy Involving Two Macrostates*). If the macrostate entropy $\tilde{s}(u, m)$ is concave, as in the case of the 2D Ising model (see Example 5.4), then it can be expressed as the Legendre–Fenchel transform of a free energy function $\varphi(\beta, \eta)$ involving two variables or *conjugated fields*: one for h_n, the other for M_n. By analogy with the definition $\varphi(\beta)$, $\varphi(\beta, \eta)$ is constructed as

$$\varphi(\beta, \eta) = \lim_{n\to\infty} -\frac{1}{n} \ln \int_{\Lambda_n} \mathrm{e}^{-n\beta h_n - n\eta M_n} P(\mathrm{d}\omega). \tag{196}$$

The concavity of $\tilde{s}(u, m)$ then implies

$$\tilde{s}(u, m) = \inf_{\beta,\eta} \{\beta u + \eta m - \varphi(\beta, \eta)\}. \tag{197}$$

The free energy $\varphi(\beta, \eta)$ can be put in the more familiar form

$$\varphi(\beta, \eta') = \lim_{n\to\infty} -\frac{1}{n} \ln \int_{\Lambda_n} \mathrm{e}^{-n\beta[h_n - \eta' M_n]} P(\mathrm{d}\omega) \tag{198}$$

by defining $\eta' = -\eta/\beta$. In this form, $\varphi(\beta, \eta')$ is nothing but the standard canonical free energy $\varphi(\beta)$ of the modified mean Hamiltonian $h'_n = h_n - \eta' M_n$ involving the magnetic field η'. This field is the parameter of the canonical ensemble which is conjugated to the mean magnetization constraint $M_n = m$ of the microcanonical ensemble, whereas β is the usual canonical parameter conjugated to the mean energy constraint $h_n = u$.

The thermodynamic equivalence of the microcanonical and canonical ensemble is also the basis for the equivalence of Gibbs's entropy and Boltzmann's entropy [99]. This is the subject of the next example.

Example 5.8 (*Boltzmann Versus Gibbs Entropy*). Gibbs's canonical entropy is defined for a discrete set of microstates ω as

$$S_G(\beta) = -\sum_{\omega\in\Lambda_n} P_\beta(\omega) \ln P_\beta(\omega), \tag{199}$$

where $P_\beta(\omega)$ is the canonical probability distribution. From the definition of this distribution, given in Eq. (180), we can write

$$S_G(\beta) = n\beta \langle h_n \rangle_\beta + \ln Z_n(\beta) + \ln |\Lambda_n|, \tag{200}$$

where $\langle\cdot\rangle_\beta$ denotes the expectation with respect to P_β. The thermodynamic limit of this expression is

$$s_G(\beta) = \lim_{n\to\infty} \frac{S_G(\beta)}{n} = \beta u_\beta - \varphi(\beta) + \ln |\Lambda|, \tag{201}$$

assuming that u_β is the unique concentration point of h_n with respect to P_β, that is, the unique equilibrium mean energy at inverse temperature β. This assumption is justified rigorously if $s(u)$ is strictly concave; see Example 5.9. In this case, it is known that $\varphi(\beta)$ is differentiable for all $\beta \in \mathbb{R}$ and $\varphi'(\beta) = u_\beta$ by Legendre duality, so that

$$s_G(\beta) = s(u_\beta) + \ln|\Lambda|. \tag{202}$$

Thus, in the thermodynamic limit, the Gibbs entropy $s_G(\beta)$ is equal (up to a constant) to the Boltzmann entropy $s(u)$ evaluated at the equilibrium mean energy value u_β. This holds again if $s(u)$ is strictly concave. If $s(u)$ is nonconcave or is concave but has an affine part, then u_β need not be unique for a given β; see Example 5.9.

There are many more issues about the equivalence of the microcanonical and canonical ensembles that could be discussed; see, e.g., [82,83]. One that deserves mention, but would take us too far to explain completely is that the two ensembles can be conceived as being equivalent or nonequivalent by comparing the equilibrium sets, $\mathcal{E}^u$ and $\mathcal{E}_\beta$, of each ensemble. This macrostate approach to the problem of ensemble equivalence was studied by Ellis, Haven and Turkington [62] (see also [100]), and has been illustrated so far for a model of 2D turbulence [84], as well as some mean-field spin models [70,72], including a toy spin model [101] based on the nonconvex rate function studied in Example 4.7. The essential result illustrated by these models is, in a simplified form, that the microcanonical and canonical ensembles are equivalent at the macrostate level if and only if they are equivalent at the thermodynamic level, that is, if and only if $s(u)$ is concave. Thus all models with a concave entropy $s(u)$ have equivalent microcanonical and canonical ensembles at the macrostate level. For a simple introduction to these results, see [102]; for the treatment of ensembles other than microcanonical and canonical, see [62].

The last example of this section explains how the concavity of $s(u)$ determines the behavior of the canonical equilibrium mean energy u_β as a function of β. This example illustrates in the simplest way possible the theory of nonequivalent ensembles developed by Ellis, Haven and Turkington [62], and completes Example 5.5.

Example 5.9 (*Canonical Equilibrium Mean Energy Revisited*). We have seen in Example 5.5 that the equilibrium mean energy u_β in the canonical ensemble is given by the zero(s) of the rate function $I_\beta(u)$ displayed in (193). The behavior of u_β as a function of β is determined from this rate function by analyzing the concavity of $s(u)$. Three cases must be distinguished [102]. For simplicity, we assume that $s(u)$ is differentiable in all three cases.

- $s(u)$ *is strictly concave*: In this case, $I_\beta(u)$ has a unique zero u_β for all $\beta \in \mathbb{R}$ such that $\varphi'(\beta) = u_\beta$ and $s'(u_\beta) = \beta$. Furthermore, the range of u_β coincides with the domain of $s(u)$. These two results imply that there is a bijection between u and β: to any u in the domain of $s(u)$, there exists a unique β such that $u = u_\beta$; to any $\beta \in \mathbb{R}$, there exists a unique u in the domain of $s(u)$ such that $u = u_\beta$. This bijection is an expression of the Legendre duality between $s(u)$ and $\varphi(\beta)$.
- $s(u)$ *is nonconcave*: In this case, $I_\beta(u)$ has more than one critical point satisfying $s'(u) = \beta$. The global minimum u_β of $I_\beta(u)$ is one such critical point satisfying the additional condition $s(u_\beta) = s^{**}(u_\beta)$. If $s(u) \neq s^{**}(u)$, then $u \neq u_\beta$ for all $\beta \in \mathbb{R}$. Therefore, all u such that $s(u) \neq s^{**}(u)$ do not appear in the canonical ensemble as equilibrium values of h_n. This explains at the level of the mean energy why there is a first-order phase transition in the canonical ensemble when $s(u)$ is nonconcave.
- $s(u)$ *is concave with an affine part of slope* β_c: In this case, $I_\beta(u)$ has a unique zero u_β for all $\beta \neq \beta_c$. For $\beta = \beta_c$, $I_\beta(u)$ has a whole range of zeros corresponding to the range of affinity of $s(u)$. This range is a coexistence region for the mean energy h_n.

The case of a nonconcave $s(u)$ is illustrated in Fig. 17.

5.6. Existence of the thermodynamic limit

Proving that a large deviation principle holds for a macrostate is equivalent to proving the existence of a thermodynamic limit for an entropy function in the microcanonical ensemble, or a free energy function in the canonical ensemble. This equivalence is obvious in the microcanonical ensemble, for the statement $P(M_n \in \mathrm{d}m) \asymp \mathrm{e}^{n\tilde{s}(m)}\,\mathrm{d}m$ is equivalent to the existence of the limit

$$\lim_{n\to\infty} \frac{1}{n} \ln P(M_n \in \mathrm{d}m) = \tilde{s}(m). \tag{203}$$

In particular, $P(h_n \in \mathrm{d}u) \asymp \mathrm{e}^{ns(u)}\,\mathrm{d}u$ if and only if the limit (142) defining $s(u)$ exists. To establish the same correspondence in the canonical ensemble, note that the existence of the following macrostate free energy for M_n:

$$\tilde{\varphi}(\eta) = \lim_{n\to\infty} -\frac{1}{n} \ln \int_{\Lambda_n} \mathrm{e}^{-n\eta M_n(\omega)}\, P(\mathrm{d}\omega) \tag{204}$$

implies the existence of $\tilde{s}(m)$, since

$$\tilde{s}(m) \leq \tilde{s}^{**}(m) = \inf_\eta \{\eta m - \tilde{\varphi}(\eta)\}. \tag{205}$$

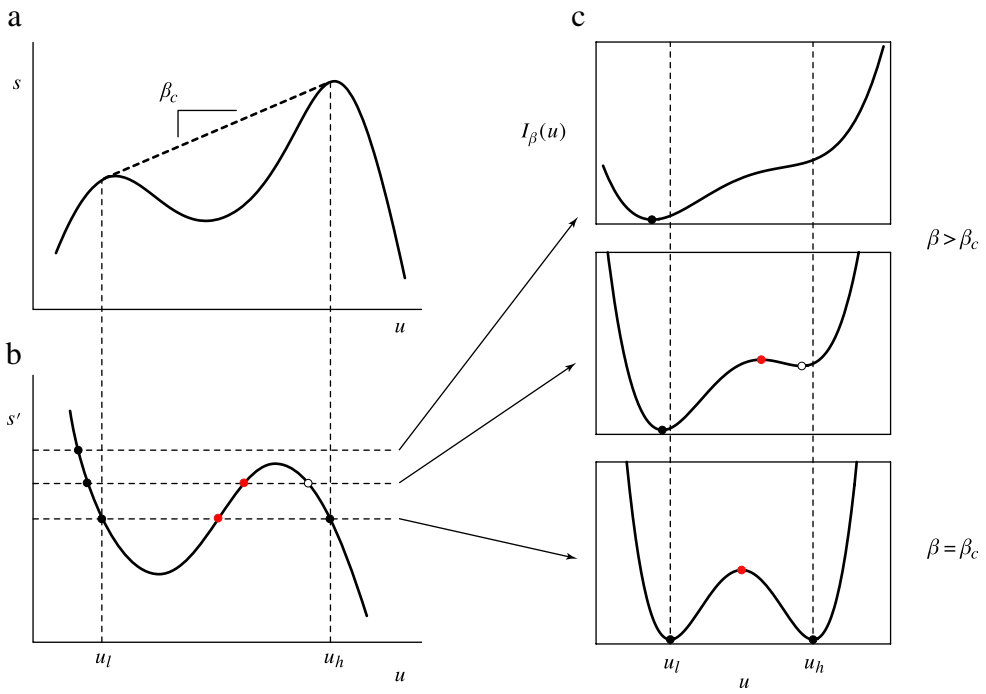

Fig. 17. (a) Generic nonconcave entropy function $s(u)$. (b) Derivative of $s(u)$. The nonconcavity of $s(u)$ translates into a non-monotonic derivative $s'(u)$. (c) Canonical rate function $I_\beta(u)$ for the mean energy, given in Eq. (193). The critical points of $I_\beta(u)$, for a given value of β, satisfy $s'(u) = \beta$. When $s(u)$ is nonconcave, $I_\beta(u)$ may have more than one critical point. The equilibrium mean energy u_β in the canonical ensemble, which corresponds to the global minimum of $I_\beta(u)$ (black dot), is always located in the concave region of $s(u)$, i.e., in the region where $s(u) = s^{**}(u)$. The unstable (red dots) and metastable (open dot) critical points are located in the nonconcave region, i.e., in the region where $s(u) \neq s^{**}(u)$. At $\beta = \beta_c$, $I_\beta(u)$ has two global minima corresponding to u_l and u_h (phase coexistence).

The converse of this result also holds by Varadhan's Theorem: namely, $\tilde{\varphi}(\eta)$ exists if $\tilde{s}(m)$ exists.[10] As a particular case, we thus have that $s(u)$ exists if and only if $\varphi(\beta)$ exists; that is, the thermodynamic entropy exists if and only if the thermodynamic free energy exists [54,59].

Our treatment of large deviation principles avoided, for the most part, any proofs of the thermodynamic limit. We either assumed the existence of large deviation principles in order to work out their consequences, or we established directly those large deviation principles by deriving their rate functions by contraction of other rate functions that are known to exist (e.g., the relative entropy). In the case of mean-field and long-range systems, for example, we are often led to prove that $s(u)$ exists simply by calculating this rate function with the general maximum entropy principle involving the macrostate entropy $\tilde{s}(m)$ and the energy representation function $\tilde{h}(m)$. Short-range models are more difficult to treat because, as mentioned, they necessitate the use of the empirical process, the Level-3 macrostate. The calculation of $s(u)$ for these models thus involves the contraction of an infinite-dimensional rate function—the relative entropy of the empirical process—down to the mean energy h_n.

The existence of $s(u)$ may be established in a more general way by proving the existence of the limit defining this quantity for certain classes of Hamiltonians. This method was initiated by Ruelle [53,103] and Lanford [11] (see also Griffiths [104]), who proved that the limit defining $s(u)$ does exist for interactions that are *stable*, in the sense that they do not lead to a collapse of the particles into a low energy state, and are *tempered*, in the sense that they decay sufficiently quickly at large distances so as to limit surface or boundary effects in the thermodynamic limit. The method of proof relies on the notion of super-additivity, and establishes as an added result that $s(u)$ is concave, which implies, in turn, that the microcanonical and canonical ensembles are equivalent [53,105,106]. This result can be extended to macrostates other than h_n to conclude that macrostates satisfying similar conditions of stability and temperedness have concave macrostate entropies. For an introduction to these results, the reader is referred to the work of Lanford [11], who treats a simpler class of short-range interactions and macrostates known as *finite-range*.

Proofs of the existence of the canonical free energy $\varphi(\beta)$ have also been given, notably by van Hove [107,108], Ruelle [53,109], Fisher [110], and Griffiths [111,112]. The class of short-range interactions considered in this case is essentially the same as the one mentioned before, namely, stable are tempered. One interesting aspect of these proofs is that they also establish the equivalence of the microcanonical and canonical ensembles, among other ensembles. Therefore, one condition

[10] Except possibly at boundary points; see Example 4.8.

that appears to be necessary (although not sufficient) for having nonconcave entropies and nonequivalent ensembles in the thermodynamic limit is for the interaction in a system to be long-range. Gravitating particles, unscreened plasmas, and vortex models of 2D turbulence are examples of such long-range interaction systems; see [80] for others.

6. Large deviations in nonequilibrium statistical mechanics

We turn in this section to the study of large deviations arising in physical systems that dynamically evolve in time or that are maintained in out-of-equilibrium steady states by an external forcing. The methodology that will be followed for studying these nonequilibrium systems is more or less the one that we followed in the previous section. All that changes is the type of systems studied, and the fact that in nonequilibrium statistical mechanics the object of focus is most often not the Hamiltonian or the constraints imposed on a system, but the stochastic process (Markov chain, Langevin equation, master equation, etc.) used to model that system.

The dynamical nature of nonequilibrium systems requires, of course, that we include time in the large deviation analysis, possibly as the extensive parameter controlling a large deviation principle (as is the case for the number of particles). Conceptually, this is a minor adjustment to take into account. A more fundamental difference between equilibrium and nonequilibrium is that there is no concept of statistical–mechanical ensemble for nonequilibrium systems, even those driven in out-of-equilibrium steady states [113]. That is to say, when a system is out of equilibrium, we do not know in general what the underlying probability distribution of its states is (if such a distribution indeed exists). To find it, we must define the system precisely, calculate the probability distribution of its states from first principles, and proceed from there to derive large deviation principles for observables that are functions of the system's state. There is no general principle whereby one can calculate the distribution of the system's states from the sole knowledge of the system's invariants or external constraints imposed on the system. Such a general principle is precisely what a statistical–mechanical ensemble is, and what is missing from the theory of nonequilibrium systems.

In spite of this, it is possible to formulate a number of general and interesting results for nonequilibrium systems, especially when these are modeled as Markov processes. The aim of this section is to give an overview of these results in the style of the previous section, with an emphasis on large deviations. We will see with these results that it is often possible to characterize the most probable states (trajectories or paths) of a nonequilibrium system as the minima of a rate function, and that these minima give rise to a variational principle that generalizes the maximum entropy or minimum free energy principles. The knowledge of this rate function also provides, as in the case of equilibrium systems, a complete description of the fluctuations of the system considered.

6.1. Noise-perturbed dynamical systems

The first class of large deviation results that we study concerns the fluctuations of deterministic dynamical systems perturbed by noise. The idea here is to consider a differential equation, which determines the motion of a dynamical system in time, and to perturb it with a Gaussian white noise of zero mean and small intensity (variance or power) $\epsilon \geq 0$. In the presence of noise ($\epsilon \neq 0$), the system's motion is random, but for a small noise, that random motion is expected to stay close to the unperturbed dynamics, and should converge, in the zero-noise limit $\epsilon \to 0$, to the deterministic motion determined by the unperturbed differential equation. In terms of probabilities, this means that the probability distribution of the system's trajectories or *paths* should concentrate, as $\epsilon \to 0$, around the deterministic path of the unperturbed system. This concentration effect is akin to a Law of Large Numbers, so in the spirit of large deviation theory it is natural to inquire about the scaling of that concentration with ϵ; that is, how is the probability decaying around its maximum as $\epsilon \to 0$? The answer, as might be expected, is that the decay has the form of a large deviation principle.

6.1.1. Formulation of the large deviation principle

The study of large deviations of random paths gives rise to a mathematical difficulty that we encountered before when we treated the continuous version of Sanov's Theorem: a trajectory is a function, which means that the probabilities that we must handle are probabilities over a function space. A rigorous mathematical treatment of large deviations exists in this setting (see, e.g., [6] or Chap. 5 of [13]), but for simplicity, and to give a clearer presentation of the ideas involved, we will follow the previous sections and deal with probabilities of random functions at a heuristic level. To simplify the presentation, we will also start our study with a simple, one-dimensional noise-perturbed system described by the following stochastic differential equation:

$$\dot{X}_\epsilon(t) = b(X_\epsilon) + \sqrt{\epsilon}\eta(t), \quad X_\epsilon(0) = x_0. \tag{206}$$

In this expression, $X_\epsilon \in \mathbb{R}$, b is a real function (sufficiently well-behaved), and $\eta(t)$ is a Gaussian white noise process characterized by its mean $\langle \eta(t) \rangle = 0$ and correlation function $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$. The real constant $\epsilon \geq 0$ is the large deviation parameter controlling the intensity of the noise. For the mathematically minded, Eq. (206) should properly be interpreted as the Itô form

$$\mathrm{d}X_\epsilon(t) = b(X_\epsilon)\,\mathrm{d}t + \sqrt{\epsilon}\,\mathrm{d}W(t) \tag{207}$$

involving the Brownian or Wiener motion $W(t)$ [114]. What we consider in (206) is the naive yet standard version of Eq. (207), obtained by heuristically viewing Gaussian white noise as the time-derivative of Brownian motion.

The probability that we wish to investigate is the probability of a trajectory or path of the random dynamical system described by Eq. (206), that is, the probability of a given realization $\{x(t)\}_{t=0}^{\tau}$ of that equation, extending from an initial time $t = 0$ to some time $\tau > 0$. Of course, one cannot speak of the probability of a *single* trajectory, but only of a *set* of trajectories, and this is where the difficulty mentioned above comes in. To avoid it, one may consider the probability that the system's trajectory lies in some cylinder or "tube" enclosing a given trajectory $\{x(t)\}_{t=0}^{\tau}$, or any other finite set of trajectories.

This way of making sense of probabilities in trajectory space will not be followed here; instead, we will assume at a heuristic level that there is a probability density $P[x]$ over the different paths $\{x(t)\}_{t=0}^{\tau}$ of the system. Following the physics literature, we denote this density with square brackets to emphasize that it is a functional of the whole function $x(t)$. With this notation, we then write a large deviation principle for the random paths as $P_\epsilon[x] \asymp \mathrm{e}^{-a_\epsilon J[x]}$ to mean that $P[x]$ decays exponentially with *speed* a_ϵ in such a way that $a_\epsilon \to \infty$ as $\epsilon \to 0$.[11] The rate function $J[x]$ is a functional of the paths. To be rigorous, we should write this large deviation principle as

$$P\left(\sup_{0\le t\le\tau} |X_\epsilon(t) - x(t)| < \delta\right) \asymp \mathrm{e}^{-a_\epsilon J[x]}, \quad \epsilon \to 0 \tag{208}$$

where δ is any small, positive constant. In this form, the rate function is then obtained by the limit

$$\lim_{\epsilon\to 0} \frac{1}{a_\epsilon} \ln P\left(\sup_{0\le t\le\tau} |X_\epsilon(t) - x(t)| < \delta\right) = -J[x]. \tag{209}$$

The notation $P_\epsilon[x] \asymp \mathrm{e}^{-a_\epsilon J[x]}$ is obviously more economical.

The large deviation principle associated with the specific system described by Eq. (206) was derived rigorously by Freidlin and Wentzell [6,115], and through formal path integral methods by Graham and Tél [116,117], as well as by Dykman and Krivoglaz [118] among others.[12] The result, in the path density notation, is

$$P_\epsilon[x] \asymp \mathrm{e}^{-J[x]/\epsilon}, \qquad J[x] = \frac{1}{2}\int_0^\tau [\dot{x} - b(x)]^2 \, \mathrm{d}t, \tag{210}$$

where $x(t)$ is any (absolutely) continuous path[13] satisfying the initial condition $x(0) = x_0$. The rate function $J[x]$ is sometimes called the *action functional* [6] or *entropy* of the path [5].[14] Notice that $J[x]$ is positive and has a unique zero corresponding to the deterministic path $x^*(t)$ satisfying the unperturbed equation $\dot{x}^* = b(x^*)$. Therefore, $\|X_\epsilon(t) - x^*(t)\| \to 0$ in probability as $\epsilon \to 0$. The quadratic form of $J[x]$ stems from the Gaussian nature of the noise $\eta(t)$. For other types of noise, in particular correlated (colored) noise, $P_\epsilon[x]$ may still have a large deviation form, but with a rate function which is not quadratic or local in time. More details on these correlated large deviation principles can be found in [126–130] (see also [131,132]).

6.1.2. Proofs of the large deviation principle

The large deviation result displayed in (210) can be proved in many different ways. The simplest, perhaps, is to approximate the trajectories $\{x(t)\}_{t=0}^{\tau}$ in the spirit of path integral techniques by discrete-time trajectories $\{x_i\}_{i=1}^{n}$ involving n points equally spaced between $t = 0$ and $t = \tau$ at interval Δt; see Fig. 18. This discretization or "time-slicing" procedure has the effect of transforming the Markov stochastic process of Eq. (206) into a Markov chain, for which it is relatively easy to compute the probability density $p(x_1, x_2, \ldots, x_n)$ given the properties of the noise $\eta(t)$. The probability density $P_\epsilon[x]$ is then obtained from $p(x_1, x_2, \ldots, x_n)$ by taking the double limit $n \to \infty$, $\Delta t \to 0$. For more details, see Chap. 2 of [133].

More interesting from the point of view of large deviation theory is the fact that $P_\epsilon[x]$ can be obtained from the Gärtner–Ellis Theorem by calculating the functional Legendre–Fenchel transform of the scaled cumulant generating functional of $X_\epsilon(t)$, defined as

$$\lambda[k] = \lim_{\epsilon\to 0} \epsilon \ln \left\langle \mathrm{e}^{k\cdot X_\epsilon/\epsilon} \right\rangle, \tag{211}$$

---

11 The term "speed" in this context has, of course, nothing to do with the time derivative of the position. The term is used in large deviation theory because a_ϵ determines how quickly P_ϵ decays to zero with ϵ; see Appendices B and D.

12 Onsager and Machlup [119] derived this result for linear equations as far back as 1953 (see also [120–122]).

13 That $x(t)$ should be a continuous path does not contradict the fact that the random paths of stochastic equations driven by Gaussian white noise are nondifferentiable with probability 1. Remember that $P_\epsilon[x]$ is a formal notation for the probability that a given random path lies inside an infinitesimal tube whose center follows a given smooth path $x(t)$. Thus, what we are interested in is the probability that a random path *is close* to some smooth path $x(t)$, not the probability that a random path *follows exactly* some smooth path $x(t)$ [123].

14 The form of $J[x]$ presented here is the form obtained in the Itô interpretation of the stochastic equation. In the Stratonovich interpretation, there is an additional term involving the components of the vector $b(x)$, which vanishes in the zero-noise limit. The difference amounts to a Jacobian term in path integrals involving $P_\epsilon[x]$ [124,125].

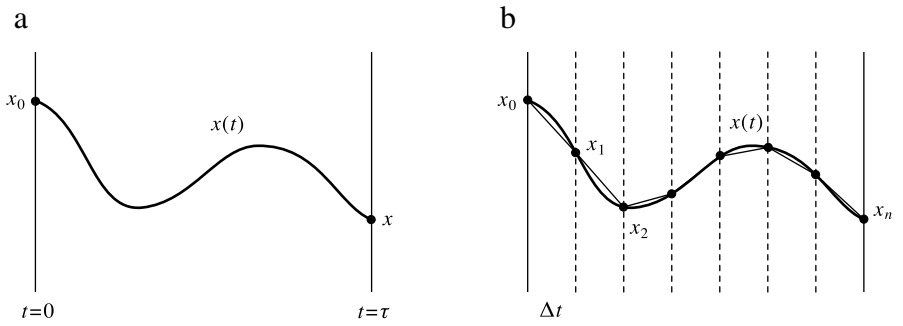

Fig. 18. (a) Trajectory starting at x_0 at time $t = 0$ and ending at x at time $t = \tau$. (b) Time-discretization of the continuous-time trajectory.

where

$$\left\langle e^{k\cdot X_\epsilon/\epsilon}\right\rangle = \int \mathcal{D}[x]\, P_\epsilon[x]\, \exp\left(\frac{1}{\epsilon}\int_0^\tau k(t)x(t)\,dt\right). \tag{212}$$

The path integral defining the expectation value above can be solved exactly for Gaussian white noise to obtain an explicit expression for $\lambda[k]$ [133,134]. In this case, the measure $\mathcal{D}[x]$ on the space of trajectories is the Wiener measure. Alternatively, $\lambda[k]$ can be obtained in the context of the Donsker–Varadhan theory from the generator of the Markov process defined by Eq. (206) (see, e.g., [5,17,28]). In both cases, $J[x]$ is then expressed as

$$J[x] = \inf_k\{k\cdot x - \lambda[k]\}. \tag{213}$$

The same rate function can also be derived from the large deviation point of view by applying the contraction principle to the rate function governing the fluctuations of the scaled noise $\eta_\epsilon(t) = \sqrt{\epsilon}\eta(t)$ [13]. We show in the next example how to obtain the latter rate function from the Gärtner–Ellis Theorem. The large deviation result obtained for $\eta_\epsilon(t)$ is known in the mathematical literature as *Schilder's Theorem* [13,135]. The calculation of $J[x]$ based on the contraction principle follows the presentation of that theorem.

Example 6.1 (*Schilder's Theorem*). The properties of a Gaussian white noise $\eta(t)$ with $\langle\eta(t)\rangle = 0$ for all t and $\left\langle\eta(t)\eta(t')\right\rangle = \delta(t-t')$ are completely determined by its characteristic function:

$$G_\eta[k] = \left\langle e^{ik\cdot\eta}\right\rangle = \int \mathcal{D}[\eta]\, P[\eta]\, e^{i\int k(t)\eta(t)\,dt} = \exp\left(-\frac{1}{2}\int k(t)^2\,dt\right). \tag{214}$$

From this form of $G_\eta[k]$, the scaled cumulant generating functional of the scaled noise $\eta_\epsilon(t) = \sqrt{\epsilon}\eta(t)$ is easily found to be

$$\lambda[k] = \lim_{\epsilon\to 0}\,\epsilon\ln\left\langle e^{k\cdot\eta_\epsilon/\epsilon}\right\rangle = \frac{1}{2}\int k(t)^2\,dt. \tag{215}$$

This result is the functional analog of the log-generating function of a Gaussian random variable with zero mean and unit variance; see Example 3.1. As in that example, $\lambda[k]$ is differentiable, but now in the functional sense. By applying the Gärtner–Ellis Theorem, we then conclude that $\eta_\epsilon(t)$ satisfies a large deviation principle in the limit $\epsilon \to 0$ with a rate function given by the Legendre–Fenchel transform of $\lambda[k]$:

$$P_\epsilon[\phi] \asymp e^{-I[\phi]/\epsilon}, \qquad I[\phi] = \sup_k\{k\cdot\phi - \lambda[k]\}. \tag{216}$$

In this expression, $\phi(t)$ is a given trajectory or realization of the noise $\eta_\epsilon(t)$ starting at $\phi(0) = 0$. As in the non-functional case, we can use the differentiability of $\lambda[k]$ to reduce the Legendre–Fenchel transform above to a Legendre transform, given by

$$I[\phi] = k_\phi\cdot\phi - \lambda[k_\phi] = \int k_\phi(t)\phi(t)\,dt - \lambda[k_\phi], \tag{217}$$

where k_ϕ is the functional root of

$$\frac{\delta\lambda[k]}{\delta k(t)} = \phi(t). \tag{218}$$

In the present case, $k_\phi(t) = \phi(t)$, so that

$$I[\phi] = \phi\cdot\phi - \lambda[\phi] = \frac{1}{2}\int \phi(t)^2\,dt. \tag{219}$$

The expression of this rate function has an obvious similarity with the rate function of the sample mean of IID Gaussian random variables discussed in Example 3.1.

We are now in a position to derive the rate function $J[x]$ of $X_\epsilon(t)$, shown in Eq. (210), from the rate function $I[\phi]$ of the scaled Gaussian noise $\eta_\epsilon(t)$. The main point to observe is that $X_\epsilon(t)$ is a contraction of $\eta_\epsilon(t)$, in the sense of the contraction principle (Sec. 5.6 of [13]). This is obvious if we note that the stochastic differential Eq. (206) has for solution

$$x(t) = x_0 + \int_0^t b(x(s))\,\mathrm{d}s + \int_0^t \phi(s)\,\mathrm{d}s, \tag{220}$$

where $\phi(t)$ is, as before, a realization of $\eta_\epsilon(t)$. Let us denote this solution by the functional $f[\phi] = x$. Using the contraction principle, we then write

$$J[x] = \inf_{\phi: f[\phi]=x} I[\phi] = \inf_{\phi:\phi=\dot{x}-b(x)} I[\phi] = I[\dot{x} - b(x)], \tag{221}$$

which is exactly the result of Eq. (210) given the expression of $I[\phi]$ found in Eq. (219).

For future use, we re-write $J[x]$ as

$$J[x] = \int L(\dot{x}, x)\,\mathrm{d}t, \quad L(\dot{x}, x) = \frac{1}{2}\left[\dot{x}(t) - b(x(t))\right]^2. \tag{222}$$

The function $L(\dot{x}, x)$ is called the *Lagrangian* of the stochastic process $X_\epsilon(t)$. The next example gives the expression of $L(\dot{x}, x)$ and $J[x]$ for a more general class of stochastic differential equations involving a state-dependent diffusion term. This class is the one considered by Freidlin and Wentzell [6] (see also Sec. 5.6 of [13] and [136,137]).

Example 6.2 (*General Stochastic Differential Equation*). Let $X_\epsilon(t)$ be a flow in $\mathbb{R}^d$, $d \geq 1$, governed by the following (Itô) stochastic differential equation:

$$\mathrm{d}X_\epsilon(t) = b(X_\epsilon)\,\mathrm{d}t + \sqrt{\epsilon}\sigma(X_\epsilon)\,\mathrm{d}W(t), \quad X_\epsilon(0) = x_0, \tag{223}$$

where b is some function mapping $\mathbb{R}^d$ to itself, σ is a square, positive-definite matrix assumed to be nonsingular, and $W(t)$ is the usual Brownian motion. For this system, Freidlin and Wentzell [6] proved that $P_\epsilon[x] \asymp \mathrm{e}^{-J[x]/\epsilon}$ as $\epsilon \to 0$ with rate function

$$J[x] = \frac{1}{2}\int_0^\tau \left[\dot{x}(t) - b(x(t))\right]^T A^{-1}\left[\dot{x}(t) - b\left(x(t)\right)\right]\,\mathrm{d}t, \tag{224}$$

where $A = \sigma\sigma^T$ is the so-called *diffusion matrix*. Two technical conditions complete this large deviation result. First, to ensure the existence and uniqueness of a solution for Eq. (223), the *drift* vector $b(x)$ must be Lipschitz continuous. Second, the realizations $x(t)$ of $X_\epsilon(t)$ for which the rate function $J[x]$ exists must verify the initial condition $x(0) = x_0$, in addition to having square integrable time-derivatives.

6.1.3. Large deviations for derived quantities

Once a large deviation principle has been proved for the path density $P_\epsilon[x]$, the way becomes wide open for deriving large deviation principles for all sorts of probabilities using the contraction principle. Of particular interest is the probability density

$$P_\epsilon(x, \tau|x_0) = \int_{x(0)=x_0}^{x(\tau)=x} \mathcal{D}[x]\,P_\epsilon[x] \tag{225}$$

that the process $X_\epsilon(t)$ reaches a point x at time $t = \tau$ given that it started at a point x_0 at time $t = 0$; see Fig. 18(a). Assuming that $P_\epsilon[x]$ satisfies a large deviation principle with rate function $J[x]$, it directly follows from the contraction principle that $P(x, \tau|x_0)$ also satisfies a large deviation principle of the form

$$P_\epsilon(x, \tau|x_0) \asymp \mathrm{e}^{-V(x,\tau|x_0)/\epsilon} \tag{226}$$

with rate function given by

$$V(x, \tau|x_0) = \inf_{x(t):x(0)=x_0, x(\tau)=x} J[x]. \tag{227}$$

This rate function is also called the *quasi-potential*.

The large deviation approximation of $P(x, \tau|x_0)$ is often referred to as a *WKB approximation*, following Wentzel,[15] Kramers and Brillouin, who developed a similar approximation in the context of quantum mechanics and differential

[15] Wentzel the physicist, not to be confused with Wentzell, the mathematician mentioned earlier.

equations [133].[16] The meaning of this approximation follows exactly the interpretation of the contraction principle, in that the dominant contribution to the probability of a fluctuation—here the observation of $x(\tau) = x$ starting from $x(0) = x_0$—is the probability of the most probable path leading to that fluctuation. This most probable or *optimal* path $x^*(t)$, which is the path solving the variational problem (227), can be determined by solving the *Euler–Lagrange equation*

$$\left.\frac{\delta J[x]}{\delta x(t)}\right|_{x^*(t)} = 0, \quad x(0) = x_0, x(\tau) = x, \tag{228}$$

which has the well-known form

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = 0, \quad x(0) = x_0, x(\tau) = x \tag{229}$$

in terms of the Lagrangian $L(\dot{x}, x)$. The optimal path[17] can also be interpreted, by analogy with classical mechanics, as the solution of the *Hamilton–Jacobi equations*,

$$\dot{x} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial x}, \quad H(p, x) = \dot{x}p - L(\dot{x}, x), \tag{230}$$

which involve the *Hamiltonian* $H(p, x)$ conjugated, in the Legendre–Fenchel sense, to the Lagrangian $L(\dot{x}, x)$. These observations are put to use in the next example to determine the WKB approximation of the stationary distribution of a simple but important Markov process.

Example 6.3 (*Stationary Distribution of the Ornstein–Uhlenbeck Process*). Consider the linear equation

$$\dot{X}_\epsilon(t) = -\gamma X_\epsilon(t) + \sqrt{\epsilon}\eta(t), \tag{231}$$

and let $P_\epsilon(x)$ denote the stationary probability density of this process which solves the time-independent Fokker–Planck equation

$$\gamma\frac{\partial}{\partial x}[xP_\epsilon(x)] + \frac{\epsilon}{2}\frac{\partial^2 P_\epsilon(x)}{\partial x^2} = 0, \quad \epsilon > 0. \tag{232}$$

We know that in the weak-noise limit $\epsilon \to 0$, $P_\epsilon(x)$ obeys the WKB form $P_\epsilon(x) \asymp \mathrm{e}^{-V(x)/\epsilon}$. To find the expression of the quasi-potential $V(x)$, we follow Onsager and Machlup [119] and solve the Euler–Lagrange Eq. (229) with the terminal conditions $x(-\infty) = 0$ and $x(\tau) = x$. The solution is $x^*(t) = x\mathrm{e}^{\gamma(t-\tau)}$. Inserting this back into $J[x]$ yields $V(x) = J[x^*] = \gamma x^2$. This result can also be obtained in a more direct way by expressing the force $b(x)$ as the derivative of a potential, i.e., $b(x) = -U'(x)$ with $U(x) = \gamma x^2/2$. In this case, it is known that $V(x) = 2U(x)$ (see Sec. 4.3 of [6]).

The previous example can be generalized as follows. If the zero-noise limit of the stochastic differential equation given by Eq. (223) has a unique attracting fixed point x_s, then the quasi-potential $V(x)$ associated with the stationary distribution of that equation is obtained in general by

$$V(x) = \inf_{x(t):x(t_1)=x_s, x(t_2)=x} J[x], \tag{233}$$

where t_1 and t_2 are, respectively, the starting and ending times of the trajectory $x(t)$ (see Sec. 4.2 of [6]). In this variational formula, the two endpoints of the interval $[t_1, t_2]$ are not fixed, which means that they are variables of the variational problem. In many cases, however, it is possible to solve the infimum by letting $t_1 \to -\infty$ and by fixing t_2, as we have done in the previous example. For a discussion of this procedure, see Sec. 4.3 of [6]. Bertini et al. [140] give an interesting derivation of the above formula by studying the rate function associated with the time-reverse image of the trajectories $x(t)$ determined by Eq. (223).

The example that follows presents another important problem for which the WKB approximation is useful, namely, that of estimating the average time of escape from an attractor.

Example 6.4 (*Exit Time from an Attractor*). An attracting fixed point x_s of a dynamical system does not remain attracting in the presence of noise: for $\epsilon \neq 0$, there is a non-zero probability, however small, that a trajectory starting in the vicinity of x_s will be "pushed" by the noise out of some bounded region D enclosing x_s; see Fig. 19(a). The probability that such an escape occurs is very small, and decreases to zero as $\epsilon \to 0$. Consequently, the random time needed for the system to reach the boundary ∂D of D, which is defined as

$$\tau_\epsilon = \inf\{t : x(t) \in \partial D\}, \tag{234}$$

should increase in some probabilistic sense as $\epsilon \to 0$. This time τ_ϵ is called the *escape-* or *exit-time* from D.

16 The WKB approximation is also referred to as the eikonal approximation.

17 The optimal path is also called the *maximum likelihood path* between two fixed endpoints, the *phenomenological path* [134] or the *instanton* [138,139].

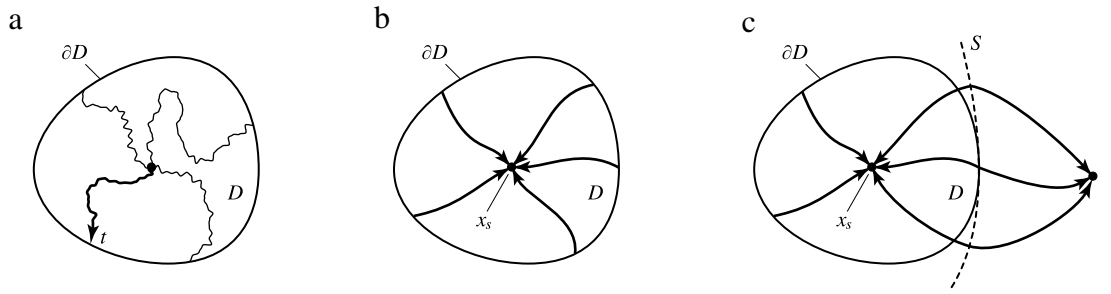

Fig. 19. (a) Random paths reaching the boundary ∂D of a region D in a time t. (b) Stable fixed point x_s located in D. (c) Separatrix S delimiting the basins of two attracting fixed points. Some paths on the boundary ∂D lie on the separatrix and are attracted by one or the other fixed point depending on whether they start on the left or right of S.

The calculation of τ_ϵ is a classical problem in nonequilibrium statistical mechanics (see, e.g., [114,141]), and was solved on the mathematical front by Freidlin and Wentzell [6], who treated it in the context of the general stochastic Eq. (223). Their main result assumes that the unperturbed dynamics associated with Eq. (223) has a single attracting fixed point x_s located in D, and that all the points on ∂D are attracted to x_s, so that the case of a boundary ∂D lying on a separatrix is excluded; see Fig. 19(c). Under these assumptions, the following limit then holds:

$$\lim_{\epsilon\to 0} P\left(\mathrm{e}^{(V^*-\delta)/\epsilon} < \tau_\epsilon < \mathrm{e}^{(V^*+\delta)/\epsilon}\right) = 1, \tag{235}$$

where

$$V^* = \inf_{x\in\partial D}\inf_{t\geq 0} V(x,t|x_s) \tag{236}$$

and δ is any small positive constant. Moreover,

$$\lim_{\epsilon\to 0} \epsilon \ln\langle\tau_\epsilon\rangle = V^*. \tag{237}$$

The first limit shown in (235) states that the most probable escape time scales as $\tau_\epsilon \asymp \mathrm{e}^{V^*/\epsilon}$ as $\epsilon \to 0$. From this concentration result, the second limit follows.

The complete proof of these results is quite involved; see Sec. 4.2 of [6] or Sec. 5.7 of [13]. However, there is a simple argument due to Kautz [142] that can be used to understand the second result stating that $\langle\tau_\epsilon\rangle \asymp \mathrm{e}^{V^*/\epsilon}$. The essential observation is that the average time $\langle\tau_\epsilon\rangle$ of escape is roughly proportional to the escape rate r_ϵ, which is itself proportional to the probability $P_\epsilon^{\mathrm{esc}}$ of escaping D. Thus $\langle\tau_\epsilon\rangle \propto 1/P_\epsilon^{\mathrm{esc}}$, where

$$P_\epsilon^{\mathrm{esc}} = \int_{\partial D} \mathrm{d}x \int_0^\infty \mathrm{d}t\, P_\epsilon(x,t|x_s). \tag{238}$$

Applying Laplace's approximation to this integral yields $P_\epsilon^{\mathrm{esc}} \asymp \mathrm{e}^{-V^*/\epsilon}$ with V^* given by Eq. (236), and therefore $\langle\tau_\epsilon\rangle \asymp \mathrm{e}^{V^*/\epsilon}$, as Eq. (237). The result of Freidlin and Wentzell is more precise, since it provides a Law of Large Numbers for τ_ϵ, not just an estimate for $\langle\tau_\epsilon\rangle$.

The two previous examples are representative of the way large deviation techniques can be applied for calculating stationary probability densities $P_\epsilon(x)$ and fixed-time probability densities $P_\epsilon(x,t|x_0)$, as well as exit times and exit points. The second example, in particular, can be used to derive a whole class of Arrhenius-type results of the form $\tau_\epsilon \asymp \mathrm{e}^{V^*/\epsilon}$ for diffusion- or thermally-induced escape processes, including Kramers's classical result for the escape time of a Brownian particle trapped in a potential [114,141]. In the specific context of systems perturbed by thermal noise, the variational principle expressed by Eq. (236) is often referred to as the *principle of minimum available energy*, since V^* can be shown to be proportional to the *activation energy*, that is, the minimum energy required to induce the escape [142,143]. An application of this principle for Josephson junctions is discussed by Kautz [142,143].

For practical applications, it is important to note that Freidlin–Wentzell can be generalized to nonlinear systems having multiple attractors A_i, $i = 1, 2, \ldots$. For these, the escape time $\tau_{\epsilon,i}$ from a domain D_i of attraction of A_i is estimated as

$$\tau_{\epsilon,i} \asymp \mathrm{e}^{V_i^*/\epsilon}, \qquad V_i^* = \inf_{x\in\partial D_i}\inf_{t\geq 0} V(x,t|x_i), \tag{239}$$

where x_i is an initial point chosen inside A_i. For more than one attractor, the quasi-potential $V(x)$ characterizing the stationary distribution $P_\epsilon(x)$ over the whole state-space is also estimated as

$$V(x) = \inf_i V_i(x), \tag{240}$$

where $V_i(x)$ is the quasi-potential of $P_\epsilon(x)$ restricted to the attractor A_i, i.e., the quasi-potential of a stationary probability density obtained by initiating paths inside A_i [6,137]. One important characteristic of many-attractor systems is that $V(x)$ is in general nonconvex, in addition to being nondifferentiable at points x lying on a separatrix (see, e.g., [6,117,144,145]). Mathematically, this arises because the infimum of Eq. (240) switches abruptly on a separatrix from one (generally smooth) quasi-potential $V_i(x)$ to another. A similar switching phenomenon was observed in the simpler context of sample means in Example 4.7.

6.1.4. Experimental observations of large deviations

Optimal paths and exit times are not just mathematical constructs—they can be, and have been, observed experimentally. The reader is referred to the extensive work of Dykman, Luchinsky, McClintock and collaborators [146–149] for a discussion of many properties of optimal paths observed in analog electronic circuits, including symmetry properties of these paths with respect to time inversion [148], and their singular patterns near coexisting attractors [147]. All of these topics are reviewed in the excellent survey paper [150], which also discusses experimental measurements of exit times.

6.2. Phenomenological models of fluctuations

Equilibrium statistical mechanics is a static theory of thermodynamics fluctuations: it provides a basis for calculating the probability of fluctuations of given macrostates, but says nothing about how these fluctuations arise in time. To describe the dynamics of these fluctuations, we must consider dynamical models of many-particle systems, and infer from these models the dynamical—and possibly stochastic—equations that govern the evolution of the macrostates that we are interested in studying. Such a microstate-to-macrostate reduction of the dynamics of a many-body system is, as is well known, very difficult (if not impossible) to work out in practice, and so more modest approaches to this problem are usually sought. The most basic is the phenomenological approach, which consists in assuming that the time evolution of a macrostate, say M_n, follows a given stochastic dynamics of the form

$$\dot{M}_n(t) = b(M_n) + \xi_n(t), \tag{241}$$

where $b(M_n)$ is a force field, and $\xi_n(t)$ is a noise term that models the fluctuations of $M_n(t)$. The term "phenomenological" indicates that the dynamics of M_n is postulated on the basis of a number of physical and mathematical principles, rather than being derived directly from an n-particle dynamics. Among these principles, we note the following:

1. The unperturbed dynamics $\dot{m} = b(m)$ should represent the macroscopic (most probable) evolution of $M_n(t)$.
2. The intensity of the noise $\xi_n(t)$ should vanish as $n \to \infty$ to reflect the fact that the fluctuations of M_n vanish in the thermodynamic limit.
3. Given that the fluctuations of M_n arise from the cumulative and (we assume) short-time correlated interactions of n particles, the noise $\xi_n(t)$ should be chosen to be a Gaussian white noise with zero mean.
4. The stationary probability distribution associated with Eq. (241) should match the equilibrium probability distribution of M_n determined by the equilibrium ensemble used to describe the n-particle system in equilibrium.

The second and third points imply that $\xi_n(t)$ should satisfy $\langle \xi_n(t) \rangle = 0$ and $\langle \xi_n(t)\xi_n(t') \rangle = b_n \delta(t - t')$, with $b_n \to 0$ as $n \to \infty$. The precise dependence of b_n on n is determined self-consistently, following the last point, by matching the large-n form of the stationary distribution of Eq. (241) with the large deviation form of the equilibrium (ensemble) probability distribution of M_n. This is explained in the next example.

Example 6.5 (*Equilibrium Fluctuations*). Consider a macrostate M_n satisfying an equilibrium large deviation principle of the form

$$p(M_n = m) \asymp \mathrm{e}^{-a_n I(m)}, \tag{242}$$

where $I(m)$ is any of the rate functions arising in the microcanonical or canonical ensemble, and a_n is the speed of the large deviation principle.[18] If $I(m)$ has a unique global minimum, then we know from Example 6.3 that the stochastic dynamics

$$\dot{M}_n(t) = -\frac{1}{2} I'(M_n) + \xi_n(t) \tag{243}$$

has a stationary density given by $p(M_n = m) \asymp \mathrm{e}^{-I(m)/b_n}$ for small b_n. By matching this asymptotic with the large deviation principle of (242), we then obtain $b_n = a_n^{-1}$. Thus, if the speed a_n of the large deviation principle is the number n of particles, as is typically the case, then $b_n = n^{-1}$. Near phase transitions, the speed of a large deviation principle may change (see Examples 5.4 and 5.6), and this should be reflected in b_n.

Models of fluctuation dynamics based on the phenomenological model of Eq. (241) or the more specific equation found in (243), based on the rate function $I(m)$, are used to answer a variety of questions, such as:

[18] Recall that in the microcanonical ensemble, $I(m)$ is interpreted as an entropy function, whereas in the canonical ensemble, $I(m)$ is interpreted as a free energy function; see Section 5.

- What is the most probable *fluctuation path* $\{m(t)\}_{t=0}^{\tau}$ connecting over a time τ the equilibrium or stationary value m^* of M_n to some other value $m \neq m^*$?
- What is the most probable *decay path* connecting the nonequilibrium state $M_n(0) = m$ to the equilibrium state $M_n(\tau) = m^*$?
- Is there a relationship between a given fluctuation path and its corresponding decay path? For instance, are decay paths the time-reverse image of fluctuation paths?
- What is the typical or expected time of return to equilibrium? That is, what is the typical or expected time τ for which $M_n(\tau) = m^*$ given that $M_n(0) = m \neq m^*$?
- If $I(m)$ has local minima in addition to global minima, what is the typical time of decay from a local minimum to a global minimum? In other words, what is the typical decay time from a metastable state?

It should be clear, from our experience of the last subsection, that all of these questions can be answered within the framework of the Freidlin–Wentzell theory of differential equations perturbed by noise. In the thermodynamic limit, fluctuation and decay paths are optimal paths, and can be determined as such by the variational principle of Eq. (227), the Lagrangian equation (229) or its Hamiltonian counterpart, Eq. (230). These equations also hold the key for comparing the properties of decay and fluctuation paths. As for the calculation of decay times from nonequilibrium states, including metastable states, it closely follows the calculation of exit times that we have discussed in the previous subsection (see also [151,152]).

Other results about nonequilibrium fluctuations can be translated in much the same way within the framework of the Freidlin–Wentzell theory. The minimum dissipation principle of Onsager and Machlup [119], for example, which states that the fluctuation and decay paths of equilibrium systems minimize some dissipation function, can be re-interpreted in terms of the variational principle of Eq. (227), which determines the optimal paths of noise-perturbed systems. The next example is intended to clarify this point by explicitly translating the theory of Onsager and Machlup into the language of large deviations. For simplicity, we consider the dynamics of a single, one-dimensional macrostate.

Example 6.6 (*Linear Fluctuation Theory*). The linear theory of equilibrium fluctuations proposed by Onsager and Machlup [119] (see also [120,121]) is based on what is essentially a linear version of Eq. (243), obtained by assuming that $I(m)$ has a unique and locally quadratic minimum at $m^* = 0$, and that M_n fluctuates close to this equilibrium value. By approximating $I(m)$ to second order around its minimum

$$I(m) \approx am^2, \quad a = \frac{I''(0)}{2} > 0, \tag{244}$$

we thus write

$$\dot{M}_n(t) = -aM_n(t) + \xi_n(t). \tag{245}$$

The scaling of $\xi_n(t)$ with n is not specified by Onsager and Machlup [119], but it is obvious from their analysis that, if Eq. (245) is to have a macroscopic limit, then the variance of the noise should scale inversely with the number n of particles, as explained in Example 6.5. In this case, we can write the path probability density $P_n[m]$ of $M_n(t)$ as

$$P_n[m] \asymp \mathrm{e}^{-nJ[m]}, \quad J[m] = \frac{1}{2}\int_0^\tau [\dot{m}(t) + am(t)]^2 \,\mathrm{d}t \tag{246}$$

in the limit of large n, which is more or less what Onsager and Machlup obtain in [119]. The Lagrangian of this rate function can be re-written as

$$L(\dot{m}, m) = \Phi(\dot{m}) + \Psi(m) + \frac{\dot{I}(m)}{2} \tag{247}$$

by defining what Onsager and Machlup call the *dissipation functions* $\Phi(\dot{m}) = \dot{m}^2/2$ and $\Psi(m) = a^2m^2/2$. With this form of $L(\dot{m}, m)$, a fluctuation path is then characterized as a path that globally minimizes

$$\int_0^\tau [2\Phi(\dot{m}) + 2\Psi(m) + \dot{I}(m)]\,\mathrm{d}t = I(m(\tau)) - I(m(0)) + 2\int_0^\tau [\Phi(\dot{m}) + \Psi(m)]\,\mathrm{d}t \tag{248}$$

subject to the terminal conditions $m(0) = 0$ and $m(\tau) = m \neq 0$. This variational principle is equivalent to the general variational principle of Eq. (227), and is what Onsager and Machlup refer to as the *minimum dissipation principle* [119]. The decay path bringing an initial fluctuation $m(0) = m \neq 0$ back to the equilibrium point $m^* = 0$ also satisfies this principle, but with the terminal conditions exchanged, i.e., with $m(0) = m$ and $m(\tau) = 0$.

From the symmetry of the associated Lagrange equation, it can be shown that the decay path is the time-reverse image of the corresponding fluctuation path. This holds, in general, whenever the dynamics of $M_n(t)$ is derived from a quasi-potential $I(m)$, that is, when m^* is an *equilibrium state* in the thermodynamic sense. When the dynamics of $M_n(t)$ involves external forces or non-conservative forces (in more than one dimension), the forward and backward optimal paths need not be the time-reverse of one another; see [150] for examples. In this case, m^* is called a stationary state rather than an equilibrium state.

The linear model of Onsager and Machlup serves as a template for constructing and studying other models of fluctuation dynamics, including models of nonequilibrium steady states (see, e.g., [153–155]), and for ultimately building a general theory of nonequilibrium processes (see, e.g., [12,140,156–159]). In going beyond this model, one may replace the linear force $b(m) = -am$ by nonlinear forces, consider noise processes with a non-zero mean or noise processes that are correlated in time, in addition to studying several (possibly coupled) macrostates rather than just one as we did in the previous example. One may also model the fluctuations of a field $\rho(x, t)$ using a general equation of the form

$$\partial_t \rho(x, t) = D(\rho(x, t)) + \xi(x, t), \tag{249}$$

where D is some operator acting on $\rho(x, t)$, and $\xi(x, t)$ is a space-time noise process. Stochastic field equations of this form are known as *hydrodynamic equations*, and are used to model turbulent fluids [160–162], as well as the macroscopic dynamics of particles evolving and interacting on lattices [12,163,164]; see Section 6.4. Note that for a field $\rho(x, t)$, the analog of the path probability density $P[m]$ of $M_n(t)$ is the functional probability density $P[\rho] = P(\{\rho(x, t)\}_{t=0}^{\tau})$, which gives the probability density that $\rho(x, t)$ follows a given "trajectory" or history $\{\rho(x, t)\}_{t=0}^{\tau}$ in some function space.

In the next subsection, we will apply methods inspired from the results of Onsager and Machlup to study the fluctuations of physical observables defined as time-averages over the paths of stochastic systems.

6.3. Additive processes and fluctuation relations

The large deviation results that we have surveyed so far were mostly concerned with the trajectories or paths of stochastic processes, and the probability density of these paths. Here we shall be concerned with random variables defined on these paths as additive functionals of the form

$$A_\tau[x] = \frac{1}{\tau}\int_0^\tau f(x(t))\,\mathrm{d}t, \tag{250}$$

where f is a smooth function mapping the state $x(t)$ of some stochastic process $X(t)$ to $\mathbb{R}^d$, $d \geq 1$. The random variable $A[x]$ is called the *time-average* of $f(x(t))$ over the time interval $[0, \tau]$. The usual problem that we are concerned with is to investigate whether, for a given stochastic process $X(t)$, A_τ satisfies a large deviation principle and, if so, to determine its rate function.

6.3.1. General results

As in the case of sample means of random variables, a large deviation principle can be derived for A_τ, at least in principle, via the Gärtner–Ellis Theorem. The scaled cumulant generation function in this case is

$$\lambda(k) = \lim_{\tau\to\infty}\frac{1}{\tau}\ln\left\langle \mathrm{e}^{\tau k\cdot A_\tau}\right\rangle, \quad k \in \mathbb{R}^d, \tag{251}$$

where

$$\left\langle \mathrm{e}^{\tau k\cdot A_\tau}\right\rangle = \int \mathrm{e}^{\tau k\cdot a} P(A_\tau \in \mathrm{d}a) = \int \mathcal{D}[x]\,P[x]\,\mathrm{e}^{\tau k\cdot A_\tau[x]} \tag{252}$$

and $P[x]$ is, as before, the probability density over the paths $\{x(t)\}_{t=0}^{\tau}$ extending from $t = 0$ to $t = \tau$. Provided that $\lambda(k)$ exists and is differentiable, we then have

$$P(A_\tau \in \mathrm{d}a) \asymp \mathrm{e}^{-\tau I(a)}\mathrm{d}a, \qquad I(a) = \sup_k\{k\cdot a - \lambda(k)\}. \tag{253}$$

If $X(t)$ is an ergodic Markov process, the large deviations of A_τ can be determined, also in principle, using the Donsker–Varadhan theory of Markov additive processes. In this case, $\lambda(k)$ is evaluated as the logarithm of the largest eigenvalue of the operator $L_k = L + k\cdot f$, L being the generator of the stochastic process $X(t)$ [2–4]; see also Sec. V.A of [28] and [165].[19] This result is the continuous-time generalization of the result of Section 4.3 stating that, for an ergodic Markov chain, $\lambda(k)$ is given by the logarithm of the largest eigenvalue of the "tilted" transition matrix Π_k. An important example of additive random variables, which has been extensively studied by Donsker and Varadhan [2–4], is presented next.

Example 6.7 (*Occupation Measure*). Let $1_A(x)$ denote the indicator function for the set A which equals 1 if $x \in A$ and 0 otherwise. The time-average of this function, given by

$$M_\tau(A) = \frac{1}{\tau}\int_0^\tau 1_A(x(t))\,\mathrm{d}t, \tag{254}$$

[19] There is a mathematical difficulty that will not be discussed here, namely, that the largest eigenvalue of L_k has to be isolated in order for the large deviation principle to hold.

gives the fraction of the time τ that the path $\{x(t)\}_{t=0}^{\tau}$ spends in A, and plays, as such, the role of the empirical vector for continuous-time dynamics. To make the connection more obvious, take A to be an infinitesimal interval $[x, x+\mathrm{d}x]$ anchored at the point x. Then $M_\tau(\mathrm{d}x) = M_\tau([x, x+\mathrm{d}x])$ "counts" the number of times $x(t)$ goes inside $[x, x+\mathrm{d}x]$. The density version of $M_\tau(\mathrm{d}x)$, defined as

$$L_\tau(x) = \frac{1}{\tau}\int_0^\tau \delta(x(t)-x)\,\mathrm{d}t, \tag{255}$$

"counts", similarly as for the empirical density defined in Eq. (90), the number of times that $x(t)$ hits a given point x as opposed to an interval of points.[20]

For many stochastic processes, L_τ is observed to converge in probability to a given stationary density in the long-time limit $\tau \to \infty$. The fluctuations of L_τ around this concentration point can be characterized by a rate function, which can formally be expressed via the Gärtner–Ellis Theorem as

$$I[\mu] = \sup_k\{\mu\cdot k - \lambda[k]\}, \tag{256}$$

where

$$\mu\cdot k = \int_{\mathbb{R}^d}\mu(x)k(x)\,\mathrm{d}x \tag{257}$$

and

$$\lambda[k] = \lim_{\tau\to\infty}\frac{1}{\tau}\ln\left\langle \exp\left\{\int_0^\tau k(x(t))\,\mathrm{d}t\right\}\right\rangle. \tag{258}$$

This result can be found in Gärtner [17]. More explicit expressions for $I[\mu]$ can be obtained from the general Legendre–Fenchel transform shown in Eq. (256) by considering specific random processes $X(t)$. In the particular case of an ergodic Markov process with generator G, Donsker and Varadhan obtained

$$I[\mu] = -\inf_{u>0}\left\langle\frac{Gu}{u}\right\rangle_\mu = -\inf_{u>0}\int\mu(x)\frac{(Gu)(x)}{u(x)}\,\mathrm{d}x \tag{259}$$

as part of their general theory of large deviations of Markov processes [2–4] (see also Sec. V.B of [28]). This rate function is the continuous-time analog of the rate function presented in Example 4.5. As in that example, the minimum and zero of $I[\mu]$ is the stationary probability density ρ^* of the ergodic Markov process generated by G. The rate function $I[\mu]$ characterizes the fluctuations of L_τ around that concentration point.

The next example shows how a large deviation principle can be derived for A_τ when the stochastic process $X(t)$ falls in the Freidlin–Wentzell framework of stochastic differential equations perturbed by Gaussian noise. The large deviation principle that one obtains in this case applies in the limit of vanishing noise, which is different from the $\tau\to\infty$ limit that we have just considered.

Example 6.8. Consider a general Markov process $X_\epsilon(t)$ arising, as in Section 6.1, as the solution of a dynamical system perturbed by a Gaussian white noise of strength ϵ, and let $A_\tau[x]$ be a time average defined over $X_\epsilon(t)$. From the Freidlin–Wentzell theory, we know that the large deviation principle $P[x] \asymp \mathrm{e}^{-J[x]/\epsilon}$ applies in the small noise limit $\epsilon\to 0$, with rate functional $J[x]$ given by Eq. (210). Since A_τ is a functional of $X_\epsilon(t)$, the contraction principle immediately implies that A_τ also satisfies a large deviation principle in the limit $\epsilon\to 0$, with rate function $I(a)$ given by the contraction of $J[x]$:

$$I(a) = \inf_{x(t):A_\tau[x]=a} J[x]. \tag{260}$$

As always, we can use Lagrange's multiplier method to transform this constrained maximization into an unconstrained optimization problem. This was done in Example 4.10 for the contraction of Sanov's Theorem, as well as in Example 5.1, which illustrated the maximum entropy principle. The unconstrained functional that we have to optimize in the present case is $K[x] = J[x] - \beta A_\tau[x]$, and involves the Lagrange multiplier β which takes care of the constraint $A_\tau[x] = a$.

Although the large deviation principle obtained in the previous example applies, strictly speaking, in the limit $\epsilon\to 0$, it can often be transformed into a large deviation principle for A_τ in the limit $\tau\to\infty$ by studying the extensivity of $I(a)$ with τ (see, e.g., [154,157]). The large deviation principle that one obtains in this case applies in the dual limit $\tau\to\infty$ and $\epsilon\to 0$, which means that it is only an approximation of the large deviation principle that governs the fluctuations of A_τ in the limit $\tau\to\infty$ for an *arbitrary* noise power ϵ, i.e., without the limit $\epsilon\to\infty$. In technical terms, this means that the knowledge of the optimal path solving the variational principle (260) is in general not sufficient to derive the long-time large deviations of A_τ for any noise power. The only exception to this statement, noted by Onsager and Machlup [119], are linear stochastic

[20] $L_\tau(x)$ is a density, so the number of times that $x(t)$ hits the interval $[x, x+\mathrm{d}x]$ is actually $L_\tau(x)\,\mathrm{d}x$.

differential equations, i.e., linear Langevin equations. For these, the evaluation of a path integral by its most probable path actually gives the exact value of the path integral for all $\epsilon > 0$, up to a normalization constant, which can usually be omitted for the purpose of deriving large deviation results.

6.3.2. Fluctuation relations

The next example is concerned with the large deviations of an additive process often studied in nonequilibrium statistical mechanics. This example is also our point of departure for studying large deviations of *nonequilibrium observables*, that is, random variables defined in the context of nonequilibrium systems, and for introducing an important class of results known as *fluctuation relations* or *fluctuation theorems*.

Example 6.9 (*Work Fluctuations for a Brownian Particle [166]*). Consider a Brownian particle immersed in a fluid, and subjected to the "pulling" force of a harmonic potential moving at constant velocity v_p. The dynamics of the particle is modeled, in the overdamped limit, by the Langevin equation

$$\dot{x}(t) = -[x(t) - v_p t] + \zeta(t), \tag{261}$$

where $x(t)$ denotes the position of the particle at time t, with $x(0) = 0$, and $\zeta(t)$ is a Gaussian white noise characterized by $\langle \zeta(t) \rangle = 0$ for all t and $\langle \zeta(t)\zeta(t') \rangle = 2\delta(t - t')$.[21] The *work per unit time* or *intensive work* W_τ done by the pulling force $F(t) = -[x(t) - v_p t]$ over an interval of time $[0, \tau]$ has for expression:

$$W_\tau = \frac{1}{\tau} \int_0^\tau v_p F(t)\, \mathrm{d}t = -\frac{v_p}{\tau} \int_0^\tau [x(t) - v_p t]\, \mathrm{d}t. \tag{262}$$

The large deviation principle governing the fluctuations of this additive process is found, following the Gärtner–Ellis Theorem, by calculating the scaled cumulant generating function $\lambda(k)$ of W_τ. This calculation can be performed using various methods (e.g., characteristic functions [166], differential equation techniques [167], path integrals [154], etc.), which all lead to

$$\lambda(k) = ck + ck^2 = ck(1 + k), \tag{263}$$

where $c = (v_p)^2$. Since this function is quadratic, the rate function of W_τ given by the Legendre–Fenchel transform of $\lambda(k)$ must also be quadratic, which implies that the fluctuations of W_τ are Gaussian. To be more precise, let $p(W_\tau = w)$ denote the probability density of W_τ. Then

$$p(W_\tau = w) \asymp \mathrm{e}^{-\tau I(w)}, \qquad I(w) = \inf_k \{kw - \lambda(k)\} = \frac{(w - c)^2}{4c}. \tag{264}$$

The main conclusion that we draw from this result is that positive amounts of work done by the pulling force on the Brownian particle are exponentially more probable than negative amounts of equal magnitude, since $c > 0$ for $v_p \neq 0$. To make this more obvious, consider the probability ratio

$$R_\tau(w) = \frac{p(W_\tau = w)}{p(W_\tau = -w)}. \tag{265}$$

Given the quadratic form of $I(w)$, it is easy to see that

$$R_\tau(w) \asymp \mathrm{e}^{\tau[I(-w) - I(w)]} = \mathrm{e}^{\tau w}. \tag{266}$$

Accordingly, the probability that $W_\tau = w > 0$ is, in the large time limit, exponential larger than the probability that $W_\tau = -w$.

The study of the probability ratio $R_\tau(w)$ for physical observables other than the work W_τ defined above has become an active topic of study in nonequilibrium statistical mechanics; see [168–171] for theoretical surveys of this topic, and [172,173] for more experimental surveys. The importance of $R_\tau(w)$ is justified by two observations. The first is that $R_\tau(w)$ provides a measure of how "out of equilibrium" a system is, since it yields information about the positive–negative asymmetry of nonequilibrium fluctuations that arises in general because of the irreversibility of the fluctuations paths [174,175]. The second observation is that the precise exponential form of $R_\tau(w)$ displayed in (266) appears to be a general law characterizing the fluctuations of several different nonequilibrium observables, not just the work W_τ considered above. A precise formulation of this law, now commonly referred to as the *fluctuation theorem* [176,177], can be given as follows. Let A_τ denote a nonequilibrium observable integrated over a time interval τ. For simplicity, let us assume that A_τ is a real random variable, and that $p(A_\tau = a)$ is non-zero for all $a \in \mathbb{R}$. Then A_τ is said to satisfy the fluctuation theorem if

$$R_\tau(a) = \frac{p(A_\tau = a)}{p(A_\tau = -a)} \asymp \mathrm{e}^{\tau c a}, \tag{267}$$

[21] Dimensional units are used here; see [166] for the full, physical version of this equation.

in the limit of large τ, with c a constant independent of a and τ. Equivalently, A_τ satisfies the fluctuation theorem if

$$\varrho(a) = \lim_{\tau\to\infty} \frac{1}{\tau} \ln R_\tau(a) = ca. \tag{268}$$

The expression "fluctuation theorem" should be used, strictly speaking, when the asymptotic result of (267) or (268) is proved for a specific nonequilibrium observable. When this result is experimentally or numerically verified rather than being proved, the expression "fluctuation relation" is more appropriate.

The first observation of a fluctuation relation was reported by Evans, Cohen and Morriss [178], who numerically studied the fluctuations of sheared fluids. Based on these results, Gallavotti and Cohen [176,177] then proved a fluctuation theorem for the entropy rate of chaotic, deterministic systems, which was later extended to general Markov processes by Kurchan [179], Lebowitz and Spohn [180], and Maes [181]. These results form the basis of several experimental studies of fluctuation relations arising in the context of particles immersed in fluids [182,183], electrical circuits [183–185], granular media [186–190], turbulent fluids [191,192], and the effusion of ideal gases [193], among other systems. The next example gives the essence of the fluctuation theorem for the entropy rate of Markov processes. This example is based on the results of Lebowitz and Spohn [180], and borrows some notations from Gaspard [194] (see also [195,196]). For a discussion of the entropy production rate based on the Donsker–Varadhan rate function of the empirical measure, the reader is referred to [197].

Example 6.10 (*Entropy Production*). Let $\sigma = \sigma_1, \sigma_2, \ldots, \sigma_n$ be the trajectory of a discrete-time ergodic Markov chain starting in the state σ_1 at time 1 and ending with the state σ_n at time n. Denote by σ^R the *time-reversed* version of σ obtained by reversing the order in which the states $\sigma_1, \sigma_2, \ldots, \sigma_n$ are visited in time, that is, $\sigma^R = \sigma_n, \sigma_{n-1}, \ldots, \sigma_1$. If the Markov chain is *reversible*, that is, if $P(\sigma) = P(\sigma^R)$ for all σ, then the *entropy production rate* of the Markov chain, defined as

$$W_n(\sigma) = \frac{1}{n} \ln \frac{P(\sigma)}{P(\sigma^R)}, \tag{269}$$

equals zero for all σ. Accordingly, to study the irreversibility of the Markov chain, we may study how W_n fluctuates around its mean $\langle W_n\rangle$, as well as how the mean differs from zero.

For an ergodic Markov chain, the Asymptotic Equipartition Theorem mentioned in Section 4.5 directly implies [194,198]

$$\lim_{n\to\infty} \langle W_n\rangle = h^R - h, \tag{270}$$

where h is the *forward* entropy rate defined in Section 4.5, and h^R is the *backward entropy* rate, defined as

$$h^R = \lim_{n\to\infty} -\frac{1}{n} \sum_\sigma P(\sigma) \ln P(\sigma^R). \tag{271}$$

To find the rate function governing the fluctuations of W_n around its mean, note that

$$\left\langle \mathrm{e}^{nkW_n} \right\rangle = \sum_\sigma P(\sigma) \frac{P(\sigma)^k}{P(\sigma^R)^k} = \sum_\sigma P(\sigma^R) \frac{P(\sigma)^{k+1}}{P(\sigma^R)^{k+1}}. \tag{272}$$

Summing over the time-reversed trajectories σ^R instead of σ leads to

$$\left\langle \mathrm{e}^{nkW_n} \right\rangle = \sum_{\sigma^R} P(\sigma^R) \frac{P(\sigma)^{k+1}}{P(\sigma^R)^{k+1}} = \sum_{\sigma^R} P(\sigma) \frac{P(\sigma^R)^{k+1}}{P(\sigma)^{k+1}}. \tag{273}$$

Thus $\langle \mathrm{e}^{nkW_n}\rangle = \langle \mathrm{e}^{n(-1-k)W_n}\rangle$ and

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{nkW_n} \right\rangle = \lambda(-1-k) \tag{274}$$

for all $k \in \mathbb{R}$.[22] From the Gärtner–Ellis Theorem, we therefore obtain

$$I(w) = \sup_k \{kw - \lambda(k)\} = \sup_k \{kw - \lambda(-1-k)\} = I(-w) - w \tag{275}$$

and, consequently, $\varrho(w) = I(-w) - I(w) = w$. This shows, as announced, that W_n satisfies the fluctuation theorem.

The fluctuation theorem for W_n can be generalized to continuous-state and continuous-time Markov processes by adapting the definition of the entropy production rate to these processes, and by studying the time reversibility of their

[22] The convergence of $\lambda(k)$ for all $k \in \mathbb{R}$ follows from the assumption that the space of σ is finite; see Section 4.3.

paths or, equivalently, the time reversibility of the master equation governing the evolution of probabilities defined on these paths (see, e.g., [174,179,180,199–202]). In the case of ergodic, continuous-time Markov processes, the symmetry of $\lambda(k)$ expressed in Eq. (274) can be related to a time-inversion symmetry of the processes' generator [179,180]. This time-inversion symmetry can also be used to prove fluctuation theorems for other nonequilibrium observables related to the entropy production rate, such as the work [174,179].

6.3.3. Fluctuation relations and large deviations

Fluctuation relations and fluctuation theorems are intimately linked to large deviation principles, as is obvious from the theory and examples studied so far. In a sense, one implies the other. This equivalence was partly observed by Gallavotti and Cohen in their original derivation of the fluctuation theorem [176,177] (see also [203,204]), and can be made more explicit by examining the chain of equalities displayed in (275). To put these equalities in a more general perspective, let us consider a general nonequilibrium observable A_τ integrated over a time τ, and let $\lambda(k)$ be its scaled cumulant generating function. By re-stating the result of (275) for A_τ, it is first obvious that, if $\lambda(k)$ satisfies the conditions of the Gärtner–Ellis Theorem (differentiability and steepness), in addition to the symmetry property $\lambda(k) = \lambda(-k - c)$ for all $k \in \mathbb{R}$ and c a real constant, then $\varrho(a) = ca$. By inverting the Legendre–Fenchel transform involved in (275), we also obtain the converse result, namely that, if A_τ satisfies a large deviation principle with rate function $I(a)$, and $\varrho(a) = ca$ for some real constant c, then $\lambda(k) = \lambda(-k - c)$. By combining these two results, we thus see that the symmetry $\lambda(k) = \lambda(-c - k)$ is essentially equivalent to having a fluctuation theorem for A_τ.

When applying these results, it is important to note that the symmetry property $\lambda(k) = \lambda(-k-c)$ can be satisfied even if the generating function of A_τ satisfies the same property but only approximately in the limit of large τ. In Example 6.10, it so happens that this property is satisfied exactly by the generating function. Moreover, for processes having a countably-infinite or continuous state space, $\lambda(k)$ does not exist in general for all $k \in \mathbb{R}$, but only for a convex subset of $\mathbb{R}$ (see Examples 4.3 and 4.8). In this case, results similar to those above apply but in a pointwise sense. That is, if $\lambda(k)$ is differentiable at k and satisfies the symmetry $\lambda(k) = \lambda(-k - c)$ for the same value k, then $\varrho(a) = ca$ for a such that $a = \lambda'(k)$.[23] To formulate a converse to this result, we can follow the same arguments as above to prove that, if $\lambda(k) \neq \lambda(-k - c)$ for at least one value k, then $\varrho(a)$ is not proportional to a for at least one value a. Therefore, if $\lambda(k)$ satisfies the symmetry property only for a subset of $\mathbb{R}$, then $\varrho(a) = ca$ only for a subset of the values of A_τ. In this case, we say that A_τ satisfies an *extended* fluctuation theorem.

To make sure that extended fluctuation theorems are not confused with the fluctuations theorems defined at the start of this section, it is common to refer to the latter ones as *conventional* fluctuation theorems [171,205,206]. Thus an observable A_τ is said to satisfy a *conventional* fluctuation theorem if its fluctuation function $\varrho(a)$, defined by the limit of (268), is linear in a, i.e., if $\varrho(a) = ca$. If $\varrho(a)$ is a nonlinear function of a, then A_τ is said to satisfy an *extended* fluctuation theorem. The reader will judge by him- or herself whether these definitions are useful. In the end, it should be clear that the rate function $I(a)$ completely characterizes the fluctuations of A_τ in the long time limit, so that one might question the need to attach the terms "conventional" or "extended" to $I(a)$. Indeed, one might even question the need to define $\varrho(a)$ when one has $I(a)$.

One example of nonequilibrium observables that illustrates the notion of extended fluctuation theorem is the heat per unit time dissipated by the dragged particle of Example 6.9. For this observable, van Zon and Cohen [205,206] have shown that the symmetry on $\lambda(k)$ holds only on a bounded interval due to the fact that $\lambda(k)$ does not converge for all $k \in \mathbb{R}$. Violations of the fluctuation theorem have also been shown to arise from the choice of initial conditions [207–209], the unboundness of the observable considered [210], or the restriction of the domain of $\lambda(k)$ [211,212]. These violations are all included in large deviation theory insofar as they reflect special properties of rate functions and their associated scaled cumulant generating functions. Indeed, many of the limiting cases of rate functions that we have discussed in Section 4 do arise in the context of nonequilibrium fluctuations, and lead to violations and possible extensions of the fluctuation theorem. The extended fluctuation theorem of van Zon and Cohen [205,206], for instance, is closely related to affine rate functions, studied in Examples 3.3 and 4.8. A model for which $\lambda(k)$ is found to be nondifferentiable is discussed in [212]. Finally, a model of nonequilibrium fluctuations having a zero rate function is discussed in [167]. This model is a Markov equivalent of the sample mean of IID symmetric Lévy random variables that was considered in Example 4.2.

To close our discussion of fluctuation relations and fluctuation theorems, note that a fluctuation theorem may hold approximately for the small values of A_τ even if its associated $\lambda(k)$ does not satisfy the symmetry $\lambda(k) = \lambda(-k - c)$. This follows by noting that if a large deviation principle holds for A_τ with a rate function $I(a)$ which is differentiable at $a = 0$, then $\varrho(a) \approx -2I'(0)a$ to first order in a [186,207]. If $I(a)$ has a parabolic minimum a^*, we can also write $\varrho(a) \approx 2I''(a^*)a^*a$ to second order in $a - a^*$. In both cases, $\varrho(a)$ is linear in a, which is the defining property of conventional fluctuation relations.

6.4. Interacting particle models

Markovian models of interacting particles, such as the exclusion process, the zero-range process, and their many variants (see [163,213]), have been, and still are, extensively studied from the point of view of large deviations. The

[23] This follows by applying the local Legendre transform of Eq. (82), which we have discussed in the context of nondifferentiable points of $\lambda(k)$ and nonconvex rate functions; see Sections 4.1 and 4.4.

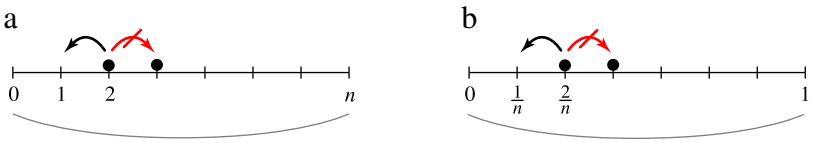

Fig. 20. (a) Exclusion process on the lattice $\mathbb{Z}_n$ and (b) rescaled lattice $\mathbb{Z}_n/n$. A particle can jump to an empty site (black arrow) but not to an occupied site (red arrow). The thin line at the bottom indicates the periodic boundary condition $\eta(0) = \eta(1)$.

interest for these models comes from the fact that their macroscopic or *hydrodynamic* behavior can be determined from their "microscopic" dynamics, sometimes in an exact way. Moreover, the typicality of the hydrodynamic behavior can be studied by deriving large deviation principles which characterize the probability of observing deviations in time from the hydrodynamic evolution [164]. The interpretation of these large deviation principles follows the Freidlin–Wentzell theory, in that a deterministic dynamical behavior—here the hydrodynamic behavior—arises as the global minimum and zero of a given (functional) rate function. From this point of view, the hydrodynamic equations, which are the equations of motion describing the hydrodynamic behavior, can be characterized as the solutions of a variational principle similar to the minimum dissipation principle of Onsager [214].

Two excellent review papers [113,215] have appeared recently on interacting particle models and their large deviations, so we will not review this subject in detail here. The next example illustrates in the simplest way possible the gist of the results that are typically obtained when studying these models. The example follows the work of Kipnis, Olla and Varadhan [216], who were the first to apply large deviation theory for studying the hydrodynamic limit of interacting particle models.

Example 6.11 (*Simple Symmetric Exclusion Process*). Consider a system of k particles moving on the lattice $\mathbb{Z}_n$ of integers ranging from 0 to n, $n > k$; see Fig. 20(a). The rules that determine the evolution of the particles are assumed to be the following:

- A particle at site i waits for a random exponential time with mean 1, then selects one of its neighbors j at random.
- The particle at i jumps to j if j is unoccupied; if j is occupied, then the particle stays at i and goes to a waiting period again before choosing another neighbor to jump to (exclusion principle).

We denote by $\eta_t(i)$ the occupation of the "site" $i \in \mathbb{Z}_n$ at time t, and by $\eta_t = (\eta_t(0), \eta_t(1), \ldots, \eta_t(n-1))$ the whole configuration or *microstate* of the system. Because of the exclusion principle, $\eta_t(i) \in \{0, 1\}$. Moreover, we impose boundary conditions on the lattice by identifying the first and last site.

The generator of the Markovian process defined by the rules above can be written explicitly by noting that there can be a jump from i to j only if $\eta(i) = 1$ and $\eta(j) = 0$. Therefore,

$$(Lf)(\eta) = \frac{1}{2} \sum_{|i-j|=1} \eta(i)[1 - \eta(j)][f(\eta^{i,j}) - f(\eta)], \tag{276}$$

where f is any function of η, and $\eta^{i,j}$ is the configuration obtained after one jump, that is, the configuration obtained by exchanging the occupied state at i with the unoccupied state at j:

$$\eta^{i,j}(k) = \begin{cases} \eta(i) & \text{if } k = j \\ \eta(j) & \text{if } k = i \\ \eta(k) & \text{otherwise.} \end{cases} \tag{277}$$

To obtain a hydrodynamic description of this dynamics, we rescale the lattice spacing by a factor $1/n$, as shown in Fig. 20(b), and take the limit $n \to \infty$ with $r = k/n$, the density of particles, fixed. Furthermore, we speed-up the time t by a factor n^2 to overcome the fact that the diffusion dynamics of the particle system "slows" down as $n \to \infty$. In this limit, it can be proved that the empirical density of the rescaled dynamics, defined by

$$\pi_t^n(x) = \frac{1}{n} \sum_{i \in \mathbb{Z}_n} \eta_{n^2 t}(i)\, \delta(x - i/n), \tag{278}$$

where x is a point of the unit circle C, weakly converges in probability to a field $\rho_t(x)$ which evolves on C according to the diffusion equation

$$\partial_t \rho_t(x) = \partial_{xx} \rho_t(x). \tag{279}$$

It can also be proved that the fluctuations of $\pi_t^n(x)$ around the deterministic field $\rho_t(x)$ follows a large deviation principle, expressed heuristically as

$$P_n[\pi_t^n = \pi_t] = P_n(\{\pi_t^n(x) = \pi_t(x)\}_{t=0}^{\tau}) \asymp \mathrm{e}^{-nI[\pi_t]}. \tag{280}$$

The interpretation of this expression follows the interpretation of the density $P_\epsilon[x]$ considered earlier: $P_n[\pi_t^n = \pi]$ is the probability density for the evolution of a field in time, so that the rate function shown in (280) is a space and time functional of that field. The expression of this rate function is relatively complicated compared to all the rate functions studied in this review. It involves two parts: a static part, which measures the "cost" of the initial field π_0^n, and a dynamic part, which measures the cost of the deviation of π_t^n from ρ_t; see [216] for the full expression of the rate function, and [12] for a discussion of its physical interpretation.

The previous example can be generalized in many different ways. One can consider asymmetric exclusion processes for which the diffusion is enhanced in one direction (see, e.g., [217,218]), or extended exclusion processes for which jumps to sites other than first neighbors are allowed. One can also consider models that allow more than one particle at each site, such as the zero-range process (see, e.g., [219,220]), or models with particle reservoirs that add and remove particles at given rates. Moreover, one can choose not to impose the exclusion rule, in which case the particles jump independently of one another [221,222].

For many of these models, large deviation principles have been derived at the level of the empirical density or density field [223–227], as well as at the level of the current [228–230], which measures the average number of particles moving on the lattice. The rate functions associated with these observables show many interesting properties. In the case of the totally asymmetric exclusion process, for instance, the rate function of the density field is nonconvex [226,227]. This provides a functional analog of nonconvex rate functions. The reader will find many details about these large deviation results in the two review papers mentioned earlier. The first one, written by Derrida [113], is useful for gaining a feeling of the mathematics involved in the derivation of large deviation results for interacting particle models. The review written by Bertini et al. [215], on the other hand, is useful for gaining an overview of the different models that have been studied, and of the theory that describes the fluctuations of these models at the macroscopic level. For a study of interacting particle systems based on the Donsker–Varadhan theory, see [231].

To close this short discussion of large deviations in interacting particle models, let us mention that Derrida and Bodineau [228] have formulated a useful calculation tool for obtaining the rate function of the current in interacting particle models, which they dubbed the *additivity principle*. This principle is close in spirit to the Freidlin–Wentzell theory (see Section 6.1), and appears to be based, as for that theory, on a Markov property of fluctuations. For a presentation of this principle and its applications, see Derrida [113].

7. Other applications

The results, techniques, and examples compiled in the previous sections make for a more or less complete toolbox that can be used to study other applications of large deviations in statistical mechanics. We conclude this review by mentioning four more important applications related to multifractals, chaotic systems, disordered systems, and quantum systems. Our discussion of these applications is far from exhaustive; our aim is merely to mention them, and to point out a few useful references for those who want to learn more about them.

7.1. Multifractals

The subject of multifractal analysis was developed independently of large deviation theory, and is typically not presented from the point of view of this theory (see, e.g., [51,232–234]). The two subjects, however, have much in common. In fact, one could say that multifractal analysis is a large deviation theory of self-similar measures, or a large deviation theory of the measure equivalent of self-processes, studied in Section 4.5. A presentation of multifractal analysis in these terms is given in [235,236], as well as in the book of Harte [237].

The idea that multifractal analysis is related to large deviation theory, or is an application of large deviation theory, becomes more obvious by noting the following:

- The two basic quantities commonly employed to characterize multifractals—the so-called *multifractal spectrum* and *structure function*—are the analogs of an entropy and a free energy function, respectively.
- The scaling limit underlying the multifractal spectrum and the structure function has the form of a large deviation limit.
- The multifractal spectrum and structure function are related by Legendre transforms.

The last point is the perhaps the most revealing: the fact that two functions are found to be related by a Legendre (or Legendre–Fenchel) transform is often the sign that a large deviation principle underlies these functions. This is the case for the entropy and the free energy of equilibrium statistical mechanics, as we have seen in Section 5, and this is the case, too, for the multifractal spectrum and the structure function.

By re-interpreting in this way multifractal analysis in terms of large deviations, we do more than just translating a theory in terms of another—we gain a rigorous formulation of multifractals, as well as a guide for deriving new results about multifractals. One case in point concerns nonconvex rate functions. It had been known for some time that the structure function of multifractal analysis, which is the analog of the function $\varphi(\beta)$ or $\lambda(k)$ studied here, can be nondifferentiable, and that the nondifferentiable points of this function signal the appearance of a multifractal analog of first-order phase transitions (see [51] and references cited therein). Some confusion reigned as to how the multifractal spectrum had to be calculated in this case. Many authors assumed that the multifractal spectrum is always the Legendre–Fenchel transform

of the structure function, and so concluded that the spectrum must be affine if the structure function is nondifferentiable [238,239]. The correct answer given by large deviation theory is more involved: the multifractal spectrum can be concave or nonconcave, in the same way that an entropy can be concave or nonconcave. If it is concave, then it can be calculated as the Legendre–Fenchel transform of the structure function, otherwise, it cannot. A recent discussion of this point can be found in [240]; see also [241–243] for mathematical examples of multifractals having nonconcave spectra.

7.2. Thermodynamic formalism of chaotic systems

The Freidlin–Wentzell theory of differential equations perturbed by noise has its analog for discrete-time dynamical maps, which was developed by Kifer [244,245]. One interesting aspect of dynamical systems, be they represented by flows or maps, is that they often give rise to large deviation principles without a perturbing noise. In many cases, the chaoticity and mixing properties of a deterministic system are indeed such that they induce a seemingly stochastic behavior of that system, which induces, in turn, a stochastic behavior of observables of that system. The study of this phenomenon is the subject of the theory of chaotic systems and ergodic theory (see, e.g., [52,246–248]), and the study of large deviations in the context of these theories is the subject of the so-called *thermodynamic formalism* developed by Ruelle [103,249] and Sinai [250,251]. For an introduction to this formalism, see [51,252].

As in the case of multifractals, the thermodynamic formalism was developed independently of large deviation theory. But it is also clear with hindsight that this formalism can be re-interpreted or recast in the language of large deviations. The basis of this interpretation can be summarized with the following basic observations:

- The so-called topological pressure, which plays a central role in the thermodynamic formalism, is a scaled cumulant generating function.
- The entropy function of an observable, as defined in the thermodynamic formalism, is a rate function.
- The topological pressure and entropy are related by Legendre transforms when the entropy is concave.
- An equilibrium state in the thermodynamic formalism has the same large deviation interpretation as an equilibrium state in equilibrium statistical mechanics: both are the solution of a variational principle which can be derived from the contraction principle.

The reader is referred to the review paper of Oono [9] for an explanation of some of these points; see also [253–255]. A number of results that establish a direct connection between dynamical systems and large deviation theory can be found in [252,256–260]. For a derivation of fluctuation theorems in the context of chaotic maps, see [261]. Finally, for a large deviation study of multiplicative processes in deterministic systems having many degrees of freedom, see [262] and references therein.

At the time of writing this review, a complete presentation of the thermodynamic formalism that refers explicitly to large deviation theory has yet to appear. The fact that chaotic maps can be thought of as continuous-state Markov chains with transition matrix given by the Frobenius operator appears to be a good starting point for establishing a direct link between the thermodynamic formalism and large deviation theory (see, e.g., [263,264]).

7.3. Disordered systems

The application of large deviation techniques for studying disordered systems focuses in the literature on two different models: random walks in random environments (see, e.g., [265–268]) and spin glasses (see, e.g., [269–271]). Large deviation principles can be derived, for both applications, at the *quenched* level, i.e., for a fixed realization of the random disorder, or at the *annealed* level, which involves an average over the disorder. An interesting question in the context of random walks in random environments is whether a large deviation arises out of an atypical state of the walk or out of the atypicality of a specific random environment. A similar question arises for spin glasses in the form of, is the equilibrium state of a spin glass obtained for a specific random interaction typical in the ensemble of all interactions? The book of den Hollander [40] and the recent review paper by Zeitouni [268] offer two good entry points to the first question; see [272,273] for a mathematical discussion of spin glasses.

From the large deviation point of view, the difference between disordered and regular systems is that generating functions defined in the context of the former systems have an extra dependence on a "disorder" variable, which implies that these generating functions are random variables themselves. Therefore, in addition to studying the "quenched" large deviations associated with a given "random" generating function (i.e., a generating function arising for a given realization of the disorder), one can study the large deviations of the generating function itself, in order to determine the most probable value of the generating function. This concentration value of the generating function often simplifies the study of "annealed" large deviations, which are obtained from generating functions averaged over the disorder. For a discussion of spin glasses which follows this point of view, see the recent book of Mézard and Montanari [274].

7.4. Quantum large deviations

Quantum systems have entered the large deviation scene relatively recently compared to classical systems: end of 1980s compared to early 1970s. Applications of large deviations for studying boson gases are described in [275–277];

quantum gases are considered in [278–280], while quantum spin systems are considered in [281–283]. For an application of Varadhan's Theorem for a class of mean-field quantum models, see [284].

A quantum version of Sanov's Theorem is presented in [285]. As for classical version of that theorem, the quantum version plays an important role in the theory of estimation and in information theory, as generalized to the quantum world [286–288]. Finally, a quantum adaptation of the Freidlin–Wentzell theory of dynamical systems perturbed by noise can be found in [289].

Acknowledgments

The writing of this review has benefited from the help of many people who provided ideas, criticisms, guidance, encouragements, as well as useful opportunities to lecture about large deviations. Among these, I would like to thank Fulvio Baldovin, Julien Barré, Freddy Bouchet, Eddie G. D. Cohen, Claude Crépeau, Thierry Dauxois, Bernard Derrida, Richard S. Ellis, Vito Latora, Michael C. Mackey, Stefano Ruffo, Attilio Stella, Julien Tailleur, Tooru Taniguchi, Bruce Turkington, the Touchette-Ostiguy family, and my colleagues of the statistical mechanics study group at QMUL. I also thank Rosemary J. Harris, Michael Kastner, and Michael K.-H. Kiessling for reading the manuscript. A special thank is also due to Ana Belinda Peñalver Peña for her more than needed support.

My intermittent work on this review has been supported over the last few years by NSERC (Canada), the Royal Society of London (Canada–UK Millennium Fellowship), and RCUK (Interdisciplinary Academic Fellowship).

Appendix A. Summary of main mathematical concepts and results

- **Large deviation principle** (Section 3): Let $\{A_n\}$ be a sequence of random variables indexed by the positive integer n, and let $P(A_n \in \mathrm{d}a) = P(A_n \in [a, a + \mathrm{d}a])$ denote the probability measure associated with these random variables. We say that A_n or $P(A_n \in \mathrm{d}a)$ satisfies a *large deviation principle* if the limit

$$I(a) = \lim_{n\to\infty} -\frac{1}{n} \ln P(A_n \in \mathrm{d}a) \tag{A.1}$$

  exists (see Appendix B for a more precise definition). The function $I(a)$ defined by this limit is called the *rate function*; the parameter n of decay is called in large deviation theory the *speed*.[24]
- **Asymptotic notation:** The existence of a large deviation principle for A_n means concretely that the dominant behavior of $P(A_n \in \mathrm{d}a)$ is a decaying exponential with n, with rate exponent $I(a)$. We summarize this property by writing

$$P(A_n \in \mathrm{d}a) \asymp \mathrm{e}^{-nI(a)}\,\mathrm{d}a. \tag{A.2}$$

  The infinitesimal element $\mathrm{d}a$ is important in this expression; if we work with the probability density $p(A_n = a)$ instead of the probability measure $P(A_n \in \mathrm{d}a)$, then the large deviation principle is expressed simply as $p(A_n = a) \asymp \mathrm{e}^{-nI(a)}$.
- **Generating function:** The *generating function* of A_n is defined as

$$W_n(k) = \langle \mathrm{e}^{nkA_n} \rangle = \int \mathrm{e}^{nka}\, P(A_n \in \mathrm{d}a), \quad k \in \mathbb{R}. \tag{A.3}$$

  In terms of the density $p(A_n)$, we have instead

$$W_n(k) = \int \mathrm{e}^{nka}\, p(A_n = a)\,\mathrm{d}a. \tag{A.4}$$

  In both expressions, the integral is over the domain of A_n.
- **Scaled cumulant generating function:** The function $\lambda(k)$ defined by the limit

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln W_n(k) \tag{A.5}$$

  is called the *scaled cumulant generating function* of A_n. It is also called the *log-generating function* or *free energy function* of A_n. The existence of this limit is equivalent to writing $W_n(k) \asymp \mathrm{e}^{n\lambda(k)}$.
- **Legendre–Fenchel transform:** The *Legendre–Fenchel transform* of a function $f(x)$ is defined by

$$g(k) = \sup_x \{kx - f(x)\}. \tag{A.6}$$

  This transform is often written in convex analysis in the compact form $g = f^*$. This transform is also sometimes written with an infimum instead of a supremum. With the supremum, the Legendre–Fenchel transform reduces to the standard *Legendre transform* when $f(x)$ is strictly convex and differentiable, for then

$$g(k) = kx(k) - f(x(k)), \tag{A.7}$$

[24] See footnote 11.

where $x(k)$ is the unique root of $f'(x) = k$. The Legendre–Fenchel transform involving the infimum reduces to the standard Legendre transform when $f(x)$ is strictly concave and differentiable. The formula of the Legendre transform, in this case, is the same as above.

- **Gärtner–Ellis Theorem:** If $\lambda(k)$ is differentiable, then A_n satisfies a large deviation principle with rate function $I(a)$ given by the Legendre–Fenchel transform of $\lambda(k)$:

$$I(a) = \sup_k \{ka - \lambda(k)\}. \tag{A.8}$$

  This Legendre–Fenchel transform is expressed in convex analysis by the shorthand notation $I = \lambda^*$. (See Section 3 for a more precise statement of this theorem.)
- **Varadhan's Theorem:** If A_n satisfies a large deviation principle with rate function $I(a)$, then its scaled cumulant generating function $\lambda(k)$ is the Legendre–Fenchel transform of $I(a)$:

$$\lambda(k) = \sup_k \{ka - I(a)\}. \tag{A.9}$$

  In shorthand notation, this is expressed as $\lambda = I^*$. (See Section 3 for a more precise statement of this theorem.)
- **Convex versus nonconvex rate functions** (Section 4): If $I(a)$ is convex, then $I = \lambda^*$. If $I(a)$ is nonconvex, then $I \neq \lambda^*$. As a corollary, rate functions that are nonconvex cannot be calculated via the Gärtner–Ellis Theorem because rate functions obtained from this theorem are always convex (strictly convex, in fact).
- **Properties of** $\lambda(k)$ (Section 3):
  1. $\lambda(0) = 0$. This follows from the normalization of probabilities.
  2. $\lambda'(0) = \lim_{n\to\infty} \langle A_n \rangle$. This property is related to the Law of Large Numbers.
  3. $\lambda''(0) = \lim_{n\to\infty} n \operatorname{var}(A_n)$. This property is related to the Central Limit Theorem.
  4. $\lambda(k)$ is convex. This implies, among other things, that $\lambda(k)$ can be nondifferentiable only at isolated points.
  5. $\lambda(k)$ is differentiable if $I(a)$ is strictly convex, i.e., convex with no linear parts.
  6. $\lambda(k)$ has at least one nondifferentiable point if $I(a)$ is nonconvex or has linear parts.
  7. Suppose that $\lambda(k)$ is differentiable. Then the value k such that $\lambda'(k) = a$ has the property that $k = I'(a)$. This is the statement of the Legendre duality between λ and I, which can be expressed in words by saying that the slopes of λ correspond to the abscissas of I, while the slopes of I correspond to the abscissas of λ. (This property can be generalized to a nondifferentiable $\lambda(k)$ and a nonconvex $I(a)$ with the concept of supporting lines [19].)
- **Contraction principle** (Section 3): Consider two sequences of random variables $\{A_n\}$ and $\{B_n\}$ such that $A_n = f(B_n)$, and assume that B_n obeys a large deviation principle with rate function I_B. Then A_n obeys a large deviation principle with rate function I_A given by

$$I_A(a) = \inf_{a:f(b)=a} I_B(b). \tag{A.10}$$

- **Connection with physics:** Entropies are rate functions; free energies are scaled cumulant generating functions.

Appendix B. Rigorous formulation of the large deviation principle

This appendix is an attempt at explaining the rigorous formulation of the large deviation principle for the benefit of physicists not versed in topology and measure theory. The formulation presented here is inspired from the work of Ellis [7,10,18,290], which is itself inspired from Varadhan [22]. For background material on topology and measure theory, the reader should consult Appendices B and D of [13].

The rigorous definition of the large deviation principle is based on four basic ingredients:

- A sequence of probability spaces $\{(\Lambda_n, \mathcal{F}_n, P_n), n \in \mathbb{N}\}$ consisting of a probability measure P_n defined on the set $\mathcal{F}_n$ of all (Borel) sets of the "event" set Λ_n.
- A sequence of random variables $\{Y_n, n \in \mathbb{N}\}$ mapping Λ_n into a complete, separable metric space $\mathcal{X}$, also known as a *Polish* space.
- A sequence $\{a_n : n \in \mathbb{N}\}$ of positive constants such that $a_n \to \infty$ as $n \to \infty$.
- A lower semi-continuous function $I(x)$ mapping $\mathcal{X}$ into $[0, \infty]$.

From the point of view of statistical mechanics, Λ_n can be thought of as the space of the microstates of an n-particle system. The set $\mathcal{F}_n$ is the set of all possible events (sets) on Λ_n, whereas P_n is a probability measure on $\mathcal{F}_n$. The fact that we are dealing with a "sequence" of probability spaces is there, of course, because we are interested in studying the behavior of P_n in the limit $n \to \infty$, which we call the thermodynamic limit. In the same vein, Y_n should be thought of as a macrostate, and $\mathcal{X}$ as the macrostate space. One can think of Y_n, for example, as the mean magnetization of a simple spin model, in which case $\mathcal{X} = [-1, 1]$. In all the applications covered in this review, $\mathcal{X}$ is a subset of $\mathbb{R}^d$, and so we need not bother with the fact that $\mathcal{X}$ is a "complete, separable metric" space. This requirement is a technicality used by mathematicians to make the theory of large deviations as general as possible.

The random variable for which we are interested to formulate a large deviation principle is Y_n. The probability measure P_n defined at the level of Λ_n is extended to Y_n via

$$P_n(Y_n \in B) = \int_{\{\omega\in\Lambda_n : Y_n(\omega)\in B\}} P_n(\mathrm{d}\omega), \tag{B.1}$$

where B is any subset of $\mathcal{X}$. Given this probability, we say that the sequence $\{Y_n, n \in \mathbb{N}\}$ satisfies a *large deviation principle* on $\mathcal{X}$ with *rate function* I and *speed* a_n if for any *closed* set C,

$$\limsup_{n\to\infty} \frac{1}{a_n} \ln P_n(Y_n \in C) \leq -\inf_{y\in C} I(y), \tag{B.2}$$

and for any *open* set O,

$$\liminf_{n\to\infty} \frac{1}{a_n} \ln P_n(Y_n \in O) \geq -\inf_{y\in O} I(y). \tag{B.3}$$

The lower semi-continuity of I guarantees that this function achieves its minimum on any closed sets (a lower semi-continuous function has closed level sets; see Chap. 5 of [21]).

The two limits (B.2) and (B.3) give a rigorous meaning to the two bounds mentioned in our formal discussion of the large deviation principle; see Section 3. To understand why the first limit involves closed sets and the second open sets, we need to invoke the notion of weak convergence. The idea, as partly explained in Section 3, is that we wish to approximate a measure μ_n by a limit measure μ such that

$$\lim_{n\to\infty} \int_{\mathcal{X}} f(y)\mu_n(\mathrm{d}y) = \int_{\mathcal{X}} f(y)\mu(\mathrm{d}y) \tag{B.4}$$

for all bounded and continuous functions $f : \mathcal{X} \to \mathbb{R}$. Though less transparent, an equivalent and often more practical way of expressing the weak convergence of μ_n to μ is provided by the so-called *Portmanteau Theorem* (see Sec. D.2 of [13]), which states that the limit (B.4) is equivalent to

$$\limsup_{n\to\infty} \mu_n(C) \leq \mu(C) \tag{B.5}$$

for all closed subsets C of $\mathcal{X}$, and

$$\liminf_{n\to\infty} \mu_n(O) \geq \mu(O) \tag{B.6}$$

for all open subsets O of $\mathcal{X}$. These two limits correspond to the two limits shown in (B.2) and (B.3), with μ_n equal to $a_n^{-1} \ln P_n$ to account for the scaling $P_n \asymp \mathrm{e}^{-a_n I}$, $I \geq 0$.

The heuristic form of the large deviation principle that we use as the basis of this review is a simplification of the rigorous formulation, in that we assume, as in Section 3, that the large deviation upper and lower bounds, defined by (B.2) and (B.3) respectively, are the same. This is a strong simplification, which happens to be verified only for so-called *I-continuity sets*, that is, sets A such that

$$\inf_{y\in\bar{A}} I(y) = \inf_{y\in A^\circ} I(y), \tag{B.7}$$

where $\bar{A}$ and A° denote, respectively, the closure and relative interior of A; see Sec. 3 of [10] or [290] for more details. In treating large deviations, we also take the simplifying step of considering probabilities of the form $P(Y_n \in dy)$, where $dy = [y, y + dy]$ with a bit of abuse of notation, in which case

$$\lim_{n\to\infty} \frac{1}{a_n} \ln P(Y_n \in dy) = -\inf_{x\in[y,y+dy]} I(x) = -I(y). \tag{B.8}$$

Finally, in most examples covered in this review, the speed a_n is equal to n. In statistical mechanics, the proportionality of a_n with n is an expression of the concept of extensivity.

Appendix C. Derivations of the Gärtner–Ellis Theorem

We give here two derivations of the Gärtner–Ellis Theorem for random variables taking values in $\mathbb{R}$. The first derivation is inspired from the work of Daniels [33] on saddle-point approximations in statistics, and is presented to reveal the link that exists between the large deviation principle, the saddle-point approximation, and Laplace's approximation.[25] The second derivation is based on a clever change of measure which goes back to Cramér [1], and which is commonly used nowadays to prove large deviation principles. None of the derivations is rigorous.

[25] The so-called Darwin–Fowler method [291] used in statistical mechanics is yet another example of saddle-point or Laplace approximation, as applied to discrete generating functions.

C.1. Saddle-point approximation

Consider a random variable $S_n(\omega)$ which is a function of a sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ of n random variables. The random variable S_n need not be a sample mean, but it is useful to think of it as being one. For simplicity, assume that the ω_i's are also real random variables, so that $\omega \in \mathbb{R}^n$. Denoting by $p(\omega)$ the probability density of ω, we write the probability density of S_n as

$$p(S_n = s) = \int_{\{\omega\in\mathbb{R}^n : S_n(\omega)=s\}} p(\omega)\,\mathrm{d}\omega = \int_{\mathbb{R}^n} \delta(S_n(\omega) - s)\, p(\omega)\,\mathrm{d}\omega = \langle \delta(S_n - s)\rangle\,, \tag{C.1}$$

just as in Eq. (6) of Section 2. Using the Laplace transform representation of Dirac's delta function,

$$\delta(s) = \frac{1}{2\pi\,\mathrm{i}} \int_{a-\mathrm{i}\infty}^{a+\mathrm{i}\infty} \mathrm{e}^{\zeta s}\,\mathrm{d}\zeta\,, \quad a \in \mathbb{R}, \tag{C.2}$$

we then write

$$p(S_n = s) = \frac{1}{2\pi\,\mathrm{i}} \int_{a-\mathrm{i}\infty}^{a+\mathrm{i}\infty} \mathrm{d}\zeta \int_{\mathbb{R}^n} \mathrm{d}\omega\, p(\omega)\, \mathrm{e}^{\zeta[S_n(\omega)-s]} = \frac{1}{2\pi\,\mathrm{i}} \int_{a-\mathrm{i}\infty}^{a+\mathrm{i}\infty} \mathrm{d}\zeta\, \mathrm{e}^{-\zeta s} \int_{\mathbb{R}^n} \mathrm{d}\omega\, p(\omega)\, \mathrm{e}^{\zeta S_n(\omega)}. \tag{C.3}$$

The integral of the Laplace transform is performed along the so-called *Bromwich contour*, which runs parallel to the imaginary axis from $\zeta = a - \mathrm{i}\infty$ to $\zeta = a + \mathrm{i}\infty$, $a \in \mathbb{R}$.

At this point, we anticipate the scaling of the large deviation principle by performing the change of variable $\zeta \to n\zeta$, and note that if

$$\lambda(\zeta) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{n\zeta S_n} \right\rangle \tag{C.4}$$

exists, then

$$p(S_n = s) \asymp \int_{a-\mathrm{i}\infty}^{a+\mathrm{i}\infty} \mathrm{d}(\zeta)\, \mathrm{e}^{-n[\zeta s - \lambda(\zeta)]} \tag{C.5}$$

with sub-exponential corrections in n. By deforming the contour so that it goes through the saddle-point ζ^* of $\zeta s - \lambda(\zeta)$, and by considering only the exponential contribution to the integral coming from the saddle-point, we then write

$$p(S_n = s) \asymp \int_{\zeta^*-\mathrm{i}\infty}^{\zeta^*+\mathrm{i}\infty} \mathrm{d}(-\mathrm{i}\zeta)\, \mathrm{e}^{-n[\zeta s - \lambda(\zeta)]} \asymp \mathrm{e}^{-n[\zeta^* s - \lambda(\zeta^*)]}. \tag{C.6}$$

The last approximation is the saddle-point approximation (see Chap. 6 of [20]). This result is completed by noting that the saddle-point ζ^* must be real, since $p(S_n = s)$ is real. Moreover, if we assume that $\lambda(\zeta)$ is analytic, then ζ^* is the unique minimum of $\zeta s - \lambda(\zeta)$ satisfying $\lambda'(\zeta^*) = s$ along the Bromwich contour. The analyticity of $\lambda(k)$ also implies, by the Cauchy–Riemann equations, that the point ζ^*, which is a minimum of $\zeta s - \lambda(\zeta)$ along the Bromwich contour, is a maximum of $\zeta s - \lambda(\zeta)$ for ζ real. Therefore, we can write

$$\lim_{n\to\infty} -\frac{1}{n} \ln p(S_n = s) = \sup_{k\in\mathbb{R}} \{ks - \lambda(k)\}. \tag{C.7}$$

This concludes our first derivation of the Gärtner–Ellis Theorem. For a discussion of cases for which $\lambda(k)$ is not analytic, see Section 4.

C.2. Exponential change of measure

We consider the same random variable $S_n(\omega)$ as in the previous derivation, but now we focus on the probability measure $P(\mathrm{d}\omega)$ instead of the probability density $p(\omega)$. We also introduce the following modification or "perturbation" of $P(\mathrm{d}\omega)$:

$$P_k(\mathrm{d}\omega) = \frac{\mathrm{e}^{nkS_n(\omega)}}{\left\langle \mathrm{e}^{nkS_n} \right\rangle} P(\mathrm{d}\omega), \tag{C.8}$$

which involves the parameter $k \in \mathbb{R}$. This probability has the same form as the probability $P_\beta(\mathrm{d}\omega)$ defining the canonical ensemble. In large deviation theory, $P_k(\mathrm{d}\omega)$ is called the *tilted measure*, and the family of such measures indexed by k is often called the *exponential family* [8].[26]

[26] In statistics and actuarial mathematics, P_k is also known as the *associated law* or *Esscher transform* of P [292].

Starting from the definition of $P_k(\mathrm{d}\omega)$, one can prove the following properties (see [8,17,18]):

Property 1: If

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{n} \ln \left\langle \mathrm{e}^{nkS_n} \right\rangle \tag{C.9}$$

exists, then

$$P_k(\mathrm{d}\omega) \asymp \mathrm{e}^{n[kS_n(\omega)-\lambda(k)]} P(\mathrm{d}\omega). \tag{C.10}$$

The so-called Radon–Nikodym derivative of $P_k(\mathrm{d}\omega)$ with respect to $P(\mathrm{d}\omega)$ is thus written as

$$\frac{\mathrm{d}P_k(\omega)}{\mathrm{d}P(\omega)} = \frac{P_k(\mathrm{d}\omega)}{P(\mathrm{d}\omega)} \asymp \mathrm{e}^{n[kS_n(\omega)-\lambda(k)]}. \tag{C.11}$$

Property 2: If $\lambda(k)$ is differentiable at k, then

$$\begin{aligned} \lim_{n\to\infty} \langle S_n \rangle_k &= \lim_{n\to\infty} \int_{\mathbb{R}^n} S_n(\omega)\, P_k(\mathrm{d}\omega) \\ &= \lim_{n\to\infty} \frac{1}{\left\langle \mathrm{e}^{nkS_n} \right\rangle} \int_{\mathbb{R}^n} S_n(\omega)\, \mathrm{e}^{nkS_n(\omega)}\, P(\mathrm{d}\omega) \\ &= \lambda'(k). \end{aligned} \tag{C.12}$$

Property 3: The value $s_k = \lambda'(k)$ is the concentration point (viz., typical value) of S_n with respect to $P_k(\mathrm{d}\omega)$, that is,

$$\lim_{n\to\infty} P_k(S_n \in [s_k, s_k + \mathrm{d}s]) = 1. \tag{C.13}$$

This limit expresses a Law of Large Numbers for S_n with respect to $P_k(\mathrm{d}\omega)$.[27]

From these properties, we obtain a large deviation principle for $P(S_n \in \mathrm{d}s)$ as follows. Starting with

$$P(S_n \in \mathrm{d}s) = \int_{\{\omega\in\mathbb{R}^n : S_n(\omega)\in \mathrm{d}s\}} P(\mathrm{d}\omega) = \int_{\{\omega\in\mathbb{R}^n : S_n(\omega)\in \mathrm{d}s\}} \frac{P(\mathrm{d}\omega)}{P_k(\mathrm{d}\omega)}\, P_k(\mathrm{d}\omega), \tag{C.14}$$

we use the first property to obtain

$$\begin{aligned} P(S_n \in \mathrm{d}s) &\asymp \int_{\{\omega\in\mathbb{R}^n : S_n(\omega)\in \mathrm{d}s\}} \mathrm{e}^{-n[kS_n(\omega)-\lambda(k)]}\, P_k(\mathrm{d}\omega) \\ &= \mathrm{e}^{-n[ks-\lambda(k)]} \int_{\{\omega\in\mathbb{R}^n : S_n(\omega)\in \mathrm{d}s\}} P_k(\mathrm{d}\omega), \end{aligned} \tag{C.15}$$

which implies

$$P(S_n \in \mathrm{d}s) \asymp \mathrm{e}^{-n[ks-\lambda(k)]}\, P_k(S_n \in \mathrm{d}s). \tag{C.16}$$

Next we choose k such that $\lambda'(k) = s$. According to the second and third properties, we must have

$$\lim_{n\to\infty} P_k(S_n \in [s, s + \mathrm{d}s]) = 1 \tag{C.17}$$

or, equivalently, $P_k(S_n \in \mathrm{d}s) \asymp \mathrm{e}^{n0}\, \mathrm{d}s$ using the asymptotic notation, so that

$$P(S_n \in \mathrm{d}s) \asymp \mathrm{e}^{-n[ks-\lambda(k)]}\, \mathrm{d}s. \tag{C.18}$$

Therefore, $P(S_n \in \mathrm{d}s) \asymp \mathrm{e}^{-nI(s)}$, where

$$I(s) = ks - \lambda(k), \quad \lambda'(k) = s. \tag{C.19}$$

We recognize in the last expression the Legendre transform of $\lambda(k)$.

This derivation can be adapted to other random variables and processes, and is useful in practice for deriving large deviation principles, as the Radon-Nikodym derivative can often be calculated explicitly. In the case of Markov processes, for example, $\mathrm{d}P_k/\mathrm{d}P$ is given by Girsanov's formula [38]. Other perturbations of P, apart from the exponential one, can also be used. The general idea at play is to change the measure (or process) P into a measure P', so as to make an unlikely event under P a typical event under P', and to use the relationship between P and P' to infer the probability of that event under P.

[27] Recall that the concentration point of S_n with respect to $P(\mathrm{d}\omega)$ is $s_0 = \lambda'(0)$; see Section 3.5.1.

Appendix D. Large deviation results for different speeds

The Gärtner–Ellis Theorem is stated and used throughout this review mostly for large deviation principles having a linear speed $a_n = n$. The following is the general version of that theorem which applies to any speed a_n such that $a_n \to \infty$ as $n \to \infty$ [8]. Consider a random variable W_n such that

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{a_n} \ln \left\langle \mathrm{e}^{a_n k W_n} \right\rangle \tag{D.1}$$

exists and is differentiable. Then $P(W_n \in \mathrm{d}w) \asymp \mathrm{e}^{-a_n I(w)}\, \mathrm{d}w$, where $I(w)$ is, as before, the Legendre–Fenchel transform of $\lambda(k)$.

The version of Varadhan's Theorem that applies to general speeds is the following [13]. Let W_n be a random variable satisfying a large deviation principle with speed a_n and rate function $I(w)$, and let f be a bounded function of W_n. Then

$$\lambda(f) = \lim_{n\to\infty} \frac{1}{a_n} \ln \left\langle \mathrm{e}^{a_n f(W_n)} \right\rangle = \sup_w \{f(w) - I(w)\}. \tag{D.2}$$

For the (unbounded) linear function $f(W_n) = kW_n$, it can also be proved, with an additional mild assumption on W_n (see, e.g., Theorem 5.1 of [10] or Theorem 4.3.1 of [13]), that

$$\lambda(k) = \lim_{n\to\infty} \frac{1}{a_n} \ln \left\langle \mathrm{e}^{a_n k W_n} \right\rangle = \sup_w \{kw - I(w)\}. \tag{D.3}$$

References

[1] H. Cramér, Sur un nouveau théorème limite dans la théorie des probabilités, in: Colloque consacré à la théorie des probabilités, vol. 3, Hermann, Paris, 1938, pp. 2–23.
[2] M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. I, Comm. Pure Appl. Math. 28 (1975) 1–47.
[3] M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. II, Comm. Pure Appl. Math. 28 (1975) 279–301.
[4] M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. III, Comm. Pure Appl. Math. 29 (4) (1976) 389–461.
[5] M.D. Donsker, S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. IV, Comm. Pure Appl. Math. 36 (2) (1983) 183–212.
[6] M.I. Freidlin, A.D. Wentzell, Random Perturbations of Dynamical Systems, in: Grundlehren der Mathematischen Wissenschaften, vol. 260, Springer-Verlag, New York, 1984.
[7] R.S. Ellis, The theory of large deviations: From Boltzmann's 1877 calculation to equilibrium macrostates in 2D turbulence, Physica D 133 (1999) 106–136.
[8] R.S. Ellis, Entropy, Large Deviations, and Statistical Mechanics, Springer, New York, 1985.
[9] Y. Oono, Large deviation and statistical physics, Progr. Theoret. Phys. Suppl. 99 (1989) 165–205.
[10] R.S. Ellis, An overview of the theory of large deviations and applications to statistical mechanics, Scand. Actuar. J. 1 (1995) 97–142.
[11] O.E. Lanford, Entropy and equilibrium states in classical statistical mechanics, in: A. Lenard (Ed.), Statistical Mechanics and Mathematical Problems, in: Lecture Notes in Physics, vol. 20, Springer, Berlin, 1973, pp. 1–113.
[12] G.L. Eyink, Dissipation and large thermodynamic fluctuations, J. Statist. Phys. 61 (3) (1990) 533–572.
[13] A. Dembo, O. Zeitouni, Large Deviations Techniques and Applications, 2nd ed., Springer, New York, 1998.
[14] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley, New York, 1991.
[15] P. Dupuis, R.S. Ellis, A Weak Convergence Approach to the Theory of Large Deviations, in: Wiley Series in Probability and Statistics, John Wiley, New York, 1997.
[16] M. Capiński, E. Kopp, Measure, Integral and Probability, in: Springer Undergraduate Mathematics Series, Springer, New York, 2005.
[17] J. Gärtner, On large deviations from the invariant measure, Theory Probab. Appl. 22 (1977) 24–39.
[18] R.S. Ellis, Large deviations for a general class of random vectors, Ann. Probab. 12 (1) (1984) 1–12.
[19] R.T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, 1970.
[20] C.M. Bender, S.A. Orszag, Advanced Mathematical Methods for Scientists and Engineers, McGraw-Hill, New York, 1978.
[21] J. van Tiel, Convex Analysis: An Introductory Text, John Wiley, New York, 1984.
[22] S.R.S. Varadhan, Asymptotic probabilities and differential equations, Comm. Pure Appl. Math. 19 (1966) 261–286.
[23] N. O'Connell, From laws of large numbers to large deviation principles, Markov Process. Related Fields 3 (4) (1997).
[24] N. O'Connell, A large deviations heuristic made precise, Math. Proc. Cambridge Philos. Soc. 128 (2000) 561–569.
[25] A. Martin-Löf, A Laplace approximation for sums of independent random variables, Z. Wahrscheinlichkeitstheor. Verwandte Geb. 59 (1) (1982) 101–115.
[26] W. Bryc, A remark on the connection between the large deviation principle and the central limit theorem, Statist. Probab. Lett. 18 (4) (1993) 253–256.
[27] P. Ney, Dominating points and the asymptotics of large deviations for random walk on $\mathbb{R}^d$, Ann. Probab. 11 (1) (1983) 158–167.
[28] J.A. Bucklew, Large Deviation Techniques in Decision, Simulation and Estimation, in: Wiley Series in Probability and Mathematical Statistics, Wiley Interscience, New York, 1990.
[29] I.N. Sanov, On the probability of large deviations of random variables, in: Select. Transl. Math. Statist. and Probability, vol. 1, Inst. Math. Statist. and Amer. Math. Soc., Providence, R.I., 1961, pp. 213–244.
[30] L. Boltzmann, Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respecktive den Sätzen über das Wärmegleichgewicht (On the relationship between the second law of the mechanical theory of heat and the probability calculus), Wiener Berichte 2 (76) (1877) 373–435.
[31] R.R. Bahadur, S.L. Zabell, Large deviations of the sample mean in general vector spaces, Ann. Probab. 7 (4) (1979) 587–621.
[32] D. Plachky, J. Steinebach, A theorem about probabilities of large deviations with an application to queuing theory, Period. Math. Hungar. 6 (1975) 343–345.
[33] H.E. Daniels, Saddlepoint approximations in statistics, Ann. Math. Stat. 25 (1954) 631–650.

[34] O.E. Barndorff-Nielsen, D.R. Cox, Asymptotic Techniques for Use in Statistics, in: Monographs on Statistics and Applied Probability, Chapman and Hall, London, 1989.
[35] R.W. Butler, Saddlepoint Approximations with Applications, Cambridge University Press, Cambridge, 2007.
[36] A. Amann, H. Atmanspacher, Introductory remarks on large deviation statistics, J. Sci. Exploration 13 (1999) 639–664.
[37] J.T. Lewis, R. Russell, An introduction to large deviations for teletraffic engineers, DIAS Report, 1996.
[38] S.R.S. Varadhan, Large deviations and entropy, in: A. Greven, G. Keller, G. Warnecke (Eds.), Entropy, in: Princeton Ser. Appl. Math., Princeton University Press, Princeton, 2003, pp. 199–214 (Chapter 9).
[39] J.D. Deuschel, D.W. Stroock, Large Deviations, Academic Press, Boston, 1989.
[40] F. den Hollander, Large Deviations, in: Fields Institute Monograph, Amer. Math. Soc., Providence, R.I., 2000.
[41] V.V. Uchaikin, V.M. Zolotarev, Chance and Stability: Stable Distributions and their Applications, VSP, Utrecht, 1999.
[42] A.V. Nagaev, Asymptotic properties of stable densities and the asymmetric large deviation problems, Statist. Probab. Lett. 61 (4) (2003) 429–438.
[43] A.V. Nagaev, Cramér's large deviations when the extreme conjugate distribution is heavy-tailed, Theory Probab. Appl. 43 (1999) 405–421.
[44] P. Ney, E. Nummelin, Markov additive processes I: Eigenvalue properties and limit theorems, Ann. Probab. 15 (1987) 561–592.
[45] P. Ney, E. Nummelin, Markov additive processes II: Large deviations, Ann. Probab. 15 (1987) 593–609.
[46] I.H. Dinwoodie, S.L. Zabell, Large deviations for exchangeable random vectors, Ann. Probab. 20 (1992) 1147–1166.
[47] I.H. Dinwoodie, Identifying a large deviation rate function, Ann. Probab. 21 (1993) 216–231.
[48] D. Ioffe, Two examples in the theory of large deviations, Statist. Probab. Lett. 18 (1993) 297–300.
[49] V. Lecomte, C. Appert-Rolland, F. van Wijland, Chaotic properties of systems with Markov dynamics, Phys. Rev. Lett. 95 (1) (2005) 010601.
[50] V. Lecomte, C. Appert-Rolland, F. van Wijland, Thermodynamic formalism for systems with Markov dynamics, J. Statist. Phys. (2007).
[51] C. Beck, F. Schlögl, Thermodynamics of Chaotic Systems: An Introduction, Cambridge University Press, Cambridge, 1993.
[52] P. Gaspard, Chaos, Scattering and Statistical Mechanics, in: Cambridge Nonlinear Science Series, vol. 9, Cambridge University Press, Cambridge, 1998.
[53] D. Ruelle, Statistical Mechanics: Rigorous Results, W.A. Benjamin, Amsterdam, 1969.
[54] J.T. Lewis, The large deviation principle in statistical mechanics: An expository account, in: A. Truman, I.M. Davies (Eds.), Stochastic Mechanics and Stochastic Processes, in: Lecture Notes in Mathematics, vol. 1325, Springer, New York, 1988, pp. 141–155.
[55] J.T. Lewis, The large deviation principle in statistical mechanics, in: Mark Kac Seminar on Probability and Physics, vol. 17, Math. Centrum Centrum Wisk. Inform., Amsterdam, 1988, pp. 85–102.
[56] J.T. Lewis, Large deviations and statistical mechanics, in: Mathematical Methods in Statistical Mechanics, in: Leuven Notes Math. Theoret. Phys. Ser. A Math. Phys., vol. 1, Leuven Univ. Press, Leuven, 1989, pp. 77–90.
[57] J.T. Lewis, C.-E. Pfister, W.G. Sullivan, Large deviations and the thermodynamic formalism: A new proof of the equivalence of ensembles, in: M. Fannes, C. Maes, A. Verbeure (Eds.), On Three Levels, Plenum Press, New York, 1994.
[58] J.T. Lewis, C.-E. Pfister, W.G. Sullivan, Entropy, concentration of probability and conditional limit theorem, Markov Process. Related Fields 1 (1995) 319–386.
[59] J.T. Lewis, C.-E. Pfister, Thermodynamic probability theory: Some aspects of large deviations, Russ. Math. Surveys 50 (2) (1995) 279–317.
[60] C.-E. Pfister, Large deviations and phase separation in the two-dimensional Ising model, Helv. Phys. Acta 64 (7) (1991) 953–1054.
[61] R. Balian, From Microphysics to Macrophysics: Methods and Applications of Statistical Physics, vol. I, Springer, Berlin, 1991.
[62] R.S. Ellis, K. Haven, B. Turkington, Large deviation principles and complete equivalence and nonequivalence results for pure and mixed ensembles, J. Statist. Phys. 101 (2000) 999–1064.
[63] A. Einstein, The theory of opalescence of homogeneous fluids and liquid mixtures near the critical state, in: J. Stachel (Ed.), in: The Collected Papers of Albert Einstein, vol. 3, Princeton University Press, Princeton, 1987, pp. 231–249.
[64] E.T. Jaynes, Information theory and statistical mechanics, Phys. Rev. 106 (1957) 620–630.
[65] E.T. Jaynes, Probability Theory: The Logic of Science, Vol. I, Cambridge University Press, Cambridge, 2003.
[66] T. Lehtonen, E. Nummelin, Level I theory of large deviations in the ideal gas, Internat. J. Theoret. Phys. 29 (1990) 621–635.
[67] T. Eisele, R.S. Ellis, Multiple phase transitions in the generalized Curie-Weiss model, J. Statist. Phys. 52 (1) (1988) 161–202.
[68] S. Orey, Large deviations for the empirical field of Curie-Weiss models, Stochastics 25 (1988) 3–14.
[69] R.S. Ellis, K. Wang, Limit theorems for the empirical vector of the Curie-Weiss-Potts model, Stochastic Process. Appl. 35 (1990) 59–79.
[70] M. Costeniuc, R.S. Ellis, H. Touchette, Complete analysis of phase transitions and ensemble equivalence for the Curie-Weiss-Potts model, J. Math. Phys. 46 (2005) 063301.
[71] J. Barré, D. Mukamel, S. Ruffo, Inequivalence of ensembles in a system with long-range interactions, Phys. Rev. Lett. 87 (2001) 030601.
[72] R.S. Ellis, H. Touchette, B. Turkington, Thermodynamic versus statistical nonequivalence of ensembles for the mean-field Blume–Emery–Griffiths model, Physica A 335 (2004) 518–538.
[73] R.S. Ellis, P. Otto, H. Touchette, Analysis of phase transitions in the mean-field Blume-Emery-Griffiths model, Ann. Appl. Probab. 15 (2005) 2203–2254.
[74] J. Barré, F. Bouchet, T. Dauxois, S. Ruffo, Large deviation techniques applied to systems with long-range interactions, J. Statist. Phys. 119 (3) (2005) 677–713.
[75] M. Kastner, O. Schnetz, On the mean-field spherical model, J. Statist. Phys. 122 (6) (2006) 1195–1214.
[76] L. Casetti, M. Kastner, Partial equivalence of statistical ensembles and kinetic energy, Physica A 384 (2) (2007) 318–334.
[77] I. Hahn, M. Kastner, The mean-field ϕ^4 model: Entropy, analyticity, and configuration space topology, Phys. Rev. E 72 (2005) 056134.
[78] I. Hahn, M. Kastner, Application of large deviation theory to the mean-field ϕ^4-model, Eur. Phys. J. B 50 (2006) 311–314.
[79] A. Campa, S. Ruffo, H. Touchette, Negative magnetic susceptibility and nonequivalent ensembles for the mean-field ϕ^4 spin model, Physica A 385 (1) (2007) 233–248.
[80] T. Dauxois, S. Ruffo, E. Arimondo, M. Wilkens (Eds.), Dynamics and Thermodynamics of Systems with Long Range Interactions, vol. 602, Springer, New York, 2002.
[81] F. Bouchet, J. Barré, Classification of phase transitions and ensemble inequivalence in systems with long range interactions, J. Statist. Phys. 118 (5) (2005) 1073–1105.
[82] A. Campa, A. Giansanti, G. Morigi, F.S. Labini (Eds.), Dynamics and Thermodynamics of Systems with Long-range Interactions: Theory and Experiments, AIP, Melville, NY, 2008.
[83] A. Campa, T. Dauxois, S. Ruffo, Statistical mechanics and dynamics of solvable models with long-range interactions, 2008.
[84] R.S. Ellis, K. Haven, B. Turkington, Nonequivalent statistical equilibrium ensembles and refined stability theorems for most probable flows, Nonlinearity 15 (2002) 239–255.
[85] C.-E. Pfister, Thermodynamical aspects of classical lattice systems, in: V. Sidoravicius (Ed.), In and Out of Equilibrium, Probability with a Physics Flavor, Birkhäuser, Boston, 2002, pp. 393–472.
[86] M. Kastner, Existence and order of the phase transition of the Ising model with fixed magnetization, J. Statist. Phys. 109 (1) (2002) 133–142.
[87] J.T. Lewis, C.-E. Pfister, G.W. Sullivan, The equivalence of ensembles for lattice systems: Some examples and a counterexample, J. Statist. Phys. 77 (1994) 397–419.
[88] L. Onsager, Crystal statistics. I. A two-dimensional model with an order-disorder transition, Phys. Rev. 65 (3–4) (1944) 117–149.
[89] D. Ioffe, Large deviations for the 2D Ising model: A lower bound without cluster expansions, J. Statist. Phys. 74 (1) (1994) 411–432.
[90] L.D. Landau, E.M. Lifshitz, Statistical Physics, 3rd ed., in: Landau and Lifshitz Course of Theoretical Physics, vol. 5, Butterworth Heinemann, Oxford, 1991.
[91] H. Touchette, Comment on First-order phase transition: Equivalence between bimodalities and the Yang–Lee theorem, Physica A 359 (2006) 375–379.
[92] J.C. Maxwell, On the dynamical evidence of the molecular constitution of bodies, Nature II (1875) 357.
[93] K. Huang, Statistical Mechanics, Wiley, New York, 1987.

[94] I. Ispolatov, E.G.D. Cohen, On first-order phase transitions in microcanonical and canonical non-extensive systems, Physica A 295 (2000) 475–487.
[95] M.K.-H. Kiessling, T. Neukirch, Negative specific heat of a magnetically self-confined plasma torus, Proc. Natl. Acad. Sci. USA 100 (4) (2003) 1510–1514.
[96] M.K.-H. Kiessling, J. Lebowitz, The micro-canonical point vortex ensemble: Beyond equivalence, Lett. Math. Phys. 42 (1997) 43–56.
[97] P.-H. Chavanis, Phase transitions in self-gravitating systems, Internat. J. Modern Phys. B 20 (22) (2006) 3113–3198.
[98] D. Lynden-Bell, Negative specific heat in astronomy, physics and chemistry, Physica A 263 (1999) 293–304.
[99] J.L. Lebowitz, Boltzmann's entropy and time's arrow, Phys. Today 46 (1993) 32–38.
[100] G.L. Eyink, H. Spohn, Negative-temperature states and large-scale, long-lived vortices in two-dimensional turbulence, J. Statist. Phys. 70 (1993) 833–886.
[101] H. Touchette, Simple spin models with non-concave entropies, Amer. J. Phys. 76 (2008) 26–30.
[102] H. Touchette, R.S. Ellis, B. Turkington, An introduction to the thermodynamic and macrostate levels of nonequivalent ensembles, Physica A 340 (2004) 138–146.
[103] D. Ruelle, Thermodynamic Formalism, 2nd ed., Cambridge University Press, Cambridge, 2004.
[104] R.B. Griffiths, Microcanonical ensemble in quantum statistical mechanics, J. Math. Phys. 6 (10) (1965) 1447–1461.
[105] L. Galgani, L. Manzoni, A. Scotti, Asymptotic equivalence of equilibrium ensembles of classical statistical mechanics, J. Math. Phys. 12 (6) (1971) 933–935.
[106] G. Gallavotti, Statistical Mechanics: A Short Treatise, Springer, New York, 1999.
[107] L. van Hove, Quelques propriétés générales de l'intégrale de configuration d'un système de particules avec interaction, Physica 15 (11–12) (1949) 951–961.
[108] L. van Hove, Sur l'intégrale de configuration pour les systèmes de particules à une dimension, Physica 16 (2) (1950) 137–143.
[109] D. Ruelle, Classical statistical mechanics of a system of particles, Helv. Phys. Acta 36 (1963) 183–197.
[110] M.E. Fisher, The free energy of a macroscopic system, Arch. Ration. Mech. Anal. 17 (5) (1964) 377–410.
[111] R.B. Griffiths, A proof that the free energy of a spin system is extensive, J. Math. Phys. 5 (9) (1964) 1215–1222.
[112] R.B. Griffiths, Rigorous results and theorems, in: C. Domb, M.S. Green (Eds.), in: Phase Transitions and Critical Phenomena, vol. 1, Academic Press, London, 1972, pp. 7–109.
[113] B. Derrida, Non-equilibrium steady states: Fluctuations and large deviations of the density and of the current, J. Stat. Mech. 2007 (07) (2007) P07023.
[114] C.W. Gardiner, Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences, 2nd ed., in: Springer Series in Synergetics, vol. 13, Springer, New York, 1985.
[115] A.D. Wentzell, M.I. Freidlin, On small random perturbations of dynamical systems, Russ. Math. Surveys 25 (1970) 1–55.
[116] R. Graham, Statistical theory of instabilities in stationary nonequilibrium systems with applications to lasers and nonlinear optics, in: G. Höhler (Ed.), Quantum Statistics in Optics and Solid-State Physics, in: Springer Tracts in Modern Physics, vol. 66, Springer, Berlin, 1973, pp. 1–97.
[117] R. Graham, T. Tél, Weak-noise limit of Fokker-Planck models and nondifferentiable potentials for dissipative dynamical systems, Phys. Rev. A 31 (2) (1985) 1109–1122.
[118] M.I. Dykman, M.A. Krivoglaz, Theory of fluctuational transitions between stable states of nonlinear oscillators, Sov. Phys. JETP 50 (1) (1979) 30–37.
[119] L. Onsager, S. Machlup, Fluctuations and irreversible processes, Phys. Rev. 91 (6) (1953) 1505–1512.
[120] D. Falkoff, Integral over path formulation of statistical theory of irreversible processes, Progr. Theoret. Phys. 16 (5) (1956) 530–532.
[121] D. Falkoff, Statistical theory of irrversible processes: Part I. Intergral over fluctuation path formulation, Ann. Phys. (N.Y.) 4 (3) (1958) 325–346.
[122] R.L. Stratonovich, Some Markov methods in the theory of stochastic processes in nonlinear dynamical systems, in: F. Moss, P.V.E. McClintock (Eds.), in: Noise in Nonlinear Dynamical Systems, vol. 1, Cambridge University Press, Cambridge, 1989, pp. 16–72.
[123] D. Dürr, A. Bach, The Onsager-Machlup function as Lagrangian for the most probable path of a diffusion process, Comm. Math. Phys. 60 (2) (1978) 153–170.
[124] W. Horsthemke, A. Bach, Onsager-Machlup function for one dimensional nonlinear diffusion processes, Z. Phys. B (Condens. Matter) 22 (2) (1975) 189–192.
[125] K.L.C. Hunt, J. Ross, Path integral solutions of stochastic equations for nonlinear irreversible processes: The uniqueness of the thermodynamic lagrangian, J. Chem. Phys. 75 (2) (1981) 976–984.
[126] A.J. Bray, A.J. McKane, T.J. Newman, Path integrals and non-Markov processes. II. Escape rates and stationary distributions in the weak-noise limit, Phys. Rev. A 41 (2) (1990) 657–667.
[127] A.J. McKane, H.C. Luckock, A.J. Bray, Path integrals and non-Markov processes. I. General formalism, Phys. Rev. A 41 (2) (1990) 644–656.
[128] H.S. Wio, P. Colet, M.S. Miguel, L. Pesquera, M.A. Rodríguez, Path-integral formulation for stochastic processes driven by colored noise, Phys. Rev. A 40 (12) (1989) 7312–7324.
[129] M.I. Dykman, K. Lindenberg, Fluctuations in nonlinear systems driven by colored noise, in: G.H. Weiss (Ed.), Contemporary Problems in Statistical Physics, SIAM, Philadelphia, 1994, pp. 41–101.
[130] S.J.B. Einchcomb, A.J. McKane, Use of Hamiltonian mechanics in systems driven by colored noise, Phys. Rev. E 51 (4) (1995) 2974–2981.
[131] F. Moss, P.V.E. McClintock (Eds.), Noise in Nonlinear Dynamical Systems, vol. 1, Cambridge University Press, Cambridge, 1989.
[132] F. Moss, P.V.E. McClintock (Eds.), Noise in Nonlinear Dynamical Systems, vol. 2, Cambridge University Press, Cambridge, 1989.
[133] H. Kleinert, Path Integrals in Quantum Mechanics, Statistics, Polymer Physics, and Financial Markets, World Scientific, Singapore, 2004.
[134] F.W. Wiegel, Introduction to Path-integral Methods in Physics and Polymer Science, World Scientific, Singapore, 1986.
[135] M. Schilder, Some asymptotic formulae for Wiener integrals, Trans. Amer. Math. Soc. 125 (1966) 63–85.
[136] R.V. Roy, Noise perturbation of nonlinear dynamical systems, in: A.H.-D. Cheng, C.Y. Yang (Eds.), Computational Stochastic Mechanics, Elsevier, Amsterdam, 1993, pp. 125–148.
[137] R.V. Roy, Large deviation theory, weak-noise asymptotics, and first-passage problems: Review and results, in: M. Lemaire, J.-L. Favre, A. Mebarki (Eds.), Applications of Statistics and Probability, A.A. Balkema, Rotterdam, 1995, pp. 1129–1135.
[138] M.I. Dykman, E. Mori, J. Ross, P.M. Hunt, Large fluctuations and optimal paths in chemical kinetics, J. Chem. Phys. 100 (8) (1994) 5735–5750.
[139] A.J. Bray, A.J. McKane, Instanton calculation of the escape rate for activation over a potential barrier driven by colored noise, Phys. Rev. Lett. 62 (5) (1989) 493–496.
[140] L. Bertini, A.D. Sole, D. Gabrielli, G. Jona-Lasinio, C. Landim, Macroscopic fluctuation theory for stationary non-equilibrium states, J. Statist. Phys. 107 (3) (2002) 635–675.
[141] N.G. van Kampen, Stochastic Processes in Physics and Chemistry, North-Holland, Amsterdam, 1992.
[142] R.L. Kautz, Thermally induced escape: The principle of minimum available noise energy, Phys. Rev. A 38 (4) (1988) 2066–2080.
[143] R.L. Kautz, Activation energy for thermally induced escape from a basin of attraction, Phys. Lett. A 125 (6–7) (1987) 315–319.
[144] R. Graham, T. Tél, Nonequilibrium potential for coexisting attractors, Phys. Rev. A 33 (2) (1986) 1322–1337.
[145] H.R. Jauslin, Nondifferentiable potentials for nonequilibrium steady states, Physica A 144 (1) (1987) 179–191.
[146] M.I. Dykman, P.V.E. McClintock, V.N. Smelyanski, N.D. Stein, N.G. Stocks, Optimal paths and the prehistory problem for large fluctuations in noise-driven systems, Phys. Rev. Lett. 68 (18) (1992) 2718–2721.
[147] M.I. Dykman, D.G. Luchinsky, P.V.E. McClintock, V.N. Smelyanskiy, Corrals and critical behavior of the distribution of fluctuational paths, Phys. Rev. Lett. 77 (26) (1996) 5229–5232.
[148] D.G. Luchinsky, P.V.E. McClintock, Irreversibility of classical fluctuations studied in analogue electrical circuits, Nature 389 (1997) 463–466.
[149] D.G. Luchinsky, R.S. Maier, R. Mannella, P.V.E. McClintock, D.L. Stein, Experiments on critical phenomena in a noisy exit problem, Phys. Rev. Lett. 79 (17) (1997) 3109–3112.
[150] D.G. Luchinsky, P.V.E. McClintock, M.I. Dykman, Analogue studies of nonlinear systems, Rep. Progr. Phys. 61 (8) (1998) 889–997.

[151] E. Olivieri, Metastability and entropy, in: A. Greven, G. Keller, G. Warnecke (Eds.), Entropy, in: Princeton Ser. Appl. Math., Princeton University Press, Princeton, 2003, pp. 233–250 (Chapter 11).
[152] E. Olivieri, M.E. Vares, Large Deviations and Metastability, in: Encyclopedia of Mathematics and Its Applications, vol. 100, Cambridge University Press, Cambridge, 2005.
[153] H. Hasegawa, Variational principle for non-equilibrium states and the Onsager-Machlup formula, Progr. Theoret. Phys. 56 (1) (1976) 44–60.
[154] T. Taniguchi, E.G.D. Cohen, Onsager-Machlup theory for nonequilibrium steady states and fluctuation theorems, J. Statist. Phys. 126 (1) (2007) 1–41.
[155] T. Taniguchi, E.G.D. Cohen, Nonequilibrium steady state thermodynamics and fluctuations for stochastic systems, J. Statist. Phys. 130 (4) (2008) 633–667.
[156] R. Graham, Onset of cooperative behavior in nonequilibrium steady states, in: G. Nicolis, G. Dewel, J.W. Turner (Eds.), Order and Fluctuations in Equilibrium and Nonequilibrium Statistical Mechanics, Wiley, New York, 1981.
[157] M. Paniconi, Y. Oono, Phenomenological framework for fluctuations around steady state, Phys. Rev. E 55 (1) (1997) 176–188.
[158] Y. Oono, M. Paniconi, Steady state thermodynamics, Progr. Theoret. Phys. Suppl. 130 (1998) 29–44.
[159] A. Suarez, J. Ross, B. Peng, K.L.C. Hunt, P.M. Hunt, Thermodynamic and stochastic theory of nonequilibrium systems: A Lagrangian approach to fluctuations and relation to excess work, J. Chem. Phys. 102 (11) (1995) 4563–4573.
[160] G. Falkovich, K. Gawedzki, M. Vergassola, Particles and fields in fluid turbulence, Rev. Modern Phys. 73 (4) (2001) 913–975.
[161] R. Chetrite, J.-Y. Delannoyand, K. Gawedzki, Kraichnan flow in a square: An example of integrable chaos, J. Statist. Phys. 126 (6) (2007) 1165–1200.
[162] M. Gourcy, A large deviation principle for 2D stochastic Navier-Stokes equation, Stochastic Process. Appl. 117 (7) (2007) 904–927.
[163] H. Spohn, Large Scale Dynamics of Interacting Particles, Springer Verlag, Heidelberg, 1991.
[164] C. Kipnis, C. Landim, Scaling Limits of Interacting Particle Systems, in: Grundlheren der mathematischen Wissenschaften, vol. 320, Springer-Verlag, Berlin, 1999.
[165] S.N. Majumdar, A.J. Bray, Large-deviation functions for nonlinear functionals of a Gaussian stationary Markov process, Phys. Rev. E 65 (5) (2002) 051112.
[166] R. van Zon, E.G.D. Cohen, Stationary and transient work-fluctuation theorems for a dragged Brownian particle, Phys. Rev. E 67 (2003) 046102.
[167] H. Touchette, E.G.D. Cohen, Fluctuation relation for a Lévy particle, Phys. Rev. E 76 (2) (2007) 020101.
[168] C. Maes, K. Netočný, B. Shergelashvili, A selection of nonequilibrium issues, in: Lecture notes from the 5th Prague Summer School on Mathematical Statistical Mechanics, 2006.
[169] R.J. Harris, G.M. Schütz, Fluctuation theorems for stochastic dynamics, J. Stat. Mech. 2007 (07) (2007) P07020.
[170] J. Kurchan, Non-equilibrium work relations, J. Stat. Mech. 2007 (07) (2007) P07005.
[171] U.M.B. Marconi, A. Puglisi, L. Rondoni, A. Vulpiani, Fluctuation-dissipation: Response theory in statistical physics, Phys. Rep. 461 (4–6) (2008) 111–195.
[172] F. Ritort, Work fluctuations, transient violations of the second law and free-energy recovery methods: Perspectives in theory and experiments, in: J. Dalibard, B. Duplantier, V. Rivasseau (Eds.), Poincaré Seminar 2003, in: Progress in Mathematical Physics, vol. 38, Birkhäuser Verlag, Basel, 2003, pp. 192–229.
[173] D.J. Evans, D.J. Searles, The fluctuation theorem, Adv. Phys. 51 (7) (2002) 1529–1585.
[174] G.N. Bochkov, Y.E. Kuzovlev, Nonlinear fluctuation-dissipation relations and stochastic models in nonequilibrium thermodynamics : I. Generalized fluctuation-dissipation theorem, Physica A 106 (3) (1981) 443–479.
[175] C. Maes, K. Netočný, Time-reversal and entropy, J. Statist. Phys. 110 (1) (2003) 269–310.
[176] G. Gallavotti, E.G.D. Cohen, Dynamical ensembles in nonequilibrium statistical mechanics, Phys. Rev. Lett. 74 (14) (1995) 2694–2697.
[177] G. Gallavotti, E.G.D. Cohen, Dynamical ensembles in stationary states, J. Statist. Phys. 80 (5) (1995) 931–970.
[178] D.J. Evans, E.G.D. Cohen, G.P. Morriss, Probability of second law violations in shearing steady states, Phys. Rev. Lett. 71 (15) (1993) 2401–2404.
[179] J. Kurchan, Fluctuation theorem for stochastic dynamics, J. Phys. A: Math. Gen. 31 (1998) 3719–3729.
[180] J.L. Lebowitz, H. Spohn, A Gallavotti-Cohen-type symmetry in the large deviation functional for stochastic dynamics, J. Statist. Phys. 95 (1999) 333–365.
[181] C. Maes, The fluctuation Theorem as a Gibbs property, J. Statist. Phys. 95 (1999) 367–392.
[182] G.M. Wang, E.M. Sevick, E. Mittag, D.J. Searles, D.J. Evans, Experimental demonstration of violations of the second law of thermodynamics for small systems and short time scales, Phys. Rev. Lett. 89 (5) (2002) 050601.
[183] D. Andrieux, P. Gaspard, S. Ciliberto, N. Garnier, S. Joubaud, A. Petrosyan, Entropy production and time asymmetry in nonequilibrium fluctuations, Phys. Rev. Lett. 98 (15) (2007) 150601.
[184] R. van Zon, S. Ciliberto, E.G.D. Cohen, Power and heat fluctuation theorems for electric circuits, Phys. Rev. Lett. 92 (13) (2004) 130601.
[185] N. Garnier, S. Ciliberto, Nonequilibrium fluctuations in a resistor, Phys. Rev. E 71 (6) (2005) 060101.
[186] S. Aumaître, S. Fauve, S. McNamara, P. Poggi, Power injected in dissipative systems and the fluctuation theorem, Eur. Phys. J. B 19 (3) (2001) 449–460.
[187] K. Feitosa, N. Menon, Fluidized granular medium as an instance of the fluctuation theorem, Phys. Rev. Lett. 92 (16) (2004) 164301.
[188] A. Puglisi, P. Visco, A. Barrat, E. Trizac, F. van Wijland, Fluctuations of internal energy flow in a vibrated granular gas, Phys. Rev. Lett. 95 (11) (2005) 110202.
[189] P. Visco, A. Puglisi, A. Barrat, E. Trizac, F. van Wijland, Injected power and entropy flow in a heated granular gas, Europhys. Lett. 72 (1) (2005) 55–61.
[190] P. Visco, A. Puglisi, A. Barrat, E. Trizac, F. van Wijland, Fluctuations of power injection in randomly driven granular gases, J. Statist. Phys. 125 (2006) 533–568.
[191] S. Ciliberto, S. Laroche, An experimental test of the Gallavotti-Cohen fluctuation theorem, J. Phys. IV (France) 8 (6) (1998) 215–219.
[192] S. Ciliberto, N. Garnier, S. Hernandez, C. Lacpatia, J.-F. Pinton, G.R. Chavarria, Experimental test of the Gallavotti-Cohen fluctuation theorem in turbulent flows, Physica A 340 (1-3) (2004) 240–250.
[193] B. Cleuren, C. Van den Broeck, R. Kawai, Fluctuation theorem for the effusion of an ideal gas, Phys. Rev. E 74 (2006) 021117.
[194] P. Gaspard, Time-reversed dynamical entropy and irreversibility in Markovian random processes, J. Statist. Phys. 117 (3) (2004) 599–615.
[195] R. Chetrite, K. Gawedzki, Fluctuation relations for diffusion processes, Comm. Math. Phys. 282 (2) (2008) 469–518.
[196] R. Chetrite, G. Falkovich, K. Gawedzki, Fluctuation relations in simple examples of non-equilibrium steady states, J. Stat. Mech. 2008 (08) (2008) P08005.
[197] C. Maes, K. Netočný, Minimum entropy production principle from a dynamical fluctuation law, J. Math. Phys. 48 (2007) 053306.
[198] D.-Q. Jiang, M. Qian, F.-X. Zhang, Entropy production fluctuations of finite Markov chains, J. Math. Phys. A: Math. Gen. 44 (9) (2003) 4176–4188.
[199] C. Maes, F. Redig, A. Van Moffaert, On the definition of entropy production, via examples, J. Math. Phys. 41 (3) (2000) 1528–1554.
[200] C. Maes, On the origin and the use of fluctuation relations for the entropy, Sem. Poincaré 2 (2003) 29–62.
[201] U. Seifert, Entropy production along a stochastic trajectory and an integral fluctuation theorem, Phys. Rev. Lett. 95 (4) (2005) 040602.
[202] A. Imparato, L. Peliti, Fluctuation relations for a driven Brownian particle, Phys. Rev. E 74 (2) (2006) 026106.
[203] G. Gallavotti, Breakdown and regeneration of time reversal symmetry in nonequilibrium statistical mechanics, Physica D 112 (1–2) (1998) 250–257.
[204] G. Gallavotti, Heat and fluctuations from order to chaos, Eur. Phys. J. B 61 (1) (2008) 1–24.
[205] R. van Zon, E.G.D. Cohen, Extension of the fluctuation theorem, Phys. Rev. Lett. 91 (2003) 110601.
[206] R. van Zon, E.G.D. Cohen, Extended heat-fluctuation theorems for a system with deterministic and stochastic forces, Phys. Rev. E 69 (2004) 056121.
[207] J. Farago, Injected power fluctuations in Langevin equation, J. Statist. Phys. 107 (2002) 781–803.
[208] A. Puglisi, L. Rondoni, A. Vulpiani, Relevance of initial and final conditions for the fluctuation relation in Markov processes, J. Stat. Mech. 2006 (08) (2006) P08010.
[209] P. Visco, Work fluctuations for a Brownian particle between two thermostats, J. Stat. Mech. 2006 (06) (2006) P06006.
[210] F. Bonetto, G. Gallavotti, A. Giuliani, F. Zamponi, Chaotic hypothesis, fluctuation theorem and singularities, J. Phys. Stat. 123 (2006) 39–54.

[211] R.J. Harris, A. Rákos, G.M. Schütz, Breakdown of Gallavotti-Cohen symmetry for stochastic dynamics, Europhys. Lett. 75 (2006) 227–233.
[212] A. Rákos, R.J. Harris, On the range of validity of the fluctuation theorem for stochastic Markovian dynamics, J. Stat. Mech. 2008 (05) (2008) P05005.
[213] T.M. Liggett, Interacting Particle Systems, in: Classics in Mathematics, Springer, New York, 2004.
[214] L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio, C. Landim, Minimum dissipation principle in stationary non-equilibrium states, J. Statist. Phys. 116 (1) (2004) 831–841.
[215] L. Bertini, A.D. Sole, D. Gabrielli, G. Jona-Lasinio, C. Landim, Stochastic interacting particle systems out of equilibrium, J. Stat. Mech. 2007 (07) (2007) P07014.
[216] C. Kipnis, S. Olla, S.R.S. Varadhan, Hydrodynamics and large deviation for simple exclusion processes, Comm. Pure Appl. Math. 42 (2) (1989) 115–137.
[217] B. Derrida, J.L. Lebowitz, Exact large deviation function in the asymmetric exclusion process, Phys. Rev. Lett. 80 (2) (1998) 209–213.
[218] B. Derrida, An exactly soluble non-equilibrium system: The asymmetric simple exclusion process, Phys. Rep. 301 (1–3) (1998) 65–83.
[219] O. Benois, C. Kipnis, C. Landim, Large deviations from the hydrodynamical limit of mean zero asymmetric zero range processes, Stochastic Process. Appl. 55 (1) (1995) 65–89.
[220] C. Landim, Hydrodynamical limit for mean zero asymmetric zero range processes, in: N. Boccara, E. Goles, S. Martinez, P. Picco (Eds.), Cellular Automata and Cooperative Systems, Kluwer, Boston, 1993.
[221] C. Kipnis, S. Olla, Large deviations from the hydrodynamical limit for a system of independent brownian particles, Stochastics 33 (1) (1990) 17–25.
[222] C. Kipnis, C. Léonard, Grandes déviations pour un système hydrodynamique asymétrique de particules indépendantes, Ann. Inst. Poincaré B 31 (1) (1995) 223–248.
[223] C. Landim, An overview on large deviations of the empirical measure of interacting particle systems, Ann. Inst. Poincaré A 55 (2) (1991) 615–635.
[224] B. Derrida, J.L. Lebowitz, E.R. Speer, Free energy functional for nonequilibrium systems: An exactly solvable case, Phys. Rev. Lett. 87 (15) (2001) 150601.
[225] B. Derrida, J.L. Lebowitz, E.R. Speer, Large deviation of the density profile in the steady state of the open symmetric simple exclusion process, J. Statist. Phys. 107 (3) (2002) 599–634.
[226] B. Derrida, J.L. Lebowitz, E.R. Speer, Exact free energy functional for a driven diffusive open stationary nonequilibrium system, Phys. Rev. Lett. 89 (3) (2002) 030601.
[227] B. Derrida, J.L. Lebowitz, E.R. Speer, Exact large deviation functional of a stationary open driven diffusive system: The asymmetric exclusion process, J. Statist. Phys. 110 (3) (2003) 775–810.
[228] T. Bodineau, B. Derrida, Current fluctuations in nonequilibrium diffusive systems: An additivity principle, Phys. Rev. Lett. 92 (18) (2004) 180601.
[229] T. Bodineau, B. Derrida, Distribution of current in nonequilibrium diffusive systems and phase transitions, Phys. Rev. E 72 (6) (2005) 066110.
[230] T. Bodineau, B. Derrida, Current large deviations for asymmetric exclusion processes with open boundaries, J. Statist. Phys. 123 (2) (2006) 277–300.
[231] T. Bodineau, G. Giacomin, From dynamic to static large deviations in boundary driven exclusion particle systems, Stochastic Process. Appl. 110 (1) (2004) 67–81.
[232] G. Paladin, A. Vulpiani, Anomalous scaling laws in multifractal objects, Phys. Rep. 156 (1987) 147–225.
[233] J.L. McCauley, Introduction to multifractals in dynamical systems theory and fully developed fluid turbulence, Phys. Rep. 189 (5) (1990) 225–266.
[234] K. Falconer, Techniques in Fractal Geometry, Wiley, New York, 1997.
[235] G. Zohar, Large deviations formalism for multifractals, Stochastic Process. Appl. 79 (1999) 229–242.
[236] D. Veneziano, Large deviations of multifractal measures, Fractals 10 (2002) 117–129.
[237] D. Harte, Multifractals: Theory and Applications, CRC Press, New York, 2001.
[238] H. Tominaga, H. Hata, T. Horita, H. Mori, K. Tomita, Linearities of the $f(\alpha)$ spectrum at bifurcations of chaos in dissipative differential systems, Progr. Theoret. Phys. 84 (1990) 18–22.
[239] H. Hata, T. Horita, H. Mori, T. Morita, K. Tomita, Singular local structures of chaotic attractors and q-phase transitions of spatial scaling structures, Progr. Theoret. Phys. 81 (1989) 11–16.
[240] H. Touchette, C. Beck, Nonconcave entropies in multifractals and the thermodynamic formalism, J. Statist. Phys. 125 (2006) 455–471.
[241] R. Riedi, An improved multifractal formalism and self-similar measures;, J. Math. Anal. Appl. 189 (2) (1995) 462–490.
[242] B. Testud, Transitions de phase dans l'analyse multifractale de mesures auto-similaires, C.R. Acad. Sci. Paris Ser. I 340 (9) (2005) 653–658.
[243] B. Testud, Phase transitions for the multifractal analysis of self-similar measures, Nonlinearity 19 (5) (2006) 1201–1217.
[244] Y. Kifer, Random perturbations of dynamical systems, in: Progress in Probability and Statistics, vol. 16, Birkhäuser, Boston, 1988.
[245] Y. Kifer, Large deviations in dynamical systems and stochastic processes, Trans. Amer. Math. Soc. 321 (2) (1990) 505–524.
[246] V.M. Alekseev, M.V. Yakobson, Symbolic dynamics and hyperbolic dynamic systems, Phys. Rep. 75 (1981) 290–325.
[247] J.-P. Eckmann, D. Ruelle, Ergodic theory of chaos and strange attractors, Rev. Modern Phys. 57 (1985) 617–656.
[248] A. Lasota, M.C. Mackey, Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics, in: Applied Mathematical Sciences, vol. 97, Springer, New York, 1994.
[249] D. Ruelle, The thermodynamic formalism for expanding maps, Comm. Math. Phys. 125 (2) (1989) 239–262.
[250] Y.G. Sinai, Gibbs measures in ergodic theory, Russ. Math. Surveys 27 (4) (1972) 21–69.
[251] Y.G. Sinai, Topics in Ergodic Theory, Princeton University Press, Princeton, 1994.
[252] G. Keller, Equilibrium States in Ergodic Theory, in: London Math. Soc. Student Texts, vol. 42, Cambridge University Press, Cambridge, 1998.
[253] T. Kai, K. Tomita, Statistical mechanics of deterministic chaos: The case of one-dimensional discrete process, Progr. Theoret. Phys. 64 (1980) 1532–1550.
[254] H. Mori, H. Hata, T. Horita, T. Kobayashi, Statistical mechanics of dynamical systems, Progr. Theoret. Phys. Suppl. 99 (1989) 1–63.
[255] Y. Takahashi, Y. Oono, Towards the statistical mechanics of chaos, Progr. Theoret. Phys. 71 (1984) 851–854.
[256] L.-S. Young, Some large deviation results for dynamical systems, Trans. Amer. Math. Soc. 318 (2) (1990) 525–543.
[257] A.O. Lopes, Entropy and large deviation, Nonlinearity 3 (1990) 527–546.
[258] S. Waddington, Large deviation asymptotics for Anosov flows, Ann. Inst. Poincaré C 13 (4) (1996) 445–484.
[259] M. Pollicott, R. Sharp, M. Yuri, Large deviations for maps with indifferent fixed points, Nonlinearity 11 (1998) 1173–1184.
[260] L.-S. Young, Entropy in dynamical systems, in: A. Greven, G. Keller, G. Warnecke (Eds.), Entropy, in: Princeton Ser. Appl. Math., Princeton University Press, Princeton, 2003, pp. 313–327 (Chapter 16).
[261] C. Maes, E. Verbitskiy, Large deviations and a fluctuation symmetry for chaotic homeomorphisms, Comm. Math. Phys. 233 (2003) 137–151.
[262] J. Tailleur, Grandes déviations, physique statistique et systèmes dynamiques, Ph.D. Thesis, Université Pierre et Marie Curie, Paris, 2007.
[263] Y. Oono, A heuristic approach to the Kolmogorov entropy as a disorder parameter, Progr. Theoret. Phys. 60 (1978) 1944–1946.
[264] Y. Oono, Y. Takahashi, Chaos, external noise and Fredholm theory, Progr. Theoret. Phys. 63 (1980) 1804–1807.
[265] N. Gantert, O. Zeitouni, Large deviations for one-dimentional random walk in a random environment — A survey, in: P. Revesz, B. Toth (Eds.), Proceedings of the Conference on Random Walks, in: Bolyai Society Mathematical Studies, vol. 9, 1999, pp. 127–165.
[266] F. Comets, N. Gantert, O. Zeitouni, Quenched, annealed and functional large deviations for one-dimensional random walk in random environment, Probab. Theory Related Fields 118 (1) (2000) 65–114.
[267] S.R.S. Varadhan, Large deviations for random walks in a random environment, Comm. Pure Appl. Math. 56 (8) (2003) 1222–1245.
[268] O. Zeitouni, Random walks in random environments, J. Phys. A: Math. Gen. 39 (40) (2006) R433–R464.
[269] T.C. Dorlas, J.R. Wedagedera, Large deviations and the random energy model, Internat. J. Modern Phys. B 15 (2001) 1–15.
[270] T.C. Dorlas, W.M.B. Dukes, Large deviation approach to the generalized random energy model, J. Phys. A: Math. Gen. 35 (20) (2002) 4385–4394.
[271] M. Talagrand, Large deviations, Guerra's and A.S.S. schemes, and the Parisi hypothesis, J. Statist. Phys. 126 (4) (2007) 837–894.
[272] M. Talagrand, Spin Glasses: A Challenge for Mathematicians, Springer, New York, 2003.
[273] A. Bovier, Statistical Mechanics of Disordered Systems: A Mathematical Perspective, Cambridge University Press, Cambridge, 2006.

[274] M. Mézard, A. Montanari, Information, Physics, and Computation, Oxford University Press, 2009.
[275] W. Cegla, J.T. Lewis, G.A. Raggio, The free energy of quantum spin systems and large deviations, Comm. Math. Phys. 118 (2) (1988) 337–354.
[276] M. van den Berg, J.T. Lewis, J.V. Pule, The large deviation principle and some models of an interacting boson gas, Comm. Math. Phys. 118 (1) (1988) 61–85.
[277] T.C. Dorlas, P. Martin, J. Pule, Long cycles in a perturbed mean field model of a boson gas, J. Statist. Phys. 121 (3) (2005) 433–461.
[278] M. Lenci, Classical billiards and quantum large deviations, Ph.D. Thesis, Rutgers University, New Brunswick, N.J., 1999.
[279] J.L. Lebowitz, M. Lenci, H. Spohn, Large deviations for ideal quantum systems, J. Math. Phys. 41 (3) (2000) 1224–1243.
[280] G. Gallavotti, J.L. Lebowitz, V. Mastropietro, Large deviations in rarefied quantum gases, J. Statist. Phys. 108 (5) (2002) 831–861.
[281] F. Hiai, M. Mosonyi, T. Ogawa, Large deviations and Chernoff bound for certain correlated states on a spin chain, J. Math. Phys. 48 (12) (2007) 123301.
[282] M. Lenci, L. Rey-Bellet, Large deviations in quantum lattice systems: One-phase region, J. Statist. Phys. 119 (3) (2005) 715–746.
[283] K. Netočný, F. Redig, Large deviations for quantum spin systems, J. Statist. Phys. 117 (3) (2004) 521–547.
[284] D. Petz, G.A. Raggio, A. Verbeure, Asymptotics of Varadhan-type and the Gibbs variational principle, Comm. Math. Phys. 121 (2) (1989) 271–282.
[285] I. Bjelaković, J.-D. Deuschel, T. Krüger, R. Seiler, R. Siegmund-Schultze, A. Szkola, A quantum version of Sanov's Theorem, Comm. Math. Phys. 260 (3) (2005) 659–671.
[286] M. Keyl, Quantum state estimation and large deviations, Rev. Math. Phys. 18 (1) (2006) 19–60.
[287] K.M.R. Audenaert, J. Calsamiglia, R. Muñoz-Tapia, E. Bagan, L. Masanes, A. Acin, F. Verstraete, Discriminating states: The quantum chernoff bound, Phys. Rev. Lett. 98 (16) (2007) 160501.
[288] R. Ahlswede, V.M. Blinovsky, Large deviations in quantum information theory, Probab. Inform. Trans. 39 (4) (2003) 373–379.
[289] P. Blanchard, P. Combe, M. Sirugue, M. Sirugue-Collin, Estimates of quantum deviations from classical mechanics using large deviation results, in: Quantum Probability and Applications II, in: Lecture Notes in Mathematics, vol. 1136, Springer, 1985, pp. 104–111.
[290] R.S. Ellis, The theory of large deviations and applications to statistical mechanics, in: Lectures for the International Seminar on Extreme Events in Complex Dynamics, 2006.
[291] R.H. Fowler, Statistical Mechanics, 2nd ed., Cambridge University Press, Cambridge, 1966.
[292] W. Feller, An Introduction to Probability Theory and its Applications, vol. II, Wiley, New York, 1970.