Part III

Mathematical aspects

6

The theory of large deviations and applications to statistical mechanics

Richard S. Ellis

Department of Mathematics and Statistics,
University of Massachusetts Amherst, USA

6.1 Introduction

The theory of large deviations studies the exponential decay of probabilities in certain random systems. It has been applied to a wide range of problems in which detailed information on rare events is required. One is often interested not only in the probability of rare events but also in the characteristic behavior of the system as the rare event occurs. For example, in applications to queueing theory and communication systems, the rare event could represent an overload or breakdown of the system. In this case, large deviation methodology can lead to an efficient redesign of the system so that the overload or breakdown does not occur. In applications to statistical mechanics, the theory of large deviations gives precise, exponential-order estimates that are perfectly suited for asymptotic analysis.

This paper will present a number of topics in the theory of large deviations and several applications to statistical mechanics, all united by the concept of relative entropy. This concept entered human culture through the first large deviation calculation in science, carried out by Ludwig Boltzmann. Stated in a modern terminology, his discovery was that the relative entropy expresses the asymptotic behavior of certain multinomial probabilities. This statistical interpretation of entropy has the following crucial physical implication (Ellis 2006, Section 1.1). Entropy is a bridge between a microscopic level, on which physical systems are defined in terms of the complicated interactions among the individual constituent particles, and a macroscopic level, on which the laws describing the behavior of the system are formulated.

Boltzmann and, later, Gibbs asked a fundamental question. How can one use probability theory to study equilibrium properties of physical systems such as an ideal gas, a ferromagnet, or a fluid? These properties include such phenomena as phase transitions, for example, the liquid–gas transition and spontaneous magnetization in a ferromagnet. Another example arises in the study of freely evolving, inviscid fluids, for which one wants to describe coherent states. These are steady, stable mean flows composed of one or more vortices that persist amidst the turbulent fluctuations of the vorticity field. The answer to this fundamental question, which led to the development of classical equilibrium statistical mechanics, is that one studies equilibrium properties via probability measures on configuration space known today as the microcanonical ensemble and the canonical ensemble. For background in statistical mechanics, I recommend Ellis (2006), Lanford (1973), and Wightman (1979), which cover a number of topics relevant to these notes.

Boltzmann's calculation of the asymptotic behavior of multinomial probabilities in terms of relative entropy was carried out in 1877 as a key component of his paper that gave a probabilistic interpretation of the Second Law of Thermodynamics (Boltzmann 1877). This momentous calculation represents a revolutionary moment in human culture, during which both statistical mechanics and the theory of large deviations were born. Boltzmann based his work on the hypothesis that atoms exist. Although this hypothesis is universally accepted today, one might be surprised to learn that it was highly controversial during Boltzmann's time (Lindley 2001, pp. vii–x).

Boltzmann's work is put in historical context by W. R. Everdell in his book *The First Moderns*, which traces the development of the modern consciousness in nineteenth- and twentieth-century thought (Everdell 1997). Chapter 3 focuses on the mathematicians of Germany in the 1870's—namely Cantor, Dedekind, and Frege—who "would become the first creative thinkers in any field to look at the world in a fully twentieth-century manner" (p. 31). Boltzmann is then presented as the man whose investigations in stochastics and statistics made possible the work of the two other great founders of twentieth-century theoretical physics, Planck and Einstein. As Everdell writes, "he was at the center of the change" (p. 48).

Like many areas of mathematics, the theory of large deviations has both a left hand and a right hand; the left hand provides heuristic insight while the right hand provides rigorous proofs. Although the theory is applicable in many diverse settings, the right-hand technicalities can be formidable. Recognizing this, I would like to supplement the rigorous, right-hand formulation of the theory with a number of basic results presented in a left-hand format useful to the applied researcher. For a review of the theory emphasizing applications to statistical mechanics, Touchette (2009) is recommended. The Web document Ellis (2008) is an expanded version of this paper containing additional technical results and applications.

Here is an overview of this paper. In Section 6.2, a basic probabilistic model for random variables having a finite state space is introduced. In Section 6.3, we explain Boltzmann's discovery of the asymptotic behavior of multinomial probabilities in terms of relative entropy. Section 6.4 proves a conditioned limit theorem involving relative entropy that elucidates a basic issue arising in many areas of application. What is the most likely way for an unlikely event to happen? In Section 6.5, we introduce the general concepts of a large deviation principle and a Laplace principle together with related results. In the last two sections, the theory of large deviations is applied to several problems in statistical mechanics. In Section 6.6, the theory is used to study equilibrium properties of a basic model of ferromagnetism known as the Curie–Weiss model, which is a mean-field approximation to the much more complicated Ising model. We then use the insights gained in treating the Curie–Weiss model to derive the phase-transition structure of two other basic models, the Curie–Weiss–Potts model and the mean-field Blume–Capel model. Our work in the preceding section leads in Section 6.7 to the formulation of a general procedure for applying the theory of large deviations to the analysis of an extensive class of statistical-mechanical models, an analysis that allows us to address the fundamental problem of equivalence and nonequivalence of ensembles.

6.2 A basic probabilistic model

In this section, we introduce a basic probabilistic model for random variables having a finite state space. In later sections, a number of questions in the theory of large deviations will be investigated in the context of this model. Let $\alpha \geq 2$ be an integer, $y_1 < y_2 < \ldots < y_\alpha$ a set of α real numbers, and $\rho_1, \rho_2, \ldots, \rho_\alpha$ a set of α positive real numbers summing to 1. We think of $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$ as the set of possible outcomes of a random experiment in which each individual outcome y_k has the probability ρ_k of occurring. The vector $\rho = (\rho_1, \rho_2, \ldots, \rho_\alpha)$ is an element of the set of probability vectors

$$\mathcal{P}_\alpha = \left\{ \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_\alpha) \in \mathbb{R}^\alpha : \gamma_k \geq 0, \sum_{k=1}^{\alpha} \gamma_k = 1 \right\}.$$

Any vector $\gamma \in \mathcal{P}_\alpha$ also defines a probability measure on the set of subsets of Λ via the formula $\gamma = \sum_{k=1}^{\alpha} \gamma_k \, \delta_{y_k}$, where for $y \in \Lambda$, $\delta_{y_k}\{y\} = 1$ if $y = y_k$ and equals 0 otherwise. For $B \subset \Lambda$, we define $\gamma\{B\} = \sum_{k=1}^{\alpha} \gamma_k \, \delta_{y_k}\{B\} = \sum_{y_k \in B} \gamma_k$.

For each positive integer n, the configuration space for n independent repetitions of the experiment is $\Omega_n = \Lambda^n$, a typical element of which is denoted by $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$. For each $\omega \in \Omega_n$, we define

$$P_n\{\omega\} = \prod_{j=1}^{n} \rho\{\omega_j\} = \prod_{j=1}^{n} \rho_{k_j} \quad \text{if } \omega_j = y_{k_j} \, .$$

We then extend this to a probability measure on the set of subsets of Ω_n by defining

$$P_n\{B\} = \sum_{\omega \in B} P_n\{\omega\} \text{ for } B \subset \Omega_n \, .$$

P_n is called the product measure with one-dimensional marginals ρ and is written ρ^n.

An important special case occurs when each ρ_k equals $1/\alpha$. Then for each $\omega \in \Omega_n$, $P_n\{\omega\} = 1/\alpha^n$, and for any subset B of Ω_n, $P_n\{B\} = \text{card}(B)/\alpha^n$, where card$(B)$ denotes the cardinality of B, i.e. the number of elements in B.

We return to the general case. With respect to P_n, the coordinate functions $X_j(\omega) = \omega_j, j = 1, 2, \ldots, n$, are independent identically distributed (i.i.d.) random variables with common distribution ρ; that is, for any subsets $B_1, B_2, \ldots, B_n$ of Λ,

$$\begin{aligned} &P_n\{\omega \in \Omega_n : X_j(\omega) \in B_j \text{ for } j = 1, 2, \ldots, n\} \\ &= \prod_{j=1}^{n} P_n\{\omega \in \Omega_n : X_j(\omega) \in B_j\} = \prod_{j=1}^{n} \rho\{B_j\} \, . \end{aligned}$$

Example 6.1. Random phenomena that can be studied via this basic model include standard examples such as coin tossing and die tossing and also include a discrete ideal gas.

(a) *Coin tossing.* In this case, $\Lambda = \{1, 2\}$ and $\rho_1 = \rho_2 = 1/2$.
(b) *Die tossing.* In this case, $\Lambda = \{1, 2, \ldots, 6\}$ and each $\rho_k = 1/6$.
(c) *Discrete ideal gas.* Consider a discrete ideal gas consisting of n identical, non-interacting particles, each having α equally likely energy levels $y_1, y_2, \ldots, y_\alpha$; in this case each ρ_k equals $1/\alpha$. The coordinate functions X_j represent the random energy levels of the molecules of the gas. The statistical independence of these random variables reflects the fact that the molecules of the gas do not interact. ■

It is worthwhile to reiterate the basic probabilistic model in the context of the general framework in probability theory, which involves five quantities $(\Omega, \mathcal{F}, P, X_j, \Lambda)$. Ω is a configuration space, $\mathcal{F}$ is a σ-algebra of subsets of Ω (i.e. a class of subsets of Ω containing Ω and closed under complements and countable unions), P is a probability measure on $\mathcal{F}$ (i.e. a countably additive set function satisfying $P\{\Omega\} = 1$), and X_j is a sequence of random variables (i.e. measurable functions) taking values in a state space Λ. The triplet $(\Omega, \mathcal{F}, P)$ is called a probability space. In the present section, we make the following choices for $n \in I\!N$: (a) $\Omega = \Omega_n = \Lambda^n$, where $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$; (b) $\mathcal{F}$ is the set of all subsets of Λ^n; and (c) P is the product measure $P_n = \rho^n$, where $\rho = \sum_{k=1}^{\alpha} \rho_k \delta_{y_k}$, each $\rho_k > 0$, and $\sum_{k=1}^{\alpha} \rho_k = 1$; $X_j(\omega) = \omega_j$ for $\omega \in \Omega_n$.

In the next section, we examine Boltzmann's discovery of a statistical interpretation of entropy.

6.3 Boltzmann's discovery and relative entropy

In its original form, Boltzmann's discovery concerns the asymptotic behavior of certain multinomial coefficients. For the purpose of applications in this paper, it is advantageous to formulate it in terms of a probabilistic quantity known as the empirical vector. We use the notation of the preceding section. Thus, let $\alpha \geq 2$ be an integer; let $y_1 < y_2 < \ldots < y_\alpha$ be a set of α real numbers; let $\rho_1, \rho_2, \ldots, \rho_\alpha$ be a set of α positive real numbers summing to 1; let Λ be the set $\{y_1, y_2, \ldots, y_\alpha\}$; and let P_n be the product measure on $\Omega_n = \Lambda^n$ with one-dimensional marginals $\rho = \sum_{k=1}^{\alpha} \rho_k \delta_{y_k}$. For $\omega = (\omega_1, \omega_2, \ldots, \omega_n) \in \Omega_n$, we let $\{X_j, j = 1, \ldots, n\}$ be the coordinate functions defined by $X_j(\omega) = \omega_j$. The X_j form a sequence of i.i.d. random variables with common distribution ρ.

We now turn to the object under study in the present section. For $\omega \in \Omega_n$ and $y \in \Lambda$, define

$$L_n(y) = L_n(\omega, y) = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j(\omega)}\{y\}.$$

Thus $L_n(\omega, y)$ counts the relative frequency with which y appears in the configuration ω; in symbols, $L_n(\omega, y) = n^{-1} \cdot \text{card}\{j \in \{1, \ldots, n\} : \omega_j = y\}$. We then define the empirical vector

$$\begin{aligned} L_n = L_n(\omega) &= (L_n(\omega, y_1), \ldots, L_n(\omega, y_\alpha)) \\ &= \frac{1}{n} \sum_{j=1}^{n} \left(\delta_{X_j(\omega)}\{y_1\}, \ldots, \delta_{X_j(\omega)}\{y_\alpha\}\right). \end{aligned}$$

L_n equals the sample mean of the i.i.d. random vectors $(\delta_{X_j(\omega)}\{y_1\}, \ldots, \delta_{X_j(\omega)}\{y_\alpha\})$. It takes values in the set of probability vectors

$$\mathcal{P}_\alpha = \left\{ \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_\alpha) \in \mathbb{R}^\alpha : \gamma_k \geq 0, \sum_{k=1}^{\alpha} \gamma_k = 1 \right\}.$$

The limiting behavior of L_n is straightforward to determine. Let $\| \cdot \|$ denote the Euclidean norm on $\mathbb{R}^\alpha$. For any $\gamma \in \mathcal{P}_\alpha$ and $\varepsilon > 0$, we define the open ball

$$B(\gamma, \varepsilon) = \{\nu \in \mathcal{P}_\alpha : \|\gamma - \nu\| < \varepsilon\}.$$

Since the X_j have the common distribution ρ, for each $y_k \in \Lambda$

$$E^{P_n}\{L_n(y_k)\} = E^{P_n}\left\{\frac{1}{n}\sum_{j=1}^{n}\delta_{X_j}\{y_k\}\right\} = \frac{1}{n}\sum_{j=1}^{n}P_n\{X_j = y_k\} = \rho_k\,,$$

where E^{P_n} denotes the expectation with respect to P_n. Hence, by the weak law of large numbers for the sample means of i.i.d. random variables, for any $\varepsilon > 0$

$$\lim_{n\to\infty} P_n\{L_n \in B(\rho, \varepsilon)\} = 1\,. \tag{6.1}$$

It follows that for any $\gamma \in \mathcal{P}_\alpha$ not equal to ρ and for any $\varepsilon > 0$ satisfying $0 < \varepsilon < \|\rho - \gamma\|$,

$$\lim_{n\to\infty} P_n\{L_n \in B(\gamma, \varepsilon)\} = 0\,. \tag{6.2}$$

As we will see, Boltzmann's discovery implies that these probabilities converge to 0 exponentially fast in n. The exponential decay rate is given in terms of the relative entropy, which we now define.

Definition 6.1. (Relative entropy.) *Let $\rho = (\rho_1, \ldots, \rho_\alpha)$ denote the probability vector in $\mathcal{P}_\alpha$ in terms of which the basic probabilistic model is defined. The relative entropy of $\gamma \in \mathcal{P}_\alpha$ with respect to ρ is defined by*

$$I_\rho(\gamma) = \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k}\,.$$

Several properties of the relative entropy are given in the next lemma. The proof is typical of proofs of analogous results involving relative entropy (e.g. Proposition 6.4.3) in that we use a global, convexity-based inequality rather than calculus to determine where I_ρ attains its infimum over $\mathcal{P}_\alpha$. In the present case the global inequality is that for $x \geq 0$, $x \log x \geq x - 1$ with equality if and only if $x = 1$.

Lemma 6.3.1. *For $\gamma \in \mathcal{P}_\alpha$, $I_\rho(\gamma)$ measures the discrepancy between γ and ρ in the sense that $I_\rho(\gamma) \geq 0$ and $I_\rho(\gamma) = 0$ if and only if $\gamma = \rho$. Thus $I_\rho(\gamma)$ attains its infimum of 0 over $\mathcal{P}_\alpha$ at the unique measure $\gamma = \rho$. In addition, I_ρ is strictly convex on $\mathcal{P}_\alpha$; that is, for $0 < \lambda < 1$ and any $\mu \neq \nu$ in $\mathcal{P}_\alpha$,*

$$I_\rho(\lambda\mu + (1-\lambda)\nu) < \lambda I_\rho(\mu) + (1-\lambda)I(\nu)\,.$$

Proof. For $x \geq 0$, the graph of the strictly convex function $x \log x$ has the tangent line $y = x - 1$ at $x = 1$. Hence $x \log x \geq x - 1$, with equality if and only if $x = 1$. It follows that for any $\gamma \in \mathcal{P}_\alpha$,

$$\frac{\gamma_k}{\rho_k} \log \frac{\gamma_k}{\rho_k} \geq \frac{\gamma_k}{\rho_k} - 1\,, \tag{6.3}$$

with equality if and only if $\gamma_k = \rho_k$. Multiplying this inequality by ρ_k and summing over k yields

$$I_\rho(\gamma) = \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k} \geq \sum_{k=1}^{\alpha} (\gamma_k - \rho_k) = 0 \,.$$

We now prove that $I_\rho(\gamma) = 0$ if and only if $\gamma = \rho$. If $\gamma = \rho$, then the definition of the relative entropy shows that $I_\rho(\gamma) = 0$. Now assume that $I_\rho(\gamma) = 0$. Then

$$\begin{aligned} 0 &= \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k} \\ &= \sum_{k=1}^{\alpha} \left(\gamma_k \log \frac{\gamma_k}{\rho_k} - (\gamma_k - \rho_k) \right) \\ &= \sum_{k=1}^{\alpha} \rho_k \left(\frac{\gamma_k}{\rho_k} \log \frac{\gamma_k}{\rho_k} - \left(\frac{\gamma_k}{\rho_k} - 1 \right) \right) . \end{aligned}$$

We now use the facts that $\rho_k > 0$ and that for $x \geq 0$, $x \log x \geq x - 1$ with equality if and only if $x = 1$. It follows that for each k, $\gamma_k = \rho_k$ and thus that $\gamma = \rho$. This completes the proof that $I_\rho(\gamma) \geq 0$ and $I_\rho(\gamma) = 0$ if and only if $\gamma = \rho$, which is the first assertion in the proposition.

Since

$$I_\rho(\gamma) = \sum_{k=1}^{\alpha} \rho_k \frac{\gamma_k}{\rho_k} \log \frac{\gamma_k}{\rho_k} \,,$$

the strict convexity of I_ρ is a consequence of the strict convexity of $x \log x$ for $x \geq 0$. ■

We are now ready to give the first formulation of Boltzmann's discovery, which we state using a heuristic notation and which we label, in recognition of its formal status, as a pseudo-theorem. However, the formal calculations used to motivate the pseudo-theorem can easily be turned into a rigorous proof of an asymptotic theorem. That theorem is stated in Theorem 6.3.3. From Boltzmann's momentous discovery, both the theory of large deviations and the Gibbsian formulation of equilibrium statistical mechanics grew. The notation $P_n\{L_n \in d\gamma\}$ represents the probability that L_n is close to γ.

Pseudo-theorem 6.3.2. (Boltzmann's discovery—formulation 1.) *For any* $\gamma \in \mathcal{P}_\alpha$,

$$P_n\{L_n \in d\gamma\} \approx \exp[-nI_\rho(\gamma)] \quad \text{as } n \to \infty \,.$$

Heuristic proof. Since $\gamma \in \mathcal{P}_\alpha$, $\sum_{k=1}^{\alpha} \gamma_k = 1$. By elementary combinatorics,

$$\begin{aligned} P_n\{L_n \in d\gamma\} &= P_n\left\{\omega \in \Omega_n : L_n(\omega) \sim \frac{1}{n}(n\gamma_1, n\gamma_2, \ldots, n\gamma_\alpha)\right\} \\ &\approx P_n\{\text{card}\{\omega_j = y_1\} \sim n\gamma_1, \ldots, \text{card}\{\omega_j = y_\alpha\} \sim n\gamma_\alpha\} \\ &\approx \frac{n!}{(n\gamma_1)!(n\gamma_2)! \cdots (n\gamma_\alpha)!} \, \rho_1^{n\gamma_1} \, \rho_2^{n\gamma_2} \, \cdots \, \rho_\alpha^{n\gamma_\alpha} \,. \end{aligned}$$

Stirling's formula in the weak form $\log(n!) = n\log n - n + \mathrm{O}(\log n)$ yields

$$
\begin{aligned}
&\frac{1}{n}\log P_n\{L_n \in d\gamma\} \\
&\approx \frac{1}{n}\log\left(\frac{n!}{(n\gamma_1)!(n\gamma_2)!\cdots(n\gamma_\alpha)!}\right) + \sum_{k=1}^{\alpha}\gamma_k \log\rho_k \\
&= \frac{1}{n}\log\left(\frac{n^n e^{-n}}{(n\gamma_1)^{n\gamma_1}e^{-n\gamma_1}\cdots(n\gamma_\alpha)^{n\gamma_\alpha}e^{-n\gamma_\alpha}}\right) + \sum_{k=1}^{\alpha}\gamma_k\log\rho_k + \mathrm{O}\left(\frac{\log n}{n}\right) \\
&= \frac{1}{n}\log\left(\frac{1}{\gamma_1^{n\gamma_1}\cdots\gamma_\alpha^{n\gamma_\alpha}}\right) + \sum_{k=1}^{\alpha}\gamma_k\log\rho_k + \mathrm{O}\left(\frac{\log n}{n}\right) \\
&= -\sum_{k=1}^{\alpha}\gamma_k\log\gamma_k + \sum_{k=1}^{\alpha}\gamma_k\log\rho_k + \mathrm{O}\left(\frac{\log n}{n}\right) \\
&= -\sum_{k=1}^{\alpha}\gamma_k\log\frac{\gamma_k}{\rho_k} + \mathrm{O}\left(\frac{\log n}{n}\right) \;=\; -I_\rho(\gamma) + \mathrm{O}\left(\frac{\log n}{n}\right).
\end{aligned}
$$

The term $\mathrm{O}(\log n/n)$ converges to 0 as $n \to \infty$. Hence, multiplying both sides of the last equation by n and exponentiating yields the result. ■

Pseudo-theorem 6.3.2 has the following interesting consequence. Let γ be any vector in $\mathcal{P}_\alpha$ which differs from ρ. Since $I_\rho(\gamma) > 0$ (Lemma 6.3.1), it follows that

$$P_n\{L_n \in d\gamma\} \approx \exp[-nI_\rho(\gamma)] \to 0 \ \text{ as } n \to \infty,$$

a limit which, if rigorous, would imply eqn (6.2).

Let A be a Borel subset of $\mathcal{P}_\alpha$, i.e. A is a member of the Borel σ-algebra of $\mathcal{P}_\alpha$, which is the smallest σ-algebra containing the open sets. The class of Borel subsets includes all open subsets of $\mathcal{P}_\alpha$ and all closed subsets of $\mathcal{P}_\alpha$. If ρ is not contained in the closure of A, then by the weak law of large numbers

$$\lim_{n\to\infty} P_n\{L_n \in A\} = 0,$$

and, by analogy with the heuristic asymptotic result given in Pseudo-theorem 6.3.2, we expect that these probabilities converge to 0 exponentially fast with n. This is in fact the case. In order to express the exponential decay rate of such probabilities in terms of the relative entropy, we introduce the notation $I_\rho(A) = \inf_{\gamma\in A} I_\rho(\gamma)$. The range of $L_n(\omega)$ for $\omega \in \Omega_n$ is the set of probability vectors having the form k/n, where $k \in \mathbb{R}^\alpha$ has nonnegative integer coordinates summing to n; hence the cardinality of the range does not exceed n^α. Since

$$P_n\{L_n \in A\} = \sum_{\gamma\in A} P_n\{L_n \in d\gamma\} \approx \sum_{\gamma\in A}\exp[-nI_\rho(\gamma)]$$

and

$$\exp[-nI_\rho(A)] \leq \sum_{\gamma \in A} \exp[-nI_\rho(\gamma)] \leq n^\alpha \exp[-nI_\rho(A)] \, ,$$

one expects that to exponential order,

$$P_n\{L_n \in A\} \approx \exp[-nI_\rho(A)] \ \text{ as } n \to \infty \, . \tag{6.4}$$

As formulated in Corollary 6.3.4, this asymptotic result is indeed valid. It is a consequence of the following rigorous reformulation of Boltzmann's discovery, known as Sanov's Theorem, which expresses the large deviation principle for the empirical vectors L_n. That concept is defined in general in Definition 6.2, and a general form of Sanov's Theorem is stated in Theorem 6.5.6. The special case of Sanov's Theorem stated next is proved in (Ellis 2006, Theorem VIII.2).

Theorem 6.3.3. (Boltzmann's discovery—formulation 2.) *The sequence of empirical vectors L_n satisfies the large deviation principle on $\mathcal{P}_\alpha$ with rate function I_ρ in the following sense.*

(a) Large deviation upper bound. *For any closed subset F of $\mathcal{P}_\alpha$,*

$$\limsup_{n \to \infty} \frac{1}{n} \log P_n\{L_n \in F\} \leq -I_\rho(F) \, .$$

(b) Large deviation lower bound. *For any open subset G of $\mathcal{P}_\alpha$,*

$$\liminf_{n \to \infty} \frac{1}{n} \log P_n\{L_n \in G\} \geq -I_\rho(G) \, .$$

Comments on the proof. For $\gamma \in \mathcal{P}_\alpha$ and $\varepsilon > 0$, $B(\gamma, \varepsilon)$ denotes the open ball with center γ and radius ε, and $\overline{B}(\gamma, \varepsilon)$ denotes the corresponding closed ball. Since $\mathcal{P}_\alpha$ is a compact subset of $\mathbb{R}^\alpha$, any closed subset F of $\mathcal{P}_\alpha$ is automatically compact. By a standard covering argument, it is not hard to show that the large deviation upper bound holds for any closed set F provided that one obtains the large deviation upper bound for any closed ball $\overline{B}(\gamma, \varepsilon)$:

$$\limsup_{n \to \infty} \frac{1}{n} \log P_n\{L_n \in \overline{B}(\gamma, \varepsilon)\} \leq -I_\rho(\overline{B}(\gamma, \varepsilon)) \, .$$

Likewise, the large deviation lower bound holds for any open set G provided one obtains the large deviation lower bound for any open ball $B(\gamma, \varepsilon)$:

$$\liminf_{n \to \infty} \frac{1}{n} \log P_n\{L_n \in B(\gamma, \varepsilon)\} \geq -I_\rho(B(\gamma, \varepsilon)) \, .$$

The bounds in the last two equations can be proved via combinatorics and Stirling's formula as in the heuristic proof of Pseudo-theorem 6.3.2; one can easily adapt the calculations given in (Ellis 2006, Section I.4). The details are omitted in the present work. ■

Given A a Borel subset of $\mathcal{P}_\alpha$, we denote by A° the interior of A relative to $\mathcal{P}_\alpha$ and by $\overline{A}$ the closure of A. For a class of Borel subsets, the next corollary gives a rigorous version of the asymptotic formula (6.4). This class consists of sets A such that $\overline{A^\circ}$ equals $\overline{A}$. Any open ball $B(\gamma, \varepsilon)$ or closed ball $\overline{B}(\gamma, \varepsilon)$ satisfies this condition. For any Borel subset A, the continuity of I_ρ on $\mathcal{P}_\alpha$ guarantees that $I_\rho(A^\circ) = I_\rho(\overline{A^\circ})$. Hence if $\overline{A^\circ} = \overline{A}$, then $I_\rho(A^\circ) = I_\rho(\overline{A})$, which is the hypothesis of Theorem 6.5.2. The next corollary is a consequence of that theorem and the large deviation principle given in Theorem 6.3.3.

Corollary 6.3.4. *Let A be any Borel subset of $\mathcal{P}_\alpha$ satisfying $\overline{A^\circ} = \overline{A}$. Then*

$$\lim_{n \to \infty} \frac{1}{n} \log P_n\{L_n \in A\} = -I_\rho(A) .$$

The next corollary of Theorem 6.3.3 allows one to conclude that a large class of probabilities involving L_n converge to 0. The general version of this corollary given in Theorem 6.5.3 is extremely useful in applications. For example, we will use it in Section 6.6 to analyze several lattice spin models and in Section 6.7 to motivate the definitions of the sets of equilibrium macrostates for the canonical ensemble and the microcanonical ensemble for a general class of systems (Theorems 6.7.1(c) and 6.7.2(c)).

Corollary 6.3.5. *Let A be any Borel subset of $\mathcal{P}_\alpha$ such that $\overline{A}$ does not contain ρ. Then $I_\rho(\overline{A}) > 0$, and for some $C < \infty$*

$$P_n\{L_n \in A\} \leq C \exp[-nI_\rho(\overline{A})/2] \to 0 \quad \textit{as } n \to \infty .$$

Proof. Since $I_\rho(\gamma) > I_\rho(\rho) = 0$ for any $\gamma \neq \rho$, the positivity of $I_\rho(\overline{A})$ follows from the continuity of I_ρ on $\mathcal{P}_\alpha$. The second assertion is an immediate consequence of the large deviation upper bound applied to $\overline{A}$ and the positivity of $I_\rho(\overline{A})$. ■

Take any $\varepsilon > 0$. Applying Corollary 6.3.5 to the complement of the open ball $B(\rho, \varepsilon)$ yields $P_n\{L_n \notin B(\rho, \varepsilon)\} \to 0$ or, equivalently,

$$\lim_{n \to \infty} P_n\{L_n \in B(\rho, \varepsilon)\} = 1 .$$

Although this rederives the weak law of large numbers for L_n expressed in eqn (6.1), this second derivation relates the order-1 limit for L_n to the point $\rho \in \mathcal{P}_\alpha$ at which the rate function I_ρ attains its infimum. In this context, we call ρ the equilibrium value of L_n with respect to the measures P_n.

In the next section, we will present a limit theorem for L_n whose proof is based on the precise, exponential-order estimates given by the large deviation principle in Theorem 6.3.3.

6.4 The most likely way for an unlikely event to happen

In this section, we prove a conditioned limit theorem that elucidates a basic issue arising in many areas of application. What is the most likely way for an unlikely event to happen? For example, in applications to queueing theory and communication systems, the unlikely event could represent an overload or breakdown of the system. If one knows the most likely way that the overload could occur, then one could efficiently redesign the system so that the overload does not happen.

We use the notation of the preceding section. Thus, let $\alpha \geq 2$ be an integer; let $y_1 < y_2 < \ldots < y_\alpha$ be a set of α real numbers; let $\rho_1, \rho_2, \ldots, \rho_\alpha$ be a set of α positive real numbers summing to 1; let Λ be the set $\{y_1, y_2, \ldots, y_\alpha\}$; and let P_n be the product measure on $\Omega_n = \Lambda^n$ with one-dimensional marginals $\rho = \sum_{k=1}^{\alpha} \rho_k \delta_{y_k}$. For $\omega = (\omega_1, \omega_2, \ldots, \omega_n) \in \Omega_n$, we let $\{X_j, j = 1, \ldots, n\}$ be the coordinate functions defined by $X_j(\omega) = \omega_j$. The X_j form a sequence of i.i.d. random variables with common distribution ρ. For $\omega \in \Omega_n$ and $y \in \Lambda$, we also define

$$L_n(y) = L_n(\omega, y) = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j(\omega)}\{y\}$$

and the empirical vector

$$L_n = L_n(\omega) = (L_n(\omega, y_1), \ldots, L_n(\omega, y_\alpha)) = \frac{1}{n} \sum_{j=1}^{n} \left(\delta_{X_j(\omega)}\{y_1\}, \ldots, \delta_{X_j(\omega)}\{y_\alpha\} \right).$$

L_n takes values in the set of probability vectors

$$\mathcal{P}_\alpha = \left\{ \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_\alpha) \in \mathbb{R}^\alpha : \gamma_k \geq 0, \sum_{k=1}^{\alpha} \gamma_k = 1 \right\}.$$

The main result in this section is the conditioned limit theorem for L_n given in Theorem 6.4.1. This theorem has the bonus of giving insight into a basic construction in statistical mechanics. As we show in Section 5 of Ellis (2008), it motivates the form of the canonical ensemble for the discrete ideal gas and, by extension, for any statistical-mechanical system characterized by conservation of energy. These unexpected theorems are the first indication of the power of Boltzmann's discovery, which gives precise exponential-order estimates for probabilities of the form $P_n\{L_n \in A\}$.

The conditioned limit theorem that we will consider has the following form. Suppose that one is given a particular set A for which $P_n\{L_n \in A\} > 0$ for all sufficiently large n. One wants to determine a set B belonging to a certain class (e.g. open balls) such that the conditioned limit

$$\lim_{n \to \infty} P_n\{L_n \in B \,|\, L_n \in A\} = \lim_{n \to \infty} P_n\{L_n \in B \cap A\} \cdot \frac{1}{P_n\{L_n \in A\}} = 1$$

is valid. Since, to exponential order,

$$P_n\{L_n \in B \cap A\} \cdot \frac{1}{P_n\{L_n \in A\}} \approx \exp[-n(I_\rho(B \cap A) - I_\rho(A))],$$

one should obtain the conditioned limit if B satisfies $I_\rho(B \cap A) = I_\rho(A)$. If one can determine the point in A where the infimum of I_ρ is attained, then one picks B to contain this point. As we explain in Proposition 6.4.3, such a minimizing point can be determined for an important class of sets A.

In order to formulate the conditioned limit theorem, we define

$$S_n = \sum_{j=1}^{n} X_j \quad \text{and} \quad \bar{y} = \sum_{k=1}^{\alpha} y_k \rho_k = E^{P_n}\{X_1\}\,.$$

By the weak law of large numbers, for any $\varepsilon > 0$

$$\lim_{n\to\infty} P_n\{|S_n/n - \bar{y}| \geq \varepsilon\} = 0\,.$$

Given a small positive number a, we choose z to satisfy $y_1 < z - a < z < \bar{y}$. The conditioned limit theorem involves positive numbers $\{\rho_k^*, k = 1, \ldots, \alpha\}$ summing to 1 and satisfying

$$\rho_k^* = \lim_{n\to\infty} P_n\{X_1 = k \,|\, S_n/n \in [z-a, z]\}\,. \tag{6.5}$$

By the law of large numbers, the event on which we are conditioning—namely, that $S_n/n \in [z-a, z]$ for all n—is a rare event converging to 0 as $n \to \infty$. A similar result would hold if we assumed that $S_n/n \in [z, z+a]$, where $\bar{y} < z < z + a < y_\alpha$.

The limit (6.5) will be seen to follow from the following more easily answered question: conditioned on the event $\{S_n/n \in [z-a, z]\}$, determine the most likely configuration $\rho^* = (\rho_1^*, \ldots, \rho_\alpha^*)$ of L_n in the limit $n \to \infty$. In other words, we want $\rho^* \in \mathcal{P}_\alpha$ such that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P_n\{L_n \in B(\rho^*, \varepsilon) \,|\, S_n/n \in [z-a, z]\} = 1\,.$$

In Theorem 6.4.1 we give the form of ρ^*, which depends on z through a parameter β. In order to indicate this dependence, we write $\rho^{(\beta)}$ in place of ρ^*. The form of $\rho^{(\beta)}$ is independent of a. In the proof of the analogous result given in Ellis (2008), we write $-\beta$ instead of β in the definition of $\rho^{(\beta)}$ in order to be consistent with conventions in statistical mechanics.

Theorem 6.4.1. *Let $z \in (y_1, \bar{y})$ be given, and choose $a > 0$ such that $z - a > y_1$. The following conclusions hold.*

(a) *There exists a unique $\rho^{(\beta)} \in \mathcal{P}_\alpha$ such that for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, S_n/n \in [z-a, z]\} = 1\,. \tag{6.6}$$

The quantity $\rho^{(\beta)} = (\rho_1^{(\beta)}, \ldots, \rho_\alpha^{(\beta)})$ has the form

$$\rho_k^{(\beta)} = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\,\rho_j} \cdot \exp[\beta y_k]\,\rho_k\,,$$

where $\beta = \beta(z) < 0$ is the unique value of β satisfying $\sum_{k=1}^{\alpha} y_k \rho_k^{(\beta)} = z$.

(b) *For any continuous function f mapping $\mathcal{P}_\alpha$ into $\mathbb{R}$,*

$$\lim_{n\to\infty} E^{P_n}\{f(L_n) \,|\, S_n/n \in [z-a, z]\} = f(\rho^{(\beta)})\,.$$

(c) *For each* $k \in \{1, \ldots, \alpha\}$,

$$\lim_{n\to\infty} P_n\{X_1 = y_k \,|\, S_n/n \in [z-a, z]\} = \rho_k^{(\beta)} \,.$$

In order to prove the theorem, for $t \in \mathbb{R}$ we introduce

$$c(t) = \log E^{P_n}\{e^{tX_1}\} = \log\left(\sum_{k=1}^{\alpha} \exp[ty_k]\,\rho_k\right). \tag{6.7}$$

The function c, which equals the logarithm of the moment-generating function of X_1, is also known as the cumulant-generating function of X_1. In order to show that $\rho^{(\beta)}$ is well defined, we need the following lemma.

Lemma 6.4.2. *The cumulant-generating function c has the following properties.*

(a) $c''(t) > 0$ *for all* t, *i.e. c is strictly convex on* $\mathbb{R}$.
(b) $c'(0) = \sum_{k=1}^{\alpha} y_k \rho_k = \bar{y}$.
(c) $c'(t) \to y_1$ *as* $t \to -\infty$, *and* $c'(t) \to y_\alpha$ *as* $t \to \infty$.
(d) *The range of* $c'(t)$ *for* $t \in \mathbb{R}$ *is the open interval* (y_1, y_α), *which is the interior of the smallest interval containing the support* $\{y_1, y_2, \ldots, y_\alpha\}$.

Proof.

(a) We define

$$\langle y\rangle_t = \frac{1}{\sum_{j=1}^{\alpha}\exp[ty_j]\,\rho_j}\cdot\sum_{k=1}^{\alpha} y_k\,\exp[ty_k]\,\rho_k$$

and

$$\langle (y-\langle y\rangle_t)^2\rangle_t = \frac{1}{\sum_{j=1}^{\alpha}\exp[ty_j]\,\rho_j}\cdot\sum_{k=1}^{\alpha}(y_k-\langle y\rangle_t)^2\,\exp[ty_k]\,\rho_k\,,$$

and calculate

$$c'(t) = \frac{1}{\sum_{j=1}^{\alpha}\exp[ty_j]\,\rho_j}\cdot\sum_{k=1}^{\alpha} y_k\,\exp[ty_k]\,\rho_k = \langle y\rangle_t$$

and

$$\begin{aligned} c''(t) &= \frac{1}{\sum_{j=1}^{\alpha}\exp[ty_j]\,\rho_j}\cdot\sum_{k=1}^{\alpha} y_k^2\,\exp[ty_k]\,\rho_k - \langle y\rangle_t^2 \\ &= \langle (y-\langle y\rangle_t)^2\rangle_t \;>\; 0\,. \end{aligned}$$

The last line gives part (a). This calculation shows that $c'(t)$ equals the mean of the probability measure $\exp[ty_k]\rho_k/\sum_{j=1}^{\alpha}\exp[ty_j]\,\rho_j$, and $c''(t)$ equals the variance of this probability measure.

(b) This follows from the formula for $c'(t)$ in part (a).

(c) Since $y_1 < y_j$ for all $j = 2, \dots, \alpha$,

$$\lim_{t \to -\infty} c'(t) = \lim_{t \to -\infty} \frac{1}{\sum_{j=1}^{\alpha} \exp[t(y_j - y_1)]\, \rho_j} \cdot \sum_{k=1}^{\alpha} y_k \exp[t(y_k - y_1]\, \rho_k = y_1 .$$

One similarly proves that $\lim_{t \to \infty} c'(t) = y_\alpha$.

(d) According to part (a), $c'(t)$ is a strictly increasing function of t. Hence the limits in part (c) show that the range of $c'(t)$ for $t \in \mathbb{R}$ equals the open interval (y_1, y_α). This completes the proof of the lemma. ■

We now prove Theorem 6.4.1.

Proof of Theorem 6.4.1. We first prove that $\rho^{(\beta)}$ is well defined. This follows immediately from Lemma 6.4.2, which shows that there exists a unique $\beta = \beta(z)$ satisfying

$$\begin{aligned} c'(\beta) &= \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \sum_{k=1}^{\alpha} y_k \exp[\beta y_k]\, \rho_k \\ &= \sum_{k=1}^{\alpha} y_k \rho_k^{(\beta)} = z , \end{aligned} \tag{6.8}$$

as claimed. Since $y_1 < z < \bar{y}$, $\beta = \beta(z)$ is negative.

Assuming the limit in part (a), we first prove the limits in parts (b) and (c). We then prove the limit in part (a).

(b) This limit follows from part (a) and the continuity of f. In order to see this, given $\delta > 0$, choose $\varepsilon > 0$ so that whenever $\gamma \in \mathcal{P}_\alpha$ lies in the open ball $B(\rho^{(\beta)}, \varepsilon)$, we have $|f(\gamma) - f(\rho^{(\beta)})| < \delta$. Then

$$\begin{aligned} &\left| E^{P_n}\{f(L_n) \mid S_n/n \in [z-a, z]\} - f(\rho^{(\beta)}) \right| \\ &\leq E^{P_n}\{|f(L_n) - f(\rho^{(\beta)})| \mid S_n/n \in [z-a, z]\} \\ &= E^{P_n}\{|f(L_n) - f(\rho^{(\beta)})|\, 1_{B(\rho^{(\beta)}, \varepsilon)}(L_n) \mid S_n/n \in [z-a, z]\} \\ &\quad + E^{P_n}\{|f(L_n) - f(\rho^{(\beta)})|\, 1_{[B(\rho^{(\beta)}, \varepsilon)]^c}(L_n) \mid S_n/n \in [z-a, z]\} \\ &\leq \delta + 2\|f\|_\infty P\{L_n \in [B(\rho^{(\beta)}, \varepsilon)]^c \mid S_n/n \in [z-a, z]\} . \end{aligned}$$

By part (a), the probability in the last line of the above equation converges to 0 as $n \to \infty$. Since $\delta > 0$ is arbitrary, part (b) is proved.

(c) By symmetry,

$$\begin{aligned} &P_n\{X_1 = y_k \mid S_n/n \in [z-a, z]\} \\ &\quad = E^{P_n}\{\delta_{X_1}\{y_k\} \mid S_n/n \in [z-a, z]\} \\ &\quad = E^{P_n}\left\{ \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j}\{y_k\} \;\middle|\; S_n/n \in [z-a, z] \right\} \\ &\quad = E^{P_n}\{L_n(y_k) \mid S_n/n \in [z-a, z]\} . \end{aligned}$$

Now define f_k to be the continuous function on $\mathcal{P}_\alpha$ that maps γ to γ_k. Part (b) yields the desired limit:

$$\begin{aligned}
&\lim_{n\to\infty} P_n\{X_1 = y_k \,|\, S_n/n \in [z-a, z]\} \\
&\quad = \lim_{n\to\infty} E^{P_n}\{L_n(y_k) \,|\, S_n/n \in [z-a, z]\} \\
&\quad = \lim_{n\to\infty} E^{P_n}\{f_k(L_n) \,|\, S_n/n \in [z-a, z]\} \\
&\quad = f_k(\rho^{(\beta)}) = \rho_k^{(\beta)} .
\end{aligned}$$

(a) In order to prove the limit

$$\lim_{n\to\infty} P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, S_n/n \in [z-a, z]\} = 1 ,$$

we rewrite the event $\{S_n/n \in [z-a, z]\}$ in terms of L_n. Define the closed convex set

$$\Gamma(z) = \left\{ \gamma \in \mathcal{P}_\alpha : \sum_{k=1}^{\alpha} y_k \gamma_k \in [z-a, z] \right\} ,$$

which contains $\rho^{(\beta)}$. Since for each $\omega \in \Omega_n$

$$\frac{1}{n} S_n(\omega) = \sum_{k=1}^{\alpha} y_k \, L_n(\omega, y_k) ,$$

it follows that $\{\omega \in \Omega_n : S_n(\omega)/n \in [z-a, z]\} = \{\omega \in \Omega_n : L_n(\omega) \in \Gamma(z)\}$ and thus that

$$P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, S_n/n \in [z-a, z]\} = P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, L_n \in \Gamma(z)\} .$$

We first motivate the desired limit, using the formal notation

$$P_n\{L_n \in A\} \approx \exp[-n I_\rho(A)] \ \text{ as } n \to \infty , \tag{6.9}$$

which was introduced in eqn (6.4). Sanov's Theorem (Theorem 6.5.6) is the rigorous statement of the exponential behavior of the distribution of L_n; a special case that applies to the current setup is given in Theorem 6.3.3. For large n, we have by eqn (6.9)

$$\begin{aligned}
&P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, S_n/n \in [z-a, z]\} \\
&\quad = P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, L_n \in \Gamma(z)\} \\
&\quad = P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \cap \Gamma(z)\} \cdot \frac{1}{P_n\{L_n \in \Gamma(z)\}} \\
&\quad \approx \exp[-n(I_\rho(B(\rho^{(\beta)}, \varepsilon) \cap \Gamma(z)) - I_\rho(\Gamma(z)))] .
\end{aligned}$$

The last expression, and thus the probability in the first line of the equation, are of order 1 provided

$$I_\rho(B(\rho^{(\beta)}, \varepsilon) \cap \Gamma(z)) = I_\rho(\Gamma(z)) . \tag{6.10}$$

Proposition 6.4.3 shows that I_ρ attains its infimum over $\Gamma(z)$ at the unique point $\rho^{(\beta)}$. This gives eqn (6.10) and motivates the fact that for large n,

$$P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, S_n/n \in [z-a, z]\} \approx 1\,.$$

We now convert these formal calculations into a proof of the limit

$$\lim_{n\to\infty} P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, S_n/n \in [z-a, z]\} = 1\,.$$

We prove this by showing that

$$\begin{aligned} &\lim_{n\to\infty} P_n\{L_n \in [B(\rho^{(\beta)}, \varepsilon)]^c \,|\, S_n/n \in [z-a, z]\} \\ &\quad = \lim_{n\to\infty} P_n\{L_n \in [B(\rho^{(\beta)}, \varepsilon)]^c \,|\, L_n \in \Gamma(z)\} = 0\,. \end{aligned} \tag{6.11}$$

The key is to use Corollary 6.3.4, which states that if A is any Borel subset of $\mathcal{P}_\alpha$ satisfying $\overline{A^\circ} = \overline{A}$, then

$$\lim_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in A\} = -I_\rho(A)\,.$$

Since both sets $[B(\rho^{(\beta)}, \varepsilon)]^c \cap \Gamma(z)$ and $\Gamma(z)$ satisfy the hypothesis of Corollary 6.3.4, it follows that

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in [B(\rho^{(\beta)}, \varepsilon)]^c \,|\, L_n \in \Gamma(z)\} \\ &= \lim_{n\to\infty} \frac{1}{n} P_n\{L_n \in [B(\rho^{(\beta)}, \varepsilon)]^c \cap \Gamma(z)\} - \lim_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in \Gamma(z)\} \\ &= -I_\rho([B(\rho^{(\beta)}, \varepsilon)]^c \cap \Gamma(z)) + I_\rho(\Gamma(z))\,. \end{aligned}$$

According to Proposition 6.4.3, I_ρ attains its infimum over $\Gamma(z)$ at the unique point $\rho^{(\beta)}$. It follows that $I_\rho([B(\rho^{(\beta)}, \varepsilon)]^c \cap \Gamma(z)) - I_\rho(\Gamma(z)) > 0$. In combination with the last equation, this yields the desired limit (6.11), showing in fact that the convergence to 0 is exponentially fast.

The proof of Theorem 6.4.1 will be complete after we prove the next proposition, which is carried out without calculus by using properties of the relative entropy. In order to motivate the proposition, we consider the calculus problem of determining critical points of $I_\rho(\gamma)$ subject to the constraints that $\sum_{k=1}^{\alpha} \gamma_k = 1$ and $\sum_{k=1}^{\alpha} y_k \gamma_k = z$. Let λ and $-\beta$ be Lagrange multipliers corresponding to these two constraints. Then, for each k,

$$\begin{aligned} 0 &= \frac{\partial(I_\rho(\gamma) + \lambda\,(\sum_{k=1}^{\alpha} \gamma_k - 1) - \beta\,(\sum_{k=1}^{\alpha} y_k \gamma_k) - z)}{\partial \gamma_k} \\ &= \log \gamma_k + 1 - \log \rho_k + \lambda - \beta y_k\,. \end{aligned}$$

It follows that $\gamma_k = \rho_k^{(\beta)}$, where $\beta = \beta(z)$ is as specified in the proposition.

Proposition 6.4.3. *Let $z \in (y_1, \bar{y})$ be given, and choose $a > 0$ such that $z - a > y_1$. Then I_ρ attains its infimum over*

$$\Gamma(z) = \{\gamma \in \mathcal{P}_\alpha : \sum_{k=1}^{\alpha} y_k \gamma_k \in [z-a, z]\}$$

at the unique point $\rho^{(\beta)} = (\rho_1^{(\beta)}, \ldots, \rho_\alpha^{(\beta)})$ defined in part (a) of Theorem 6.4.1.

Proof. We recall that for each $k \in \{1, \ldots, \alpha\}$,

$$\frac{\rho_k^{(\beta)}}{\rho_k} = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \exp[\beta y_k] = \frac{1}{\exp[c(\beta)]} \cdot \exp[\beta y_k]\,,$$

where c is the cumulant-generating function defined in eqn (6.7). Hence, for any $\gamma \in \Gamma(z)$,

$$\begin{aligned} I_\rho(\gamma) &= \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k} \;=\; \sum_{k=1}^{\alpha} \gamma_k \log \frac{\gamma_k}{\rho_k^{(\beta)}} + \sum_{k=1}^{\alpha} \gamma_k \log \frac{\rho_k^{(\beta)}}{\rho_k} \\ &= I_{\rho^{(\beta)}}(\gamma) + \beta \sum_{k=1}^{\alpha} y_k \gamma_k - c(\beta)\,. \end{aligned}$$

Since $I_{\rho^{(\beta)}}(\rho^{(\beta)}) = 0$ and $\sum_{k=1}^{\alpha} y_k \rho_k^{(\beta)} = z$, it follows that

$$I_\rho(\rho^{(\beta)}) = I_{\rho^{(\beta)}}(\rho^{(\beta)}) + \beta \sum_{k=1}^{\alpha} y_k \rho_k^{(\beta)} - c(\beta) = \beta z - c(\beta)\,. \tag{6.12}$$

Now consider any $\gamma \in \Gamma(z)$, $\gamma \neq \rho^{(\beta)}$. Since $\beta < 0$, $\sum_{k=1}^{\alpha} y_k \gamma_k \leq z$, and $I_{\rho^{(\beta)}}(\gamma) \geq 0$ with equality if and only if $\gamma = \rho^{(\beta)}$ (Lemma 6.3.1), we obtain

$$\begin{aligned} I_\rho(\gamma) &= I_{\rho^{(\beta)}}(\gamma) + \beta \sum_{k=1}^{\alpha} y_k \gamma_k - c(\beta) \\ &> \beta \sum_{k=1}^{\alpha} y_k \gamma_k - c(\beta) \geq \beta z - c(\beta) = I_\rho(\rho^{(\beta)})\,. \end{aligned}$$

We conclude that for any $\gamma \in \Gamma(z)$, $I_\rho(\gamma) \geq I_\rho(\rho^{(\beta)})$ with equality if and only if $\gamma = \rho^{(\beta)}$. Thus I_ρ attains its infimum over $\Gamma(z)$ at the unique point $\rho^{(\beta)}$. The proof of the proposition and thus the proof of Theorem 6.4.1 are complete. ■

In the next section, we formulate the general concepts of a large deviation principle and a Laplace principle. Subsequent sections will apply the theory of large deviations to study interacting systems in statistical mechanics.

6.5 Generalities: Large deviation principle and Laplace principle

In Theorem 6.3.3, we formulate Sanov's Theorem, which is the large deviation principle for the empirical vectors L_n on the space $\mathcal{P}_\alpha$ of probability vectors in $\mathbb{R}^\alpha$. In Subsection 6.6.2, we apply Sanov's Theorem to analyze the phase-transition structure of a model of ferromagnetism known as the Curie–Weiss–Potts model. Applications of the theory of large deviations to other models in statistical mechanics require large deviation principles in different settings. As we will see in Subsection 6.6.1, analyzing the phase-transition structure of the model of ferromagnetism known as the Curie–Weiss model involves Cramér's Theorem. This theorem states the large deviation principle for the sample means of i.i.d. random variables, which in the case of the Curie–Weiss model take values in the closed interval $[-1, 1]$. Analyzing the Ising model in dimensions $d \geq 2$ is much more complicated. It involves a large deviation principle on the space of translation-invariant probability measures on $\{-1, 1\}^{\mathbb{Z}^d}$ (Ellis 1995, Section 11).

In order to define the general concept of a large deviation principle, we need some notation. First, for each $n \in \mathbb{N}$ let $(\Omega_n, \mathcal{F}_n, P_n)$ be a probability space. Thus Ω_n is a set of points, $\mathcal{F}_n$ is a σ-algebra of subsets of Ω_n, and P_n is a probability measure on $\mathcal{F}_n$. An example is given by the basic model in Section 6.2, where $\Omega_n = \Lambda^n = \{y_1, y_2, \ldots, y_\alpha\}^n$, $\mathcal{F}_n$ is the set of all subsets of Ω_n, and P_n is the product measure with one-dimensional marginals ρ.

Second, let $\mathcal{X}$ be a complete, separable metric space or, as it is often called, a Polish space. Thus there exists a function m, called a metric, mapping $\mathcal{X} \times \mathcal{X}$ into $[0, \infty)$ and having the properties that for all x, y, and z in $\mathcal{X}$, $m(x, y) = m(y, x)$ (symmetry), $m(x, y) = 0 \Leftrightarrow x = y$ (identity), and $m(x, z) \leq m(x, y) + m(y, z)$ (triangle inequality); furthermore, with respect to m, any Cauchy sequence in $\mathcal{X}$ converges to an element of $\mathcal{X}$ (completeness) and $\mathcal{X}$ has a countable dense subset (separability). Elementary examples are $\mathcal{X} = \mathbb{R}^d$ for $d \in \mathbb{N}$; $\mathcal{X} = \mathcal{P}_\alpha$, the set of probability vectors in $\mathbb{R}^\alpha$; and, in the notation of the basic probabilistic model in Section 6.2, $\mathcal{X}$ equal to the closed bounded interval $[y_1, y_\alpha]$. In all three cases the metric is the Euclidean distance.

Third, for each $n \in \mathbb{N}$ let Y_n be a random variable mapping Ω_n into $\mathcal{X}$. For example, in the notation of the basic probability model in Section 6.2, with $\mathcal{X} = \mathcal{P}_\alpha$, let $Y_n = L_n$ or, with $\mathcal{X} = [y_1, y_\alpha]$, let $Y_n = \sum_{j=1}^n X_j/n$, where $X_j(\omega) = \omega_j$ for $\omega \in \Omega_n = \Lambda^n$.

Before continuing, we need several standard definitions from topology. A subset G of $\mathcal{X}$ is said to be open if for any x in G there exists $\varepsilon > 0$ such that the open ball $B(x, \varepsilon) = \{y \in \mathcal{X} : m(y, x) < \varepsilon\}$ is a subset of G. A subset F of $\mathcal{X}$ is said to be closed if the complement, F^c, is open or, equivalently, if for any sequence x_n in F converging to some $x \in \mathcal{X}$, we have $x \in F$. Finally, a subset K of $\mathcal{X}$ is said to be compact if for any sequence x_n in K, there exists a subsequence of x_n converging to a point in K. Equivalently, K is compact if, whenever K is a subset of the union of a collection $\mathcal{C}$ of open sets, then K is a subset of the union of finitely many open sets in $\mathcal{C}$. If $\mathcal{X} = \mathbb{R}^d$, then K is compact if and only if K is closed and bounded.

A class of Polish spaces arising naturally in applications is obtained by taking a Polish space $\mathcal{Y}$ and considering the space $\mathcal{P}(\mathcal{Y})$ of probability measures on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. $\mathcal{B}(\mathcal{Y})$ denotes the Borel σ-algebra of $\mathcal{Y}$, which is the σ-algebra generated by the open subsets of $\mathcal{Y}$. We say that a sequence $\{\Pi_n, n \in \mathbb{N}\}$ in $\mathcal{P}(\mathcal{Y})$ converges weakly to

$\Pi \in \mathcal{P}(\mathcal{Y})$, and write $\Pi_n \Rightarrow \Pi$, if $\int_{\mathcal{Y}} f d\Pi_n \to \int_{\mathcal{Y}} f d\Pi$ for all bounded, continuous functions f mapping $\mathcal{Y}$ into $\mathbb{R}$. A fundamental fact is that there exists a metric m on $\mathcal{P}(\mathcal{Y})$ such that $\Pi_n \Rightarrow \Pi$ if and only if $m(\Pi, \Pi_n) \to 0$ and $\mathcal{P}(\mathcal{Y})$ is a Polish space with respect to m (Ethier and Kurtz 1986, Section 3.1).

Let I be a function mapping the complete, separable metric space $\mathcal{X}$ into $[0, \infty]$. I is called a rate function if I has compact level sets, i.e. for all $M < \infty$, $\{x \in \mathcal{X} : I(x) \leq M\}$ is compact. This technical regularity condition implies that I has closed level sets or, equivalently, that I is lower semicontinuous, i.e. if $x_n \to x$, then $\liminf_{n\to\infty} I(x_n) \geq I(x)$. In particular, if $\mathcal{X}$ is compact, then the lower semicontinuity of I implies that I has compact level sets. When $\mathcal{X} = \mathcal{P}_\alpha$, an example of a rate function is the relative entropy I_ρ with respect to ρ; when $\mathcal{X} = [y_1, y_\alpha]$, any continuous function I mapping $[y_1, y_\alpha]$ into $[0, \infty)$ is a rate function.

We next define the concept of a large deviation principle. If Y_n satisfies the large deviation principle with rate function I, then we summarize this by the formal notation

$$P_n\{Y_n \in dx\} \asymp \exp[-nI(x)]\, dx\,.$$

For any subset A of $\mathcal{X}$, we define $I(A) = \inf_{x\in A} I(x)$.

Definition 6.2. (Large deviation principle.) *Let $\{(\Omega_n, \mathcal{F}_n, P_n), n \in \mathbb{N}\}$ be a sequence of probability spaces, $\mathcal{X}$ a complete, separable metric space, $\{Y_n, n \in \mathbb{N}\}$ a sequence of random variables such that Y_n maps Ω_n into $\mathcal{X}$, and I a rate function on $\mathcal{X}$. Then Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I if the following two limits hold.*

Large deviation upper bound. *For any closed subset F of $\mathcal{X}$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in F\} \leq -I(F)\,.$$

Large deviation lower bound. *For any open subset G of $\mathcal{X}$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in G\} \geq -I(G)\,.$$

We shall explore several consequences of this definition. It is reassuring that a large deviation principle has a unique rate function. The following result is proved in Dupuis and Ellis (1997, Theorem 1.3.1).

Theorem 6.5.1. *If Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I and with rate function J, then $I(x) = J(x)$ for all $x \in \mathcal{X}$.*

The next theorem gives a condition that guarantees the existence of large deviation limits.

Theorem 6.5.2. *Assume that Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I. Let A be a Borel subset of $\mathcal{X}$ having closure $\overline{A}$ and interior A° and satisfying $I(\overline{A}) = I(A^\circ)$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log P\{Y_n \in A\} = -I(A)\,.$$

Proof. We evaluate the large deviation upper bound for $F = \overline{A}$ and the large deviation lower bound for $G = A^\circ$. Since $\overline{A} \supset A \supset A^\circ$, it follows that $I(\overline{A}) \leq I(A) \leq I(A^\circ)$ and

$$\begin{aligned} I(\overline{A}) &\geq \limsup_{n\to\infty} \frac{1}{n} \log P\{Y_n \in \overline{A}\} \geq \limsup_{n\to\infty} \frac{1}{n} \log P\{Y_n \in A\} \\ &\geq \liminf_{n\to\infty} \frac{1}{n} \log P\{Y_n \in A\} \geq \liminf_{n\to\infty} \frac{1}{n} \log P\{Y_n \in A^\circ\} \geq I(A^\circ)\,. \end{aligned}$$

By hypothesis, the two extreme terms are equal to each other and to $I(A)$, and so the theorem follows. ■

The next theorem states useful facts concerning the infimum of a rate function over the entire space and the use of the large deviation principle to show the convergence of a class of probabilities to 0. Part (b) generalizes Corollary 6.3.5.

Theorem 6.5.3. *Suppose that Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I. The following conclusions hold.*

(a) *The infimum of I over $\mathcal{X}$ equals 0, and the set of $x \in \mathcal{X}$ for which $I(x) = 0$ is nonempty and compact.*
(b) *Define $\mathcal{E}$ to be the nonempty, compact set of $x \in \mathcal{X}$ for which $I(x) = 0$, and let A be a Borel subset of $\mathcal{X}$ such that $\overline{A} \cap \mathcal{E} = \emptyset$. Then $I(\overline{A}) > 0$, and for some $C < \infty$,*

$$P_n\{Y_n \in A\} \leq C \exp[-nI(\overline{A})/2] \to 0 \quad \textit{as } n \to \infty\,.$$

Proof.

(a) We evaluate the large deviation upper bound for $F = \mathcal{X}$ and the large deviation lower bound for $G = \mathcal{X}$. Since $P\{Y_n \in \mathcal{X}\} = 1$, we obtain $I(\mathcal{X}) = 0$. We now prove that I attains its infimum of 0 by considering any infimizing sequence x_n that satisfies $I(x_n) \leq 1$ and $I(x_n) \to 0$ as $n \to \infty$. Since I has compact level sets, there exists a subsequence $x_{n'}$ and a point $x \in \mathcal{X}$ such that $x_{n'} \to x$. Hence, by the lower semicontinuity of I,

$$0 = \lim I(x_{n'}) \geq I(x) \geq 0\,.$$

It follows that I attains its infimum of 0 at x and thus that the set of $x \in \mathcal{X}$ for which $I(x) = 0$ is nonempty and compact. This gives part (a).

(b) If $I(\overline{A}) > 0$, then the desired upper bound follows immediately from the large deviation upper bound. We prove that $I(\overline{A}) > 0$ by contradiction. If $I(\overline{A}) = 0$, then there exists a sequence x_n such that $\lim_{n\to\infty} I(x_n) = 0$. Since I has compact level sets and $\overline{A}$ is closed, there exists a subsequence $x_{n'}$ converging to an element $x \in \overline{A}$. Since I is lower semicontinuous, it follows that $I(x) = 0$ and thus that $x \in \mathcal{E}$. This contradicts the assumption that $\overline{A} \cap \mathcal{E} = \emptyset$. The proof of the proposition is complete. ■

We next state Cramér's Theorem, which is the large deviation principle for the sample means of i.i.d. random variables taking values in $\mathbb{R}^d$. The rate function is defined by a variational formula that in general cannot be evaluated explicitly. We denote by $\langle\cdot,\cdot\rangle$ the inner product on $\mathbb{R}^d$. The theorem is derived from the Gärtner–Ellis Theorem in Ellis (2006, Section VII.5); a direct proof is given in Ellis (2008, Section 7).

Theorem 6.5.4. (Cramér's Theorem.) *Let $\{X_j, j \in I\!N\}$ be a sequence of i.i.d. random vectors taking values in $\mathbb{R}^d$ and satisfying $E\{\exp\langle t, X_1\rangle\} < \infty$ for all $t \in \mathbb{R}^d$. We define the sample means $S_n/n = \sum_{j=1}^n X_j/n$ and the cumulant-generating function $c(t) = \log E\{\exp\langle t, X_1\rangle\}$. The following conclusions hold.*

(a) *The sequence of sample means S_n/n satisfies the large deviation principle on $\mathbb{R}^d$ with rate function $I(x) = \sup_{t\in\mathbb{R}^d}\{\langle t, x\rangle - c(t)\}$.*
(b) *I is a convex, lower semicontinuous function on $\mathbb{R}^d$, and it attains its infimum of 0 at the unique point $x_0 = E\{X_1\}$.*

A nice application of Cramér's Theorem is to derive the special case of Sanov's Theorem given in Theorem 6.3.3 (Ellis 2006, Section VIII.2). The latter states the large deviation principle for the empirical vectors of i.i.d. random variables having a finite state space. For application in Subsection 6.6.1, we next state a special case of Cramér's Theorem, for which the rate function can be given explicitly.

Corollary 6.5.5. *In the basic probability model of Section 6.2, let $\Lambda = \{-1, 1\}$ and $\rho = (\frac{1}{2}, \frac{1}{2})$, which corresponds to the probability measure $\rho = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ on Λ. For $\omega \in \Omega_n$, define $S_n(\omega) = \sum_{j=1}^n \omega_j$. Then the sequence of sample means S_n/n satisfies the large deviation principle on the closed interval $[-1, 1]$ with rate function*

$$I(x) = \tfrac{1}{2}(1-x)\log(1-x) + \tfrac{1}{2}(1+x)\log(1+x)\,. \tag{6.13}$$

Proof. In this case, $c(t) = \log(\frac{1}{2}[e^t + e^{-t}]$. The function $c(t)$ satisfies $c''(t) > 0$ for all t, and the range of c' equals $(-1, 1)$. Hence, for any $x \in (-1, 1)$, the supremum in the definition of I is attained at the unique $t = t(x)$ satisfying $c'(t(x)) = x$. One can easily verify that $t(x) = \frac{1}{2}\log[(1+x)/(1-x)]$ and that $I(x) = t(x)\cdot x - c(t(x))$ is given by eqn (6.13). When $x = 1$ or $x = -1$, the supremum in the definition of $I(x)$ is not attained but equals $\log 2$, which coincides with the value of the right side of eqn (6.13). ■

Corollary 6.5.5 is easy to motivate using the formal notation of Pseudo-theorem 6.3.2. For any $x \in [-1, 1]$, $S_n(\omega)/n \sim x$ if and only if approximately $(n/2)(1-x)$ of the ω_j's equal -1 and approximately $(n/2)(1+x)$ of the ω_j's equal 1. For any probability vector $\gamma = (\gamma_1, \gamma_2)$,

$$I_\rho(\gamma) = \gamma_1\log(2\gamma_1) + \gamma_2\log(2\gamma_2)\,.$$

Hence

$$\begin{aligned} P_n\{S_n/n \sim x\} &\approx P_n\{L_n(-1) = \tfrac{1}{2}(1-x), L_n(1) = \tfrac{1}{2}(1+x)\} \\ &\approx \exp[-nI_\rho(\tfrac{1}{2}(1-x), \tfrac{1}{2}(1+x))] = \exp[-nI(x)]\,. \end{aligned}$$

We next state a general version of Sanov's Theorem, which gives the large deviation principle for the sequence of empirical measures of i.i.d. random variables. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{Y}$ a complete, separable metric space, ρ a probability measure on $\mathcal{Y}$, and $\{X_j, j \in I\!\!N\}$ a sequence of i.i.d. random variables mapping Ω into $\mathcal{Y}$ and having the common distribution ρ. For $n \in I\!\!N$, $\omega \in \Omega$, and A any Borel subset of $\mathcal{Y}$, we define the empirical measure

$$L_n(A) = L_n(\omega, A) = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j(\omega)}\{A\}\,,$$

where for $y \in \mathcal{Y}$, $\delta_y\{A\}$ equals 1 if $y \in A$ and 0 if $y \notin A$. For each ω, $L_n(\omega, \cdot)$ is a probability measure on $\mathcal{Y}$. Hence the sequence L_n takes values in the complete, separable metric space $\mathcal{P}(\mathcal{Y})$. For $\gamma \in \mathcal{P}(\mathcal{Y})$, we write $\gamma \ll \rho$ if γ is absolutely continuous with respect to ρ, i.e. if $\rho\{A\} = 0$ for a Borel subset A of $\mathcal{Y}$, then $\gamma\{A\} = 0$. If $\gamma \ll \rho$, then $d\gamma/d\rho$ denotes the Radon–Nikodym derivative of γ with respect to ρ.

Theorem 6.5.6. (Sanov's Theorem.) *The sequence L_n satisfies the large deviation principle on $\mathcal{P}(\mathcal{Y})$ with a rate function given by the relative entropy with respect to ρ. For $\gamma \in \mathcal{P}(\mathcal{Y})$, this quantity is defined by*

$$I_\rho(\gamma) = \begin{cases} \displaystyle\int_{\mathcal{Y}} \left(\log \frac{d\gamma}{d\rho}\right) d\gamma & \textit{if } \gamma \ll \rho\,, \\ \infty & \textit{otherwise}\,. \end{cases}$$

This theorem is proved, for example, in Dembo and Zeitouni (1998, Section 6.2) and in Dupuis and Ellis (1997, Chapter 2). If the support of ρ is a finite set $\Lambda \subset \mathbb{R}$, then Theorem 6.5.6 reduces to Theorem 6.3.3.

The concept of a Laplace principle will be useful in the analysis of statistical-mechanical models. As we will see in Section 6.7, where a general class of statistical-mechanical models are studied, the Laplace principle gives a variational formula for the canonical free energy (Theorem 6.7.1(a)).

Definition 6.3. (Laplace principle.) *Let $\{(\Omega_n, \mathcal{F}_n, P_n), n \in I\!\!N\}$ be a sequence of probability spaces, $\mathcal{X}$ a complete, separable metric space, $\{Y_n, n \in I\!\!N\}$ a sequence of random variables such that Y_n maps Ω_n into $\mathcal{X}$, and I a rate function on $\mathcal{X}$. Then Y_n satisfies the Laplace principle on $\mathcal{X}$ with rate function I if for all bounded, continuous functions f mapping $\mathcal{X}$ into $\mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \log E^{P_n}\{\exp[nf(Y_n)]\} = \sup_{x\in\mathcal{X}}\{f(x) - I(x)\}\,.$$

Suppose that Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I. Then substituting $P_n\{Y_n \in dx\} \asymp \exp[-nI(x)]\,dx$ gives

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\log E^{P_n}\{\exp[nf(Y_n)]\} &= \lim_{n\to\infty}\frac{1}{n}\log\int_{\Omega_n}\exp[nf(Y_n)]\,dP_n \\
&= \lim_{n\to\infty}\frac{1}{n}\log\int_{\mathcal{X}}\exp[nf(x)]\,P_n\{Y_n\in dx\} \\
&\approx \lim_{n\to\infty}\frac{1}{n}\log\int_{\mathcal{X}}\exp[nf(x)]\,\exp[-nI(x)]\,dx \\
&= \lim_{n\to\infty}\frac{1}{n}\log\int_{\mathcal{X}}\exp[n(f(x)-I(x)]\,dx\,.
\end{aligned}$$

Since the asymptotic behavior of the last integral is determined by the largest value of the integrand, the last line of the above equation suggests the following limit:

$$\lim_{n\to\infty}\frac{1}{n}\log E^{P_n}\{\exp[nf(Y_n)]\} = \sup_{x\in\mathcal{X}}\{f(x)-I(x)\}\,.$$

Hence it is plausible that Y_n satisfies the Laplace principle with rate function I, a fact first proved by Varadhan (1966). In fact, we have the following stronger result, which shows that the large deviation principle and the Laplace principle are equivalent.

Theorem 6.5.7. *Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I if and only if Y_n satisfies the Laplace principle on $\mathcal{X}$ with rate function I.*

We have just motivated the suggestion that the large deviation principle with rate function I implies the Laplace principle with the same rate function. In order to motivate the converse, let A be an arbitrary Borel subset of $\mathcal{X}$ and consider the function

$$\varphi_A = \begin{cases} 0 & \text{if } x\in A\,, \\ -\infty & \text{if } x\in A^c\,. \end{cases}$$

Clearly φ_A is not a bounded, continuous function on $\mathcal{X}$. If it were, then evaluating the Laplace expectation $E\{\exp[nf(Y_n)]\}$ for $f=\varphi_A$ would yield the large deviation limit

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\log P_n\{Y_n\in A\} &= \lim_{n\to\infty}\frac{1}{n}\log E^{P_n}\{\exp[n\varphi_A(Y_n)]\} \\
&= \sup_{x\in A}\{\varphi_A(x)-I(x)\} \\
&= -\inf_{x\in A} I(x) = -I(A)\,.
\end{aligned}$$

The proof that the Laplace principle implies the large deviation principle involves approximating φ_A by suitable bounded, continuous functions. Theorem 6.5.7 is proved in Dupuis and Ellis (1997, Theorems 1.2.1 and 1.2.3).

We end this section by presenting two ways to obtain large deviation principles from existing large deviation principles. In the first theorem, we show that a large deviation principle is preserved under continuous mappings.

Theorem 6.5.8. (Contraction principle.) *Assume that Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I and that ψ is a continuous function mapping*

$\mathcal{X}$ into a complete, separable metric space $\mathcal{Y}$. Then $\psi(Y_n)$ satisfies the large deviation principle on $\mathcal{Y}$ with rate function

$$J(y) = \inf\{I(x) : x \in \mathcal{X}, \psi(x) = y\} = \inf\{I(x) : x \in \psi^{-1}(y)\}.$$

Proof. Since I maps $\mathcal{X}$ into $[0,\infty]$, J maps $\mathcal{Y}$ into $[0,\infty]$. The fact that J has compact level sets in $\mathcal{Y}$ follows from the definition of J (Ellis 2008, Theorem 6.12). The large deviation upper bound is proved next. If F is a closed subset of $\mathcal{Y}$, then since ψ is continuous, $\psi^{-1}(F)$ is a closed subset of $\mathcal{X}$. Hence, by the large deviation upper bound for Y_n,

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log P_n\{\psi(Y_n) \in F\} &= \limsup_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in \psi^{-1}(F)\} \\ &\leq - \inf_{x \in \psi^{-1}(F)} I(x) \\ &= - \inf_{y \in F} \{\inf\{I(x) : x \in \mathcal{X}, \psi(x) = y\} \\ &= - \inf_{y \in F} J(y) = -J(F).\end{aligned}$$

The large deviation lower bound is proved similarly. The proof of the theorem is complete. ■

In the next theorem, we show that a large deviation principle is preserved if the probability measures P_n are multiplied by suitable exponential factors and then normalized. This result will be applied in Sections 6.6 and 6.7 when we prove the large deviation principle for statistical-mechanical models with respect to the canonical ensemble (Theorems 6.6.1, 6.6.3, 6.6.5, and 6.7.1).

Theorem 6.5.9. *Assume that with respect to the probability measures P_n, Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I. Let ψ be a bounded, continuous function mapping $\mathcal{X}$ into $\mathbb{R}$. For $A \in \mathcal{F}_n$, we define new probability measures*

$$P_{n,\psi}\{A\} = \frac{1}{\int_{\mathcal{X}} \exp[-n\psi(Y_n)]\, dP_n} \cdot \int_A \exp[-n\psi(Y_n)]\, dP_n.$$

Then, with respect to $P_{n,\psi}$, Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function

$$I_\psi(x) = I(x) + \psi(x) - \inf_{y \in \mathcal{X}} \{I(y) + \psi(y)\}.$$

Proof. We omit the straightforward proof showing that since I has compact level sets and ψ is bounded and continuous, I_ψ has compact level sets. Since I_ψ maps $\mathcal{X}$ into $[0,\infty]$, it follows that I_ψ is a rate function. We complete the proof by showing that with respect to $P_{n,\psi}$, Y_n satisfies the equivalent Laplace principle with rate function I_ψ (Theorem 6.5.7). Let f be any bounded, continuous function mapping $\mathcal{X}$ into $\mathbb{R}$.

Since $f + \psi$ is bounded and continuous and since, with respect to P_n, Y_n satisfies the Laplace principle with rate function I, it follows that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log \int_{\Omega_n} \exp[nf(Y_n)]\, dP_{n,\psi} \\
&= \lim_{n\to\infty} \frac{1}{n} \log \int_{\Omega_n} \exp[n(f(Y_n) - \psi(Y_n))]\, dP_n \\
&\qquad - \lim_{n\to\infty} \frac{1}{n} \log \int_{\Omega_n} \exp[-n\psi(Y_n)]\, dP_n \\
&= \sup_{x\in\mathcal{X}} \{f(x) - \psi(x) - I(x)\} - \sup_{y\in\mathcal{X}} \{-\psi(y) - I(y)\} \\
&= \sup_{x\in\mathcal{X}} \{f(x) - I_\psi(x)\}\,.
\end{aligned}$$

Thus, with respect to $P_{n,\psi}$, Y_n satisfies the Laplace principle with rate function I_ψ, as claimed. This completes the proof. ■

This completes our discussion of the large deviation principle, the Laplace principle, and related general results. In the next section, we begin our study of statistical-mechanical models by considering the Curie–Weiss spin model and other mean-field models.

6.6 The Curie–Weiss model and other mean-field models

Mean-field models in statistical mechanics lend themselves naturally to a large deviation analysis. We illustrate this by first studying the Curie–Weiss model of ferromagnetism, one of the simplest examples of an interacting system in statistical mechanics. After treating this model, we also outline a large deviation analysis of two other mean-field models, the Curie–Weiss–Potts model and the mean-field Blume–Capel model. As we will see in the next section, using the theory of large deviations to analyze these models suggests how one can apply the theory to analyze much more complicated models.

6.6.1 Curie–Weiss model

The Curie–Weiss model is a spin model defined on the complete graph on n vertices $1, 2, \ldots, n$. It is a mean-field approximation to the Ising model and related, short-range, ferromagnetic models (Ellis 2006, Section V.9). In the Curie–Weiss model the spin at site $j \in \{1, 2, \ldots, n\}$ is denoted by ω_j, a quantity taking values in $\Lambda = \{-1, 1\}$; the value -1 represents spin-down and the value 1 spin-up. The configuration space for the model is the set $\Omega_n = \Lambda^n$, containing all configurations or microstates $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ with each $\omega_j \in \Lambda$.

Let $\rho = \frac{1}{2}(\delta_{-1} + \delta_1)$ and let P_n denote the product measure on Ω_n with one-dimensional marginals ρ. Thus $P_n\{\omega\} = 1/2^n$ for each $\omega \in \Omega_n$. For $\omega \in \Omega_n$, we define the spin per site $S_n(\omega)/n = \sum_{j=1}^n \omega_j/n$. The Hamiltonian, or energy function, is defined by

$$H_n(\omega) = -\frac{1}{2n}\sum_{i,j=1}^{n}\omega_i\omega_j = -\frac{n}{2}\left(\frac{S_n(\omega)}{n}\right)^2 , \tag{6.14}$$

and the probability of ω corresponding to an inverse temperature $\beta > 0$ is defined by the canonical ensemble

$$\begin{aligned} P_{n,\beta}\{\omega\} &= \frac{1}{Z_n(\beta)} \cdot \exp[-\beta H_n(\omega)]\, P_n\{\omega\} \\ &= \frac{1}{Z_n(\beta)} \cdot \exp\left[\frac{n\beta}{2}\left(\frac{S_n(\omega)}{n}\right)^2\right] P_n\{\omega\} , \end{aligned} \tag{6.15}$$

where $Z_n(\beta)$ is the partition function

$$Z_n(\beta) = \int_{\Omega_n} \exp[-\beta H_n(\omega)]\, P_n(d\omega) = \int_{\Omega_n} \exp\left[\frac{n\beta}{2}\left(\frac{S_n(\omega)}{n}\right)^2\right] P_n(d\omega) .$$

$P_{n,\beta}$ models a ferromagnet in the sense that the maximum of $P_{n,\beta}\{\omega\}$ over $\omega \in \Omega_n$ occurs at the two microstates having all coordinates ω_i equal to -1 or all coordinates equal to 1. Furthermore, as $\beta \to \infty$ all the mass of $P_{n,\beta}$ concentrates on these two microstates.

A distinguishing feature of the Curie–Weiss model is its phase transition. Namely, the alignment effects incorporated in the canonical ensemble $P_{n,\beta}$ persist in the limit $n \to \infty$. This is most easily seen by evaluating the $n \to \infty$ limit of the distributions $P_{n,\beta}\{S_n/n \in dx\}$. For $\beta \leq 1$ the alignment effects are relatively weak, and as we will see, S_n/n satisfies a law of large numbers, concentrating on the value 0. However, for $\beta > 1$ the alignment effects are so strong that the law of large numbers breaks down, and the limiting $P_{n,\beta}$ distribution of S_n/n concentrates on two points $\pm m(\beta)$ for some $m(\beta) \in (0,1)$ (see eqn (6.17)). The analysis of the Curie–Weiss model to be presented below can be easily modified to handle an external magnetic field h. The resulting probabilistic description of the phase transition yields the predictions of mean-field theory (Ellis 2006, Section V.9; Parisi 1988, Section 3.2).

We calculate the $n \to \infty$ limit of $P_{n,\beta}\{S_n/n \in dx\}$ by establishing a large deviation principle for the spin per site S_n/n with respect to $P_{n,\beta}$. For each n, S_n/n takes values in $[-1,1]$. According to the special case of Cramér's Theorem given in Corollary 6.5.5, with respect to the product measures P_n, S_n/n satisfies the large deviation principle on $[-1,1]$ with rate function

$$I(x) = \tfrac{1}{2}(1-x)\log(1-x) + \tfrac{1}{2}(1+x)\log(1+x) . \tag{6.16}$$

Because of the form of $P_{n,\beta}$ given in eqn (6.15), the large deviation principle for S_n/n with respect to $P_{n,\beta}$ is an immediate consequence of Theorem 6.5.9 with $\mathcal{X} = [-1,1]$ and $\psi(x) = -\frac{1}{2}\beta x^2$ for $x \in [-1,1]$. The proof of that theorem uses the exponential form of the measures to derive the equivalent Laplace principle. We record the large deviation principle in the next theorem.

Theorem 6.6.1. *With respect to the canonical ensemble $P_{n,\beta}$ defined in eqn (6.15), the spin per site S_n/n satisfies the large deviation principle on $[-1,1]$ with rate function*

$$I_\beta(x) = I(x) - \tfrac{1}{2}\beta x^2 - \inf_{y\in[-1,1]}\{I(y) - \tfrac{1}{2}\beta y^2\}\,,$$

where $I(x)$ is defined in eqn (6.16).

The limiting behavior of the distributions $P_{n,\beta}\{S_n/n \in dx\}$ is now determined by examining where the nonnegative rate function I_β attains its infimum of 0 or, equivalently, where $I(x) - \frac{1}{2}\beta x^2$ attains its infimum $\inf_{y\in[-1,1]}\{I(y) - \frac{1}{2}\beta y^2\}$ (Ellis 2006, Section IV.4). Global minimum points x^* satisfy

$$I_\beta^{'}(x^*) = 0 \quad \text{or} \quad I^{'}(x^*) = \beta x^*\,.$$

The second equation is equivalent to the mean-field equation $x^* = (I^{'})^{-1}(\beta x^*) = \tanh(\beta x^*)$ (Ellis 2006, Section V.9; Parisi 1988, Section 3.2). Figure 6.1 motivates the next theorem, which is a consequence of the following easily verified properties of I: (a) $I^{''}(0) = 1$; (b) $I^{'}$ is convex on $[0,1]$ and $\lim_{x\to 1} I^{'}(x) = \infty$; (c) $I^{'}$ is concave on $[-1,0]$ and $\lim_{x\to -1} I^{'}(x) = -\infty$.

Theorem 6.6.2. *For each $\beta > 0$, we define*

$$\begin{aligned}\mathcal{E}_\beta &= \{x \in [-1,1] : I_\beta(x) = 0\}\\ &= \left\{x \in [-1,1] : I(x) - \tfrac{1}{2}\beta x^2 \text{ is minimized}\right\}.\end{aligned}$$

The following conclusions hold.

(a) *For $0 < \beta \le 1$, $\mathcal{E}_\beta = \{0\}$.*
(b) *For $\beta > 1$, there exists $m(\beta) > 0$ such that $\mathcal{E}_\beta = \{\pm m(\beta)\}$. The function $m(\beta)$ is monotonically increasing on $(1,\infty)$ and satisfies $m(\beta) \to 0$ as $\beta \to 1^+$, and $m(\beta) \to 1$ as $\beta \to \infty$.*

According to part (b) of Theorem 6.5.3, if A is any closed subset of $[-1,1]$ such that $A \cap \mathcal{E}_\beta = \emptyset$, then $I(A) > 0$ and, for some $C < \infty$,

$$P_{n,\beta}\{S_n/n \in A\} \le C\,\exp[-nI(A)/2] \to 0 \text{ as } n \to \infty\,.$$

In combination with Theorem 6.6.2, we are led to the following weak limits:

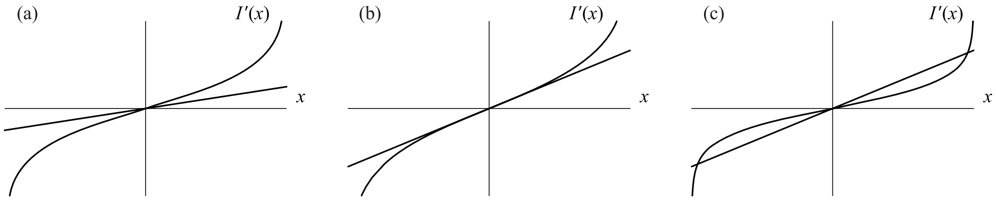

Fig. 6.1 Solutions of $I^{'}(x^*) = \beta x^*$: (a) $\beta < 1$, (b) $\beta = 1$, (c) $\beta > 1$.

$$P_{n,\beta}\{S_n/n \in dx\} \Longrightarrow \begin{cases} \delta_0 & \text{if } 0 < \beta \le 1\,, \\ \frac{1}{2}\delta_{m(\beta)} + \frac{1}{2}\delta_{-m(\beta)} & \text{if } \beta > 1\,. \end{cases} \tag{6.17}$$

We call $m(\beta)$ the spontaneous magnetization for the Curie–Weiss model and $\beta_c = 1$ the critical inverse temperature (Ellis 2006, Section IV.4). It is worth remarking that it is much easier to derive the weak limits in eqn (6.17) from the large deviation principle for S_n/n with respect to $P_{n,\beta}$ rather than to prove the weak limits directly.

The limit (6.17) justifies calling $\mathcal{E}_\beta$ the set of canonical equilibrium macrostates for the spin per site S_n/n in the Curie–Weiss model. Because $m(\beta) \to 0$ as $\beta \to 1^+$ and 0 is the unique equilibrium macrostate for $0 < \beta \le 1$, the phase transition at β_c is said to be continuous or second-order.

The phase transition in the Curie–Weiss model and in related models arises as a result of two competing microscopic effects. The first effect tends to randomize the system. It is caused by thermal excitations and is measured by entropy. The second effect tends to order the system. It is caused by attractive forces of interaction and is measured by the energy. At sufficiently low temperatures and thus for sufficiently large values of β, the energy effect predominates and a phase transition becomes possible. This phenomenology is reflected in the form of $\mathcal{E}_\beta$ given in Theorem 6.6.2. For $0 < \beta \le 1$, I_β has a unique global minimum point coinciding with the unique global minimum point of I at 0. For $\beta > 1$, I_β has two global minimum points at $\pm m(\beta)$, a structure that is consistent with the facts that $-\frac{1}{2}\beta x^2$ has two global minimum points on $[-1, 1]$ at 1 and -1 and that $m(\beta) \to 1$ as $\beta \to \infty$. For x near 0, $I(x) \sim \frac{1}{2}x^2 + \frac{1}{12}x^4$ and thus $I(x) - \frac{1}{2}\beta x^2 \sim \frac{1}{2}(1-\beta)x^2 + \frac{1}{12}x^4$. The set of global minimum points of the latter function bifurcates continuously from $\{0\}$ for $\beta \le 1$ to $\{\pm\sqrt{3(\beta-1)}\}$ for $\beta > 1$. This behavior is consistent with the continuous phase transition described in Theorem 6.6.2 and suggests that as $\beta \to 1^+$, $m(\beta) \sim \sqrt{3(\beta-1)} \to 0$.

Before we leave the Curie–Weiss model, there are several additional points that should be emphasized. The first is to focus on what makes possible the large deviation analysis of the phase transition in the model. In eqn (6.14), we write the Hamiltonian as a quadratic function of the spin per site S_n/n, which by the version of Cramér's Theorem given in Corollary 6.5.5 satisfies the large deviation principle on $[-1, 1]$ with respect to the product measures P_n. The equivalent Laplace principle allows us to convert this large deviation principle into a large deviation principle with respect to the canonical ensemble $P_{n,\beta}$. The form of the rate function I_β allows us to complete the analysis. In this context, the sequence S_n/n is called the sequence of macroscopic variables for the Curie–Weiss model. In the next section, we will generalize these steps to formulate a large deviation approach to a wide class of models in statistical mechanics.

Our large deviation analysis of the phase transition in the Curie–Weiss model has the attractive feature that it directly motivates us to attach physical importance to $\mathcal{E}_\beta$. This set is the support of the $n \to \infty$ limit of the distributions $P_{n,\beta}\{S_n/n \in dx\}$. An analogous fact is true for a large class of statistical-mechanical models (Ellis *et al.* 2000, Theorem 2.5).

As shown in eqn (6.17), the large deviation analysis of the Curie–Weiss model yields the limiting behavior of the $P_{n,\beta}$ distributions of S_n/n. For $0 < \beta \le 1$, this limit corresponds to the classical weak law of large numbers for the sample means of i.i.d. random variables and suggests examining the analogues of other classical limit results such as the central limit theorem. We end this section by summarizing these limit results for the Curie–Weiss model, referring the reader to (Ellis 2006, Section V.9) for proofs. If $\theta \in (0,1)$ and f is a nonnegative integrable function on $\mathbb{R}$, then the notation $P_{n,\beta}\{S_n/n^\theta \in dx\} \Longrightarrow f(x)\,dx$ means that the distributions of S_n/n^θ converge weakly to the probability measure on $\mathbb{R}$ having a density proportional to f with respect to the Lebesgue measure.

In the Curie–Weiss model for $0 < \beta < 1$, the alignment effects among the spins are relatively weak, and the analogue of the central limit theorem holds (Ellis 2006, Theorem V.9.4):

$$P_{n,\beta}\left\{\frac{S_n}{n^{1/2}} \in dx\right\} \Longrightarrow \exp\left[\frac{-\frac{1}{2}x^2}{\sigma^2(\beta)}\right] dx\,,$$

where $\sigma^2(\beta) = 1/(1-\beta)$. However, when $\beta = \beta_c = 1$, the limiting variance $\sigma^2(\beta)$ diverges, and the central-limit scaling $n^{1/2}$ must be replaced by $n^{3/4}$, which reflects the onset of long-range order at β_c. In this case we have (Ellis 2006, Theorem V.9.5)

$$P_{n,\beta_c}\left\{\frac{S_n}{n^{3/4}} \in dx\right\} \Longrightarrow \exp\left[-\frac{1}{12}x^4\right] dx\,.$$

Finally, for $\beta > \beta_c$, $(S_n - n\tilde{z})/n^{1/2}$ satisfies a central-limit-type theorem when S_n/n is conditioned to lie in a sufficiently small neighborhood of $\tilde{z} = m(\beta)$ or $\tilde{z} = -m(\beta)$; see Theorem 2.4 in Ellis *et al.* (1980) with $k = 1$.

The results discussed in this subsection have been extensively generalized to a number of models, including the Curie–Weiss–Potts model (Costeniuc *et al.* 2005*a*; Ellis and Wang 1990), the mean-field Blume–Capel model (Costeniuc *et al.* 2007; Ellis *et al.* 2005), and the Ising and related models (Dobrushin and Shlosman 1994; Föllmer and Orey 1987; Olla 1988), which exhibit much more complicated behavior and are much harder to analyze. For the Ising and related models, refined large deviations at the surface level have been studied; see Dembo and Zeitouni (1998, p. 339) for references.

This completes our discussion of the Curie–Weiss model. In order to reinforce our understanding of the large deviation analysis of that model, in the next two subsections we present the large deviation analysis of the Curie–Weiss–Potts model and the Blume–Capel model. This, in turn, will yield the phase-transition structure of the two models as in Theorem 6.6.2.

6.6.2 Curie–Weiss–Potts model

Let $q \geq 3$ be a fixed integer, and define $\Lambda = \{y_1, y_2, \ldots, y_q\}$, where the y_k are any q distinct vectors in $\mathbb{R}^q$; the precise values of these vectors are immaterial. The Curie–Weiss–Potts model is a spin model defined on the complete graph on n vertices $1, 2, \ldots, n$. It is a mean-field approximation to the well-known Potts model (Wu 1982). In the Curie–Weiss–Potts model, the spin at site $j \in \{1, 2, \ldots, n\}$ is denoted by ω_j, a quantity taking values in Λ. The configuration space for the model is the set $\Omega_n = \Lambda^n$ containing all microstates $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ with each $\omega_j \in \Lambda$.

Let $\rho = (1/q)\sum_{i=1}^q \delta_{y_i}$ and let P_n denote the product measure on Ω_n with one-dimensional marginals ρ. Thus $P_n\{\omega\} = 1/q^n$ for each configuration $\omega = \{\omega_i, i =$

$1, \ldots, n\} \in \Omega_n$. We also denote by ρ the probability vector in $\mathbb{R}^q$ all of whose coordinates equal q^{-1}. For $\omega \in \Omega_n$, the Hamiltonian is defined by

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

where $\delta(\omega_i, \omega_j)$ equals 1 if $\omega_i = \omega_j$ and equals 0 otherwise. The probability of ω corresponding to an inverse temperature $\beta > 0$ is defined by the canonical ensemble

$$P_{n,\beta}\{\omega\} = \frac{1}{Z_n(\beta)} \cdot \exp[-\beta H_n(\omega)]\, P_n\{\omega\}\,, \tag{6.18}$$

where $Z_n(\beta)$ is the partition function

$$Z_n(\beta) = \int_{\Omega_n} \exp[-\beta H_n(\omega)]\, P_n(d\omega) = \sum_{\omega \in \Omega_n} \exp[-\beta H_n(\omega)]\, \frac{1}{2^n}\,.$$

In order to carry out a large deviation analysis of the Curie–Weiss–Potts model, we rewrite the sequence of Hamiltonians H_n as a function of a sequence of macroscopic variables, which is a sequence of random variables that satisfies a large deviation principle. This sequence is the sequence of empirical vectors

$$L_n = L_n(\omega) = (L_n(\omega, y_1), L_n(\omega, y_2), \ldots, L_n(\omega, y_q))\,,$$

the kth component of which is defined by

$$L_n(\omega, y_k) = \frac{1}{n} \sum_{j=1}^{n} \delta(\omega_j, y_k)\,.$$

This quantity equals the relative frequency with which y_k appears in the configuration ω. The empirical vectors L_n take values in the set of probability vectors

$$\mathcal{P}_q = \left\{ \nu \in \mathbb{R}^q : \nu = (\nu_1, \nu_2, \ldots, \nu_q) : \nu_k \geq 0, \sum_{k=1}^{q} \nu_k = 1 \right\}.$$

Let $\langle \cdot, \cdot \rangle$ denote the inner product on $\mathbb{R}^q$. Since

$$\langle L_n(\omega), L_n(\omega) \rangle = \frac{1}{n^2} \sum_{i,j=1}^{n} \left(\sum_{k=1}^{q} \delta(\omega_i, y_k) \cdot \delta(\omega_j, y_k) \right) = \frac{1}{n^2} \sum_{i,j=1}^{n} \delta(\omega_i, \omega_j)\,,$$

it follows that the Hamiltonian and the canonical ensemble can be rewritten as

$$H_n(\omega) = -\frac{1}{2n} \sum_{i,j=1}^{n} \delta(\omega_i, \omega_j) = -\frac{n}{2} \langle L_n(\omega), L_n(\omega) \rangle$$

and

$$P_{n,\beta}\{\omega\} = \frac{1}{\int_{\Omega_n} \exp[(n\beta/2)\langle L_n(\omega), L_n(\omega)\rangle]\, P_n\{\omega\}} \cdot \exp\left[\frac{n\beta}{2} \langle L_n(\omega), L_n(\omega) \rangle \right] P_n\{\omega\}\,.$$

We next establish a large deviation principle for L_n with respect to $P_{n,\beta}$. According to the special case of Sanov's Theorem given in Theorem 6.3.3, with respect to the product measures P_n, L_n satisfies the large deviation principle on $\mathcal{P}_\alpha$ with the rate function being the relative entropy I_ρ. Because of the form of $P_{n,\beta}$ given in the last equation, the large deviation principle for L_n with respect to $P_{n,\beta}$ is an immediate consequence of Theorem 6.5.9 with $\mathcal{X} = \mathcal{P}_q$ and $\psi(\nu) = -\frac{1}{2}\beta\langle \nu, \nu \rangle$ for $\nu \in \mathcal{P}_q$. We record the large deviation principle in the next theorem.

Theorem 6.6.3. *With respect to the canonical ensemble $P_{n,\beta}$ defined in eqn (6.18), the empirical vector L_n satisfies the large deviation principle on $\mathcal{P}_\alpha$ with rate function*

$$I_\beta(\nu) = I_\rho(\nu) - \tfrac{1}{2}\beta\langle \nu, \nu \rangle - \inf_{\gamma \in \mathcal{P}_q}\{I_\rho(\gamma) - \tfrac{1}{2}\beta\langle \gamma, \gamma \rangle\}\,.$$

As in the Curie–Weiss model, we define the set $\mathcal{E}_\beta$ of canonical equilibrium macrostates for the empirical vector L_n in the Curie–Weiss–Potts model to be the zero set of the rate function I_β or, equivalently, the set of $\nu \in \mathcal{P}_q$ at which $I_\rho(\nu) - \frac{1}{2}\beta\langle \nu, \nu \rangle$ attains its minimum. Thus

$$\begin{aligned} \mathcal{E}_\beta &= \{\nu \in \mathcal{P}_q : I_\beta(\nu) = 0\} \\ &= \left\{\nu \in \mathcal{P}_q : I_\rho(\nu) - \tfrac{1}{2}\beta\langle \nu, \nu \rangle \text{ is minimized}\right\}. \end{aligned}$$

The structure of $\mathcal{E}_\beta$ given in the next theorem is consistent with the entropy–energy competition underlying the phase transition. Let β_c be the critical inverse temperature given in the theorem. For $0 < \beta < \beta_c$, I_β has a unique global minimum point coinciding with the unique global minimum point of I_ρ at $\rho = (q^{-1}, q^{-1}, \ldots, q^{-1})$. For $\beta > \beta_c$, I_β has q global minimum points, a structure that is consistent with the fact that $-\frac{1}{2}\beta\langle \gamma, \gamma \rangle$ has q global minimum points in $\mathcal{P}_q$ at the vectors that equal 1 in the kth coordinate and 0 in the other coordinates for $i = 1, 2, \ldots, q$. As β increases through β_c, $\mathcal{E}_\beta$ bifurcates discontinuously from $\{\rho\}$ for $0 < \beta < \beta_c$ to a set containing $q + 1$ distinct points for $\beta = \beta_c$ and to a set containing q distinct points for $\beta > \beta_c$. Because of this behavior, the Curie–Weiss–Potts model is said to have a discontinuous or first-order phase transition at β_c. The structure of $\mathcal{E}_\beta$ is given in terms of the function $\varphi : [0, 1] \to \mathcal{P}_q$ defined by

$$\varphi(s) = (q^{-1}[1 + (q-1)s], q^{-1}(1-s), \ldots, q^{-1}(1-s))\,;$$

the last $(q-1)$ components all equal $q^{-1}(1-s)$. We note that $\varphi(0) = \rho$.

Theorem 6.6.4. *We fix a positive integer $q \geq 3$. Let $\beta_c = (2(q-1)/(q-2))\log(q-1)$ and, for $\beta > 0$, let $s(\beta)$ be the largest solution of the equation $s = (1 - e^{-\beta s})/(1 + (q-1)e^{-\beta s})$. The following conclusions hold.*

(a) *The quantity $s(\beta)$ is well defined. It is positive, strictly increasing, and differentiable in β on an open interval containing $[\beta_c, \infty)$, $s(\beta_c) = (q-2)/(q-1)$, and $\lim_{\beta \to \infty} s(\beta) = 1$.*

(b) *For $\beta \geq \beta_c$, define $\nu^1(\beta) = \varphi(s(\beta))$ and let $\nu^k(\beta)$, $k = 2, \ldots, q$, denote the points in $\mathcal{P}_q$ obtained by interchanging the first and kth coordinates of $\nu^1(\beta)$. Then*

$$\mathcal{E}_\beta = \begin{cases} \{\rho\} & \text{for } 0 < \beta < \beta_c\,, \\ \{\nu^1(\beta), \nu^2(\beta), \ldots, \nu^q(\beta)\} & \text{for } \beta > \beta_c\,, \\ \{\rho, \nu^1(\beta_c), \nu^2(\beta_c), \ldots, \nu^q(\beta_c)\} & \text{for } \beta = \beta_c\,. \end{cases}$$

For $\beta \geq \beta_c$, the points in $\mathcal{E}_\beta$ are all distinct, and $\nu^k(\beta)$ is a continuous function of $\beta \geq \beta_c$.

This theorem is proved in Ellis and Wang (1990) by replacing $I_\rho(\gamma) - \frac{1}{2}\beta\langle\gamma,\gamma\rangle$ by another function that has the same global minimum points but for which the analysis is much more straightforward. The two functions are related by Legendre–Fenchel transforms. Probabilistic limit theorems for the Curie–Weiss–Potts model are proved in Ellis and Wang (1990) and Ellis and Wang (1992). Having completed our discussion of the Curie–Weiss–Potts model, we next consider a third mean-field model that exhibits different features.

6.6.3 Mean-field Blume–Capel model

We end this section by considering a mean-field version of an important spin model due to Blume and Capel (Blume 1966; Capel 1966, 1967*a*,*b*). This mean-field model is one of the simplest models that exhibits the following intricate phase-transition structure: a curve of second-order points; a curve of first-order points; and a tricritical point, which separates the two curves. A generalization of the Blume–Capel model is studied in Blume *et al.* (1971).

The mean-field Blume–Capel model is defined on the complete graph on n vertices $1, 2, \ldots, n$. The spin at site $j \in \{1, 2, \ldots, n\}$ is denoted by ω_j, a quantity taking values in $\Lambda = \{-1, 0, 1\}$. The configuration space for the model is the set $\Omega_n = \Lambda^n$, containing all microstates $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ with each $\omega_j \in \Lambda$. In terms of a positive parameter K representing the interaction strength, the Hamiltonian is defined by

$$H_{n,K}(\omega) = \sum_{j=1}^n \omega_j^2 - \frac{K}{n}\left(\sum_{j=1}^n \omega_j\right)^2$$

for each $\omega \in \Omega_n$. Let P_n be the product measure on Λ^n with identical one-dimensional marginals $\rho = \frac{1}{3}(\delta_{-1} + \delta_0 + \delta_1)$. Thus P_n assigns the probability 3^{-n} to each $\omega \in \Omega_n$. The probability of ω corresponding to an inverse temperature $\beta > 0$ and an interaction strength $K > 0$ is defined by the canonical ensemble

$$P_{n,\beta,K}\{\omega\} = \frac{1}{Z_n(\beta, K)} \cdot \exp[-\beta H_{n,K}(\omega)]\, P_n\{\omega\}\,,$$

where $Z_n(\beta, K)$ is the partition function

$$Z_n(\beta, K) = \int_{\Omega_n} \exp[-\beta H_{n,K}(\omega)]\, P_n(d\omega) = \sum_{\omega \in \Omega_n} \exp[-\beta H_{n,K}(\omega)]\, \frac{1}{3^n}\,.$$

The large deviation analysis of the canonical ensemble $P_{n,\beta,K}$ is facilitated by absorbing the noninteracting component of the Hamiltonian into the product measure P_n, and we obtain

$$P_{n,\beta,K}\{\omega\} = \frac{1}{\widetilde{Z}_n(\beta, K)} \cdot \exp\left[n\beta K\left(\frac{S_n(\omega)}{n}\right)^2\right] (P_\beta)_n\{\omega\}\,. \tag{6.19}$$

In this formula, $S_n(\omega)$ equals the total spin $\sum_{j=1}^n \omega_j$; $(P_\beta)_n$ is the product measure on Ω_n with identical one-dimensional marginals

$$\rho_\beta\{\omega_j\} = \frac{1}{Z(\beta)} \cdot \exp(-\beta\omega_j^2)\,\rho\{\omega_j\}\,; \tag{6.20}$$

$Z(\beta)$ is the normalization, equal to $\int_\Lambda \exp(-\beta\omega_j^2)\rho(d\omega_j) = (1 + 2e^{-\beta})/3$; and

$$\widetilde{Z}_n(\beta, K) = \frac{[Z(\beta)]^n}{Z_n(\beta, K)} = \int_{\Omega_n} \exp\left[n\beta K\left(\frac{S_n(\omega)}{n}\right)^2\right] (P_\beta)_n\{\omega\}\,.$$

Comparing eqn (6.19) with (6.15), we see that the mean-field Blume–Capel model has the form of a Curie–Weiss model in which β and the product measure P_n in the latter are replaced by $2\beta K$ and the β-dependent product measure $(P_\beta)_n$ in the former. For each n, S_n/n takes values in $[-1, 1]$. Hence the large deviation principle for S_n/n with respect to the canonical ensemble $P_{n,\beta,K}$ for the mean-field Blume–Capel model is proved exactly like the analogous large deviation principle for the Curie–Weiss model given in Theorem 6.6.1. By Cramér's Theorem (Theorem 6.5.4), with respect to the product measures $(P_\beta)_n$, S_n/n satisfies the large deviation principle with rate function

$$J_\beta(x) = \sup_{t\in\mathbb{R}}\{tx - c_\beta(t)\}\,, \tag{6.21}$$

where $c_\beta(t)$ is the cumulant-generating function

$$c_\beta(t) = \log \int_\Lambda \exp(t\omega_1)\, d\rho_\beta(d\omega_1) = \log\left(\frac{1 + e^{-\beta}(e^t + e^{-t})}{1 + 2e^{-\beta}}\right).$$

The following large deviation principle for S_n/n with respect to $P_{n,\beta,K}$ is an immediate consequence of Theorem 6.5.9 with $\mathcal{X} = [-1, 1]$ and $\psi(x) = \beta K x^2$ for $x \in [-1, 1]$. In this context, the sequence S_n/n is called the sequence of macroscopic variables for the mean-field Blume–Capel model.

Theorem 6.6.5. *With respect to the canonical ensemble $P_{n,\beta,K}$ defined in eqn (6.19), the spin per site S_n/n satisfies the large deviation principle on $[-1, 1]$ with rate function*

$$I_{\beta,K}(x) = J_\beta(x) - \beta K x^2 - \inf_{y\in[-1,1]}\{J_\beta(y) - \beta K y^2\}\,,$$

where $J_\beta(x)$ is defined in eqn (6.21).

As in the Curie–Weiss model and the Curie–Weiss–Potts model, we define the set $\mathcal{E}_{\beta,K}$ of canonical equilibrium macrostates for the spin per site S_n/n in the mean-field Blume–Capel model to be the zero set of the rate function $I_{\beta,K}$ or, equivalently, the set of $x \in [-1,1]$ at which $J_\beta(x) - \beta K x^2$ attains its minimum. Thus

$$\begin{aligned}\mathcal{E}_{\beta,K} &= \{x \in [-1,1] : I_{\beta,K}(x) = 0\} \\ &= \left\{x \in [-1,1] : J_\beta(x) - \beta K x^2 \text{ is minimized}\right\}.\end{aligned}$$

In the case of the Curie–Weiss model, the rate function I in Cramér's Theorem for S_n/n is defined explicitly in eqn (6.13). This explicit formula greatly facilitates the analysis of the structure of the equilibrium macrostates for the spin per site in that model. By contrast, the analogous rate function J_β in the mean-field Blume–Capel model is not given explicitly. In order to determine the structure of $\mathcal{E}_{\beta,K}$, we use the theory of Legendre–Fenchel transforms to replace $J_\beta(x) - \beta K x^2$ by another function that has the same global minimum points but for which the analysis is much more straightforward.

The critical inverse temperature for the mean-field Blume–Capel model is $\beta_c = \log 4$. The structure of $\mathcal{E}_{\beta,K}$ is given first for $0 < \beta \le \beta_c$ and second for $\beta > \beta_c$. The first theorem, proved in Theorem 3.6 of Ellis *et al.* (2005), describes the continuous bifurcation in $\mathcal{E}_{\beta,K}$ as K increases through a value $K(\beta)$. This bifurcation corresponds to a second-order phase transition.

Theorem 6.6.6. *For $0 < \beta \le \beta_c$, we define*

$$K(\beta) = \frac{1}{2\beta c''_\beta(0)} = \frac{e^\beta + 2}{4\beta}. \tag{6.22}$$

For these values of β, $\mathcal{E}_{\beta,K}$ has the following structure.

(a) *For $0 < K \le K(\beta)$, $\mathcal{E}_{\beta,K} = \{0\}$.*
(b) *For $K > K(\beta)$, there exists $m(\beta,K) > 0$ such that $\mathcal{E}_{\beta,K} = \{\pm m(\beta,K)\}$.*
(c) *$m(\beta,K)$ is a positive, increasing, continuous function for $K > K(\beta)$, and as $K \to (K(\beta))^+$, $m(\beta,K) \to 0^+$. Therefore, $\mathcal{E}_{\beta,K}$ exhibits a continuous bifurcation at $K(\beta)$.*

The next theorem, proved in Theorem 3.8 of Ellis *et al.* (2005), describes the discontinuous bifurcation in $\mathcal{E}_{\beta,K}$ for $\beta > \beta_c$ as K increases through a value $K_1(\beta)$. This bifurcation corresponds to a first-order phase transition.

Theorem 6.6.7. *For $\beta > \beta_c$, $\mathcal{E}_{\beta,K}$ has the following structure in terms of the quantity $K_1(\beta)$, denoted by $K_c^{(1)}(\beta)$ in Ellis* et al. *(2005) and defined implicitly for $\beta > \beta_c$ in Ellis* et al. *(2005, p. 2231).*

(a) *For $0 < K < K_1(\beta)$, $\mathcal{E}_{\beta,K} = \{0\}$.*
(b) *For $K = K_1(\beta)$, there exists $m(\beta, K_1(\beta)) > 0$ such that $\mathcal{E}_{\beta,K_1(\beta)} = \{0, \pm m(\beta, K_1(\beta))\}$.*
(c) *For $K > K_1(\beta)$, there exists $m(\beta,K) > 0$ such that $\mathcal{E}_{\beta,K} = \{\pm m(\beta,K)\}$.*

(d) *$m(\beta, K)$ is a positive, increasing, continuous function for $K \geq K_1(\beta)$, and as $K \to K_1(\beta)^+$, $m(\beta, K) \to m(\beta, K_1(\beta)) > 0$. Therefore, $\mathcal{E}_{\beta,K}$ exhibits a discontinuous bifurcation at $K_1(\beta)$.*

Because of the nature of the phase transitions expressed in these two theorems, we refer to the curve $\{(\beta, K(\beta)), 0 < \beta < \beta_c\}$ as the second-order curve and to the curve $\{(\beta, K_1(\beta)), \beta > \beta_c\}$ as the first-order curve. The point $(\beta_c, K(\beta_c)) = (\log 4, 3/2 \log 4)$ separates the second-order curve from the first-order curve and is called the tricritical point.

This completes our analysis of the mean-field Blume–Capel model. Probabilistic limit theorems for this model are proved in Costeniuc *et al.* (2007) and Ellis *et al.* (2005, 2009). In the next section, we discuss the equivalence and nonequivalence of ensembles for a general class of models in statistical mechanics. This discussion is based on a large deviation analysis that was inspired by the work in the present section.

6.7 Equivalence and nonequivalence of ensembles for a general class of models in statistical mechanics

Equilibrium statistical mechanics specifies two ensembles that describe the probability distribution of microstates in statistical-mechanical models. These are the microcanonical ensemble and the canonical ensemble. Particularly in the case of models of coherent structures in turbulence, the microcanonical ensemble is physically more fundamental because it expresses the fact that the Hamiltonian is a constant of the Euler dynamics underlying the model.

The introduction of two separate ensembles raises the basic problem of ensemble equivalence. As we will see in this section, the theory of large deviations and the theory of convex functions provide the perfect tools for analyzing this problem, which forces us to reevaluate a number of deep questions that have often been dismissed in the past as being physically obvious. These questions include the following. Is the temperature of a statistical-mechanical system always related to its energy in a one-to-one fashion? Are the microcanonical equilibrium properties of a system calculated as a function of the energy always equivalent to its canonical equilibrium properties calculated as a function of the temperature? Is the microcanonical entropy always a concave function of the energy? Is the heat capacity always a positive quantity? Surprisingly, the answer to each of these questions is in general no.

Starting with the work of Lynden-Bell and Wood (1968) and of Thirring (1970), physicists have come to realize in recent decades that systematic incompatibilities between the microcanonical and canonical ensembles can arise in the thermodynamic limit if the microcanonical entropy function of the system under study is nonconcave. The reason for this nonequivalence can be explained mathematically by the fact that, when applied to a nonconcave function, the Legendre–Fenchel transform is noninvolutive, i.e. performing it twice does not give back the original function but gives back its concave envelope (Ellis *et al.* 2005; Touchette *et al.* 2004). As a consequence of this property, the Legendre–Fenchel structure of statistical mechanics, traditionally used to establish a one-to-one relationship between the entropy and the free energy and between the energy and the temperature, ceases to be valid when the entropy is nonconcave.

From a more physical perspective, the explanation is even simpler. When the entropy is nonconcave, the microcanonical and canonical ensembles are nonequivalent because the nonconcavity of the entropy implies the existence of a nondifferentiable point of the free energy, and this, in turn, marks the presence of a first-order phase transition in the canonical ensemble (Ellis *et al.* 2000; Gross 1997). Accordingly, the ensembles are nonequivalent because the canonical ensemble jumps over a range of energy values at a critical value of the temperature and is therefore prevented from entering a subset of energy values that can always be accessed by the microcanonical ensemble (Ellis *et al.* 2000; Gross 1997; Thirring 1970). This phenomenon lies at the root of ensemble nonequivalence, which is observed in systems as diverse as the following. It is the typical behavior of systems, such as these, that are defined in terms of long-range interactions.

- Lattice spin models, including the Curie–Weiss–Potts model (Costeniuc *et al.* 2005*a*, 2006*a*), the mean-field Blume–Capel model (Barré *et al.* 2001, 2002; Ellis *et al.* 2004*b*, 2005), mean-field versions of the Hamiltonian model (Dauxois *et al.* 2002; Latora *et al.* 2001), and the XY model (Dauxois *et al.* 2000).
- Gravitational systems (Gross 1997; Hertel and Thirring 1971; Lynden-Bell and Wood 1968; Thirring 1970).
- Models of coherent structures in turbulence (Caglioti *et al.* 1992; Ellis *et al.* 2000, 2002; Eyink and Spohn 1993; Kiessling and Lebowitz 1997; Robert and Sommeria 1991).
- Models of plasmas (Kiessling and Neukirch 2003; Smith and O'Neil 1990).
- Model of the Lennard-Jones gas (Borges and Tsallis 2002).

Many of these models can be analyzed by the methods to be introduced in this section, which summarize the results presented in Ellis *et al.* (2000). Further developments in the theory are given in Costeniuc *et al.* (2005*b*). The reader is referred to these two papers for additional references to the large literature on ensemble equivalence for classical lattice systems and other models.

In the examples just cited, as well as in other cases, the microcanonical formulation gives rise to a richer set of equilibrium macrostates than does the canonical formulation, a phenomenon that occurs especially in the negative-temperature regimes of vorticity dynamics models (DiBattista *et al.* 1998, 2000; Eyink and Spohn 1993; Kiessling and Lebowitz 1997). For example, it has been shown computationally that the strongly reversing zonal-jet structures on Jupiter, as well as the Great Red Spot, fall into the nonequivalent range of the microcanonical ensemble with respect to the energy and circulation invariants (Turkington *et al.* 2001).

6.7.1 Large deviation analysis

The general class of models to be considered includes both spin models and models of coherent structures in turbulence. In order to simplify the presentation, we focus on spin models only. For models of coherent structures in turbulence, the definitions and theorems take slightly different forms (Ellis 2008, Section 10.1). The models to be considered are defined in terms of the following quantities.

- A sequence of probability spaces $(\Omega_n, \mathcal{F}_n, P_n)$ indexed by $n \in I\!N$, which typically represents a sequence of finite-dimensional systems. The Ω_n are the configuration spaces, $\omega \in \Omega_n$ are the microstates, and the P_n are the prior measures.
- For each $n \in I\!N$, the Hamiltonian H_n, a bounded, measurable function mapping Ω_n into $\mathbb{R}$.
- A sequence of positive scaling constants $a_n \to \infty$ as $n \to \infty$. In general, a_n equals the total number of degrees of freedom in the model. In many cases a_n equals the number of particles.

Models of coherent structures in turbulence often incorporate other dynamical invariants besides the Hamiltonian. In this case one replaces H_n in the second bullet point above by the vector of dynamical invariants and makes other corresponding changes in the theory, which are all purely notational. For simplicity, we work only with the Hamiltonian in this section.

A large deviation analysis of the general model is possible provided that there exist, as specified in the next four items, a space of macrostates, a sequence of macroscopic variables, and an interaction representation function, and provided that the macroscopic variables satisfy the large deviation principle on the space of macrostates. One can easily verify that this general setup applies to the three mean-field models considered in Section 6.6.

1. *Space of macrostates.* This is a complete, separable metric space $\mathcal{X}$, which represents the set of all possible macrostates.
2. *Macroscopic variables.* These are a sequence of random variables Y_n mapping Ω_n into $\mathcal{X}$. These functions associate a macrostate in $\mathcal{X}$ with each microstate $\omega \in \Omega_n$.
3. *Hamiltonian representation function.* This is a bounded, continuous function $\tilde{H}$ that maps $\mathcal{X}$ into $\mathbb{R}$ and enables us to write H_n, either exactly or asymptotically, as a function of the macrostate via the macroscopic variable Y_n. Namely, as $n \to \infty$,

$$H_n(\omega) = a_n \tilde{H}(Y_n(\omega)) + \mathrm{o}(a_n) \quad \text{uniformly for } \omega \in \Omega_n \,,$$

i.e.

$$\lim_{n\to\infty} \sup_{\omega\in\Omega_n} |H_n(\omega)/a_n - \tilde{H}(Y_n(\omega))| = 0 \,. \tag{6.23}$$

4. *Large deviation principle for the macroscopic variables.* There exists a function I mapping $\mathcal{X}$ into $[0,\infty]$ and having compact level sets such that, with respect to P_n, the sequence Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I and scaling constants a_n. In other words, for any closed subset F of $\mathcal{X}$,

$$\limsup_{n\to\infty} \frac{1}{a_n} \log P_n\{Y_n \in F\} \le - \inf_{x\in F} I(x) \,,$$

and for any open subset G of $\mathcal{X}$,

$$\liminf_{n\to\infty} \frac{1}{a_n} \log P_n\{Y_n \in G\} \ge - \inf_{x\in G} I(x) \,.$$

Here is a partial list of statistical-mechanical models to which the large deviation formalism has been applied. Further details are given in Costeniuc *et al.* (2005*b*, Example 2.1).

- The Miller–Robert model of fluid turbulence based on the two-dimensional Euler equations (Boucher *et al.* 2000).
- A model of geophysical flows based on equations describing barotropic, quasi-geostrophic turbulence (Ellis *et al.* 2002).
- A model of soliton turbulence based on a class of generalized nonlinear Schrödinger equations (Ellis *et al.* 2004*a*).
- Lattice spin models, including the Curie–Weiss model (Ellis 2006, Section IV.4), the Curie–Weiss–Potts model (Costeniuc *et al.* 2005*a*), the mean-field Blume–Capel model (Ellis *et al.* 2005), and the Ising model (Föllmer and Orey 1987; Olla 1988). The large deviation analysis of these models illustrates the three levels of the Donsker–Varadhan theory of large deviations, which are explained in Ellis (2006, Chapter 1).
  - *Level 1.* As we saw in Subsection 6.6.1, for the Curie–Weiss model the macroscopic variables are the sample means of i.i.d. random variables, and the large deviation principle with respect to the prior measures is the version of Cramér's Theorem given in Corollary 6.5.5. Similar comments apply to the mean-field Blume–Capel model considered in Subsection 6.6.3.
  - *Level 2.* As we saw in Subsection 6.6.2, for the Curie–Weiss–Potts model (Costeniuc *et al.* 2005*a*) the macroscopic variables are the empirical vectors of i.i.d. random variables, and the large deviation principle with respect to the prior measures is the version of Sanov's Theorem given in Theorem 6.3.3.
  - *Level 3.* For the Ising model, the macroscopic variables are an infinite-dimensional generalization of the empirical measure known as the empirical field, and the large deviation principle with respect to the prior measures is derived in Föllmer and Orey (1987) and Olla (1988). This is related to level 3 of the Donsker–Varadhan theory, which is formulated for a general class of Markov chains and Markov processes (Donsker and Varadhan 1983). A special case is treated in Ellis (2006, Chapter IX), which proves the large deviation principle for the empirical process of i.i.d. random variables taking values in a finite state space. The complicated large deviation analysis of the Ising model is outlined in Ellis (1995, Section 11).

Returning now to the general theory, we introduce the microcanonical ensemble, the canonical ensemble, and the basic thermodynamic functions associated with each ensemble: the microcanonical entropy and the canonical free energy. We then sketch the proofs of the large deviation principles for the macroscopic variables Y_n with respect to the two ensembles. As in the case of the Curie–Weiss model, the zeros of the corresponding rate functions define the corresponding sets of equilibrium macrostates, one for the microcanonical ensemble and one for the canonical ensemble. The problem of ensemble equivalence investigates the relationship between these two sets of equilibrium macrostates.

In general terms, the main result is that a necessary and sufficient condition for equivalence of ensembles to hold at the level of equilibrium macrostates is that it holds at the level of thermodynamic functions, which is the case if and only if the microcanonical entropy is concave. The necessity of this condition has the following striking formulation. If the microcanonical entropy is not concave at some value of its argument, then the ensembles are nonequivalent in the sense that the corresponding set of microcanonical equilibrium macrostates is disjoint from any set of canonical equilibrium macrostates. The reader is referred to Ellis *et al.* (2000, Section 1.4) for a detailed discussion of models of coherent structures in turbulence in which nonconcave microcanonical entropies arise.

We start by introducing the function whose support and concavity properties completely determine all aspects of ensemble equivalence and nonequivalence. This function is the microcanonical entropy, defined for $u \in \mathbb{R}$ by

$$s(u) = -\inf\{I(x) : x \in \mathcal{X}, \tilde{H}(x) = u\} . \tag{6.24}$$

Since I maps $\mathcal{X}$ into $[0, \infty]$, s maps $\mathbb{R}$ into $[-\infty, 0]$. Moreover, since I is lower semicontinuous and $\tilde{H}$ is continuous on $\mathcal{X}$, s is upper semicontinuous on $\mathbb{R}$. We define $\operatorname{dom} s$ to be the set of $u \in \mathbb{R}$ for which $s(u) > -\infty$. In general, $\operatorname{dom} s$ is nonempty, since $-s$ is a rate function (Ellis *et al.* 2000, Proposition 3.1(a)). For each $u \in \operatorname{dom} s$, $r > 0$, $n \in I\!N$, and $A \in \mathcal{F}_n$, the microcanonical ensemble is defined to be the conditioned measure

$$P_n^{u,r}\{A\} = P_n\{A \mid H_n/a_n \in [u - r, u + r]\} . \tag{6.25}$$

As shown in Ellis *et al.* (2000, p. 1027), if $u \in \operatorname{dom} s$, then for all sufficiently large n the conditioned measures $P_n^{u,r}$ are well defined.

A mathematically more tractable probability measure is the canonical ensemble. For each $n \in I\!N$, $\beta \in \mathbb{R}$, and $A \in \mathcal{F}_n$, we define the partition function

$$Z_n(\beta) = \int_{\Omega_n} \exp[-\beta H_n] \, dP_n ,$$

which is well defined and finite; the canonical free energy

$$\varphi(\beta) = -\lim_{n \to \infty} \frac{1}{a_n} \log Z_n(\beta) ;$$

and the probability measure

$$P_{n,\beta}\{A\} = \frac{1}{Z_n(\beta)} \cdot \int_A \exp[-\beta H_n] \, dP_n . \tag{6.26}$$

The measures $P_{n,\beta}$ are Gibbs states that define the canonical ensemble for the given model. Although for spin models one usually takes $\beta > 0$, in general $\beta \in \mathbb{R}$ is allowed; for example, negative values of β arise naturally in the study of coherent structures in two-dimensional turbulence.

Among other reasons, the canonical ensemble was introduced by Gibbs in the hope that in the limit $n \to \infty$ the two ensembles would be equivalent, i.e. all macroscopic properties of the model obtained via the microcanonical ensemble could be realized as macroscopic properties obtained via the canonical ensemble. However, as we will see, this is not in general the case.

The large deviation analysis of the canonical ensemble is summarized in the next theorem, Theorem 6.7.1. Part (a) of Theorem 6.7.1 shows that the limit defining $\varphi(\beta)$ exists and is given by a variational formula. Part (b) states the large deviation principle for the macroscopic variables with respect to the canonical ensemble. Part (b) is the analogue of Theorem 6.6.1 for the Curie–Weiss model. In part (c), we consider the set $\mathcal{E}_\beta$ consisting of points at which the rate function in part (b) attains its infimum of 0. The second property of $\mathcal{E}_\beta$ given in part (c) justifies calling this the set of canonical equilibrium macrostates. Part (c) is a special case of Theorem 6.5.3.

Theorem 6.7.1. *We assume that there exists a space of macrostates $\mathcal{X}$, macroscopic variables Y_n, and a Hamiltonian representation function $\tilde{H}$ satisfying*

$$\lim_{n\to\infty} \sup_{\omega\in\Omega_n} |H_n(\omega)/a_n - \tilde{H}(Y_n(\omega))| = 0\,, \tag{6.27}$$

where H_n is the Hamiltonian. We also assume that with respect to the prior measures P_n, Y_n satisfies the large deviation principle on $\mathcal{X}$ with some rate function I and scaling constants a_n. For each $\beta \in \mathbb{R}$, the following conclusions hold.

(a) *The canonical free energy $\varphi(\beta) = -\lim_{n\to\infty}(1/a_n)\log Z_n(\beta)$ exists and is given by*

$$\varphi(\beta) = \inf_{x\in\mathcal{X}}\{\beta\tilde{H}(x) + I(x)\}\,.$$

(b) *With respect to the canonical ensemble $P_{n,\beta}$ defined in eqn (6.26), Y_n satisfies the large deviation principle on $\mathcal{X}$ with scaling constants a_n and rate function*

$$I_\beta(x) = I(x) + \beta\tilde{H}(x) - \varphi(\beta)\,.$$

(c) *We define the set of canonical equilibrium macrostates*

$$\mathcal{E}_\beta = \{x \in \mathcal{X} : I_\beta(x) = 0\}\,.$$

Then $\mathcal{E}_\beta$ is a nonempty, compact subset of $\mathcal{X}$. In addition, if A is a Borel subset of $\mathcal{X}$ such that $\overline{A}\cap\mathcal{E}_\beta = \emptyset$, then $I_\beta(\overline{A}) > 0$ and, for some $C < \infty$,

$$P_{n,\beta}\{Y_n \in A\} \le C\exp[-nI_\beta(\overline{A})/2] \to 0 \quad \text{as } n\to\infty\,.$$

Proof.

(a) Once we take into account the error between H_n and $a_n\tilde{H}(Y_n)$ expressed in eqn (6.27), the proof of (a) and (b) follows from the Laplace principle. Here are the details. By eqn (6.27),

$$\begin{aligned}
&\left|\frac{1}{a_n}\log Z_n(\beta) - \frac{1}{a_n}\log\int_{\Omega_n}\exp[-\beta a_n\tilde{H}(Y_n)]\,dP_n\right| \\
&= \left|\frac{1}{a_n}\log\int_{\Omega_n}\exp[-\beta H_n]\,dP_n - \frac{1}{a_n}\log\int_{\Omega_n}\exp[-\beta a_n\tilde{H}(Y_n)]\,dP_n\right| \\
&\le |\beta|\frac{1}{a_n}\sup_{\omega\in\Omega_n}|H_n(\omega) - a_n\tilde{H}(Y_n(\omega))| \to 0 \;\text{ as } n\to\infty\,.
\end{aligned}$$

Since $\tilde{H}$ is a bounded continuous function mapping $\mathcal{X}$ into $\mathbb{R}$, the Laplace principle satisfied by Y_n with respect to P_n yields part (a):

$$\begin{aligned}\varphi(\beta) &= -\lim_{n\to\infty}\frac{1}{a_n}\log Z_n(\beta)\\ &= -\lim_{n\to\infty}\frac{1}{a_n}\log\int_{\Omega_n}\exp[-\beta a_n\tilde{H}(Y_n)]\,dP_n\\ &= -\sup_{x\in\mathcal{X}}\{-\beta\tilde{H}(x)-I(x)\}\\ &= \inf_{x\in\mathcal{X}}\{\beta\tilde{H}(x)+I(x)\}\,.\end{aligned}$$

(b) This is an immediate consequence of Theorem 6.5.9 with $\psi = \tilde{H}$.

(c) This is proved in Theorem 6.5.3. The proof of the second equation in part (c) is based on the large deviation upper bound for Y_n with respect to $P_{n,\beta}$ (part (b) of this theorem). The proof of the theorem is complete. ■

We next present the large deviation analysis of the microcanonical ensemble $P_n^{u,r}$ defined in eqn (6.25). In this context, the microcanonical entropy s defined in eqn (6.24) has the following property: $-s$ is the rate function in the large deviation principles, with respect to the prior measures P_n, for both $\tilde{H}(Y_n)$ and H_n/a_n. In order to see this, we recall that with respect to P_n, Y_n satisfies the large deviation principle with rate function I. Since $\tilde{H}$ is a continuous function mapping $\mathcal{X}$ into $\mathbb{R}$, the large deviation principle for $\tilde{H}(Y_n)$ is a consequence of the contraction principle (Theorem 6.5.8). For $u \in \mathbb{R}$, the rate function is given by

$$\inf\{I(x) : x \in \mathcal{X}, \tilde{H}(x) = u\} = -s(u)\,.$$

In addition, since

$$\lim_{n\to\infty}\sup_{\omega\in\Omega_n}|H_n(\omega)/a_n - \tilde{H}(Y_n(\omega))| = 0\,,$$

H_n/a_n inherits from $\tilde{H}(Y_n)$ the large deviation principle with the same rate function. This follows from Dupuis and Ellis (1997, Theorem 1.3.3) or can be derived as in the proof of part (a) of Theorem 6.7.1 by using the equivalent Laplace principle. We summarize this large deviation principle by the notation

$$P_n\{H_n/a_n \in [u-r, u+r]\} \asymp \exp[a_n s(u)] \quad \text{as } n\to\infty, r\to 0\,. \tag{6.28}$$

For $x \in \mathcal{X}$ and $\alpha > 0$, $B(x,\alpha)$ denotes the open ball with center x and radius α. We next motivate the large deviation principle for Y_n with respect to the microcanonical ensemble $P_n^{u,r}$ by estimating the exponential-order contribution to the probability $P_n^{u,r}\{Y_n \in B(x,\alpha)\}$ as $n\to\infty$. Specifically, we seek a function I^u such that for all $u \in \operatorname{dom} s$, all $x \in \mathcal{X}$, and all $\alpha > 0$ sufficiently small,

$$P_n^{u,r}\{Y_n \in B(x,\alpha)\} \approx \exp[-a_n I^u(x)] \quad \text{as } n\to\infty, r\to 0, \alpha\to 0\,. \tag{6.29}$$

The calculation that we present shows both the interpretative power of the large deviation notation and the value of left-handed thinking.

We first work with $x \in \mathcal{X}$ for which $I(x) < \infty$ and $\tilde{H}(x) = u$. Such an x exists, since $u \in \operatorname{dom} s$ and thus $s(u) > -\infty$. Because

$$\lim_{n\to\infty} \sup_{\omega\in\Omega_n} |H_n(\omega)/a_n - \tilde{H}(Y_n(\omega))| = 0\,,$$

for all sufficiently large n depending on r, the set of ω for which both $Y_n(\omega) \in B(x,\alpha)$ and $H_n(\omega)/a_n \in [u-r, u+r]$ is approximately equal to the set of ω for which both $Y_n(\omega) \in B(x,\alpha)$ and $\tilde{H}(Y_n(\omega)) \in [u-r, u+r]$. Since $\tilde{H}$ is continuous and $\tilde{H}(x) = u$, for all sufficiently small α compared with r this set reduces to $\{\omega : Y_n(\omega) \in B(x,\alpha)\}$. Hence, for all sufficiently small r, all sufficiently large n depending on r, and all sufficiently small α compared with r, the assumed large deviation principle for Y_n with respect to P_n and the large deviation principle for H_n/a_n summarized in eqn (6.28) yield

$$\begin{aligned} P_n^{u,r}\{Y_n \in B(x,\alpha)\} &= \frac{P_n\{\{Y_n \in B(x,\alpha)\} \cap \{H_n/a_n \in [u-r,u+r]\}\}}{P_n\{H_n/a_n \in [u-r,u+r]\}} \\ &\approx \frac{P_n\{Y_n \in B(x,\alpha)\}}{P_n\{H_n/a_n \in [u-r,u+r]\}} \\ &\approx \exp[-a_n(I(x)+s(u))]\,. \end{aligned}$$

On the other hand, if $\tilde{H}(x) \neq u$, then a similar calculation shows that for all sufficiently small r, all sufficiently small α, and all sufficiently large n, $P_n^{u,r}\{Y_n \in B(x,\alpha)\} = 0$. Comparing these approximate calculations with the desired asymptotic form (6.29) motivates the correct formula for the rate function (Ellis *et al.* 2000, Theorem 3.2):

$$I^u(x) = \begin{cases} I(x) + s(u) & \text{if } \tilde{H}(x) = u\,, \\ \infty & \text{if } \tilde{H}(x) \neq u\,. \end{cases} \tag{6.30}$$

We record the facts in the next theorem, which is proved in Ellis *et al.* (2000, Section 3). An additional complication occurs in part (b) because the large deviation principle involves the double limit $n \to 0$ followed by $r \to 0$. In part (c), we introduce the set of microcanonical equilibrium macrostates $\mathcal{E}^u$ and state a property of this set with respect to the microcanonical ensemble that is analogous to the property satisfied by the set $\mathcal{E}_\beta$ of canonical equilibrium macrostates with respect to the canonical ensemble. The proof, given in Ellis *et al.* (2000, Theorem 3.5), is similar to the proof of the analogous property of $\mathcal{E}_\beta$ given in part (c) of Theorem 6.7.1 and is omitted.

Theorem 6.7.2. *We assume that there exists a space of macrostates $\mathcal{X}$, macroscopic variables Y_n, and a Hamiltonian representation function $\tilde{H}$ satisfying eqn (6.23). We also assume that with respect to the prior measures P_n, Y_n satisfies the large deviation principle on $\mathcal{X}$ with scaling constants a_n and some rate function I. For each $u \in \operatorname{dom} s$ and any $r \in (0,1)$, the following conclusions hold.*

(a) *With respect to P_n, $\tilde{H}(Y_n)$ and H_n/a_n both satisfy the large deviation principle with scaling constants a_n and rate function $-s$.*

(b) *We consider the microcanonical ensemble $P_n^{u,r}$ defined in eqn (6.25). With respect to $P_n^{u,r}$ and in the double limit $n \to \infty$ and $r \to 0$, Y_n satisfies the large deviation* *principle on $\mathcal{X}$ with scaling constants a_n and the rate function I^u defined in eqn (6.30). That is, for any closed subset F of $\mathcal{X}$,*

$$\lim_{r\to 0} \limsup_{n\to\infty} \frac{1}{a_n} \log P_n^{u,r}\{Y_n \in F\} \leq -I^u(F)$$

and, for any open subset G of $\mathcal{X}$,

$$\lim_{r\to 0} \liminf_{n\to\infty} \frac{1}{a_n} \log P_n^{u,r}\{Y_n \in G\} \geq -I^u(G)\,.$$

(c) *We define the set of equilibrium macrostates*

$$\mathcal{E}^u = \{x \in \mathcal{X} : I^u(x) = 0\}\,.$$

Then $\mathcal{E}^u$ is a nonempty, compact subset of $\mathcal{X}$. In addition, if A is a Borel subset of $\mathcal{X}$ such that $\overline{A} \cap \mathcal{E}^u = \emptyset$, then $I^u(\overline{A}) > 0$ and there exists $r_0 > 0$ and, for all $r \in (0, r_0]$, there exists $C_r < \infty$ such that

$$P_{n,\beta}\{Y_n \in A\} \leq C_r \exp[-n\, I_\beta(\overline{A})/2] \to 0 \quad as\ n \to \infty\,.$$

This completes the large deviation analysis of the general spin model. In the next subsection, we investigate the equivalence and nonequivalence of the canonical and microcanonical ensembles.

6.7.2 Equivalence and nonequivalence of ensembles

The study of the equivalence and nonequivalence of the canonical and microcanonical ensembles involves the relationships between the two sets of equilibrium macrostates

$$\mathcal{E}_\beta = \{x \in \mathcal{X} : I_\beta(x) = 0\} \quad \text{and} \quad \mathcal{E}^u = \{x \in \mathcal{X} : I^u(x) = 0\}\,.$$

The following questions will be considered.

1. Given $\beta \in \mathbb{R}$ and $x \in \mathcal{E}_\beta$, does there exists $u \in \mathbb{R}$ such that $x \in \mathcal{E}^u$? In other words, is any canonical equilibrium macrostate realized microcanonically?
2. Given $u \in \mathbb{R}$ and $x \in \mathcal{E}^u$, does there exist $\beta \in \mathbb{R}$ such that $x \in \mathcal{E}_\beta$? In other words, is any microcanonical equilibrium macrostate realized canonically?

As we will see in Theorem 6.7.4, the answer to question 1 is always yes, but the answer to question 2 is much more complicated, involving three possibilities.

2a. *Full equivalence.* There exists $\beta \in \mathbb{R}$ such that $\mathcal{E}^u = \mathcal{E}_\beta$.
2b. *Partial equivalence.* There exists $\beta \in \mathbb{R}$ such that $\mathcal{E}^u \subset \mathcal{E}_\beta$ but $\mathcal{E}^u \neq \mathcal{E}_\beta$.
2c. *Nonequivalence.* $\mathcal{E}^u$ is disjoint from $\mathcal{E}_\beta$ for all $\beta \in \mathbb{R}$.

One of the big surprises of the theory to be presented here is that we are able to decide on which of these three possibilities occurs by examining the support and concavity properties of the microcanonical entropy $s(u)$. This is remarkable because the sets $\mathcal{E}_\beta$ and $\mathcal{E}^u$ are, in general, infinite-dimensional while the microcanonical entropy is a function on $\mathbb{R}$.

In order to begin our study of ensemble equivalence and nonequivalence, we first recall the definitions of the corresponding rate functions:

$$I_\beta(x) = I(x) + \beta\tilde{H}(x) - \varphi(\beta)\,,$$

where $\varphi(\beta)$ denotes the canonical free energy

$$\varphi(\beta) = \inf_{x\in\mathcal{X}}\{\beta\tilde{H}(x) + I(x)\}\,,$$

and

$$I^u(x) = \begin{cases} I(x) + s(u) & \text{if } \tilde{H}(x) = u\,, \\ \infty & \text{if } \tilde{H}(x) \neq u\,, \end{cases}$$

where $s(u)$ denotes the microcanonical entropy

$$s(u) = -\inf\{I(x) : x \in \mathcal{X}, \tilde{H}(x) = u\}\,.$$

Using these definitions, we see that the two sets of equilibrium macrostates have the alternate characterizations

$$\mathcal{E}_\beta = \{x \in \mathcal{X} : I(x) + \beta\tilde{H}(x) \text{ is minimized}\}$$

and

$$\mathcal{E}^u = \{x \in \mathcal{X} : I(x) \text{ is minimized subject to } \tilde{H}(x) = u\}\,.$$

Thus $\mathcal{E}^u$ is defined by the following constrained minimization problem for $u \in \mathbb{R}$:

$$\text{minimize } I(x) \text{ over } \mathcal{X} \text{ subject to the constraint } \tilde{H}(x) = u\,. \tag{6.31}$$

By contrast, $\mathcal{E}_\beta$ is defined by the following related, unconstrained minimization problem for $\beta \in \mathbb{R}$:

$$\text{minimize } I(x) + \beta\tilde{H}(x) \text{ over } x \in \mathcal{X}\,. \tag{6.32}$$

In this formulation, β is a Lagrange multiplier dual to the constraint $\tilde{H}(x) = u$. The theory of Lagrange multipliers outlines suitable conditions under which the solutions of the constrained problem (6.31) lie among the critical points of $I + \beta\tilde{H}$. However, it does not give, as we will do in Theorem 6.7.4, necessary and sufficient conditions for the solutions of eqn (6.31) to coincide with the solutions of the unconstrained minimization problem (6.32). These necessary and sufficient conditions are expressed in terms of support and concavity properties of the microcanonical entropy $s(u)$.

Before we explain this, we point out a basic property relating the two thermodynamic functions $s(u)$ and $\varphi(\beta)$.

Theorem 6.7.3. *The microcanonical entropy $s(u)$ defined in eqn (6.24) and the canonical free energy $\varphi(\beta)$ defined in part (a) of Theorem 6.7.1 are related via the Legendre–Fenchel transform*

$$\varphi(\beta) = \inf_{u\in\mathbb{R}}\{\beta u - s(u)\}\,. \tag{6.33}$$

Proof. By the variational formula given in part (a) of Theorem 6.7.1,

$$\begin{aligned}\varphi(\beta) &= \inf_{x\in\mathcal{X}}\{\beta\tilde{H}(x)+I(x)\}\\ &= \inf_{u\in\mathbb{R}}\inf\{\beta\tilde{H}(x)+I(x) : x\in\mathcal{X}, \tilde{H}(x)=u\}\\ &= \inf_{u\in\mathbb{R}}\{\beta u+\inf\{I(x) : x\in\mathcal{X}, \tilde{H}(x)=u\}\\ &= \inf_{u\in\mathbb{R}}\{\beta u - s(u)\}\,.\end{aligned}$$

This completes the proof. ■

The formula (6.33), which exhibits $\varphi(\beta)$ as a concave function on $\mathbb{R}$, is related to a fundamental lack of symmetry between the microcanonical and canonical ensembles. Although one can obtain $\varphi(\beta)$ from $s(u)$ via a Legendre–Fenchel transform, in general one cannot obtain $s(u)$ from $\varphi(\beta)$ via the dual formula $s(u) = \inf_{\beta\in\mathbb{R}}\{\beta u - \varphi(\beta)\}$ unless s is concave on $\mathbb{R}$, which in general is not the case. In fact, the concavity of s on $\mathbb{R}$ depends on properties of I and $\tilde{H}$. For example, if I is convex on $\mathcal{X}$ and $\tilde{H}$ is affine, then s is concave on $\mathbb{R}$. On the other hand, microcanonical entropies s that are not concave on $\mathbb{R}$ arise in many models involving long-range interactions, including those listed in the five bullet points at the beginning of Section 6.7. This discussion indicates that of the two thermodynamic functions, the microcanonical entropy is the more fundamental, a state of affairs that is reinforced by the results on ensemble equivalence and nonequivalence to be presented in Theorem 6.7.4.

In order to state this theorem, we need several definitions. A function f on $\mathbb{R}$ is said to be concave on $\mathbb{R}$, or simply "concave," if f maps $\mathbb{R}$ into $\mathbb{R}\cup\{-\infty\}$, $f(u) > -\infty$ for some $u \in \mathbb{R}$, and, for all u and v in $\mathbb{R}$ and all $\lambda \in (0,1)$,

$$f(\lambda u + (1-\lambda)v) \geq \lambda f(u) + (1-\lambda)f(v)\,.$$

Let $f \not\equiv -\infty$ be a function mapping $\mathbb{R}$ into $\mathbb{R}\cup\{-\infty\}$. We define $\operatorname{dom} f$ to be the set of $u \in \mathbb{R}$ for which $f(u) > -\infty$. For β and u in $\mathbb{R}$, the Legendre–Fenchel transforms f^* and f^{**} are defined by (Rockafellar 1970, p. 308)

$$f^*(\beta) = \inf_{u\in\mathbb{R}}\{\beta u - f(u)\} \quad\text{and}\quad f^{**}(u) = (f^*)^*(u) = \inf_{\beta\in\mathbb{R}}\{\beta u - f^*(\beta)\}\,.$$

As in the case of convex functions (Ellis 2006, Theorem VI.5.3), f^* is concave and upper semicontinuous on $\mathbb{R}$, and for all $u \in \mathbb{R}$ we have $f^{**}(u) = f(u)$ if and only if f is concave and upper semicontinuous on $\mathbb{R}$. If f is not concave and upper semicontinuous on $\mathbb{R}$, then f^{**} is the smallest concave, upper semicontinuous function on $\mathbb{R}$ that satisfies $f^{**}(u) \geq f(u)$ for all $u \in \mathbb{R}$ (Costeniuc *et al.* 2005*b*, Proposition A.2). In particular, if for some u, $f(u) \neq f^{**}(u)$, then $f(u) < f^{**}(u)$. We say that f is concave at $u \in \operatorname{dom} f$ if $f(u) = f^{**}(u)$ and that f is nonconcave at $u \in \operatorname{dom} f$ if $f(u) < f^{**}(u)$. These definitions are reasonable because f^{**} is concave on $\mathbb{R}$.

We now state the main theorem concerning the equivalence and nonequivalence of the microcanonical and canonical ensembles. According to part (a), canonical equilibrium macrostates are always realized microcanonically. However, according to parts (b)–(d), the converse in general is false. The three possibilities given in parts (b)–(d) depend on concavity and support properties of the microcanonical entropy. The next theorem is proved in Ellis *et al.* (2000, Section 4), where it is formulated somewhat differently. The formulation given here specializes Theorem 3.1 in Costeniuc *et al.* (2005*b*) to dimension 1.

Theorem 6.7.4. *In parts (b), (c), and (d), u denotes any point in* $\operatorname{dom} s$.

(a) Canonical is always realized microcanonically. *We define $\tilde{H}(\mathcal{E}_\beta)$ to be the set of $u \in \mathbb{R}$ having the form $u = \tilde{H}(x)$ for some $x \in \mathcal{E}_\beta$. Then, for any $\beta \in \mathbb{R}$, we have $\tilde{H}(\mathcal{E}_\beta) \subset \operatorname{dom} s$ and*

$$\mathcal{E}_\beta = \bigcup_{u \in \tilde{H}(\mathcal{E}_\beta)} \mathcal{E}^u \,.$$

(b) Full equivalence. *There exists $\beta \in \mathbb{R}$ such that $\mathcal{E}^u = \mathcal{E}_\beta$ if and only if s has a strictly supporting line at u with tangent β, i.e.*

$$s(v) < s(u) + \beta(v-u) \text{ for all } v \neq u \,.$$

(c) Partial equivalence. *There exists $\beta \in \mathbb{R}$ such that $\mathcal{E}^u \subset \mathcal{E}_\beta$ but $\mathcal{E}^u \neq \mathcal{E}_\beta$ if and only if s has a nonstrictly supporting line at u with tangent β, i.e.*

$$s(v) \leq s(u) + \beta(v-u) \text{ for all } v, \text{ with equality for some } v \neq u \,.$$

(d) Nonequivalence. *For all $\beta \in \mathbb{R}$, $\mathcal{E}^u \cap \mathcal{E}_\beta = \emptyset$ if and only if s has no supporting line at u, i.e.*

$$\text{for all } \beta \in \mathbb{R} \text{ there exists } v \text{ such that } s(v) > s(u) + \beta(v-u) \,.$$

Here are some useful criteria for full or partial equivalence of ensembles and for nonequivalence of ensembles.

- *Full or partial equivalence.* Except possibly for boundary points of $\operatorname{dom} s$, s has a supporting line at $u \in \operatorname{dom} s$ if and only if s is concave at u (Costeniuc *et al.* 2005*b*, Theorem A.5(c)), and thus, according to parts (a) and (b) of Theorem 6.7.4, full or partial equivalence of ensembles holds.
- *Full equivalence.* Assume that $\operatorname{dom} s$ is a nonempty interval and that s is strictly concave on the interior of $\operatorname{dom} s$, i.e. for all $u \neq v$ in the interior of $\operatorname{dom} s$ and all $\lambda \in (0,1)$,
$$s(\lambda u + (1-\lambda)v) > \lambda s(u) + (1-\lambda)s(v) \,.$$
Then, except possibly for boundary points of $\operatorname{dom} s$, s has a strictly supporting line at all $u \in \operatorname{dom} s$, and thus, according to part (a) of the theorem, full equivalence of ensembles holds (Costeniuc *et al.* 2005*b*, Theorem A.4(c)).
- *Nonequivalence.* Except possibly for boundary points of $\operatorname{dom} s$, s has no supporting line at $u \in \operatorname{dom} s$ if and only if s is nonconcave at u (Costeniuc *et al.* 2005*b*, Theorem A.5(c)).

The various possibilities in parts (b), (c), and (d) of Theorem 6.7.4 are illustrated in Ellis *et al.* (2004*b*) for the mean-field Blume–Capel spin model. In Ellis *et al.* (2002), the theory is applied to a model of coherent structures in two-dimensional turbulence. Numerical computations implemented for geostrophic turbulence over topography in a zonal channel demonstrate that nonequivalence of ensembles occurs over a wide range of the model parameters and that physically interesting equilibria seen microcanonically are often omitted by the canonical ensemble. The coherent structures observed in the model resemble the coherent structures observed in the midlatitude, zone-belt domains on Jupiter.

In Costeniuc *et al.* (2005*b*), the theory developed in Ellis *et al.* (2000) and summarized in Theorem 6.7.4 is extended. In Costeniuc *et al.* (2005*b*), it is shown that when the microcanonical ensemble is nonequivalent to the canonical ensemble on a subset of values of the energy, it is often possible to modify the definition of the canonical ensemble so as to recover equivalence with the microcanonical ensemble. Specifically, we give natural conditions under which one can construct a so-called Gaussian ensemble that is equivalent to the microcanonical ensemble when the canonical ensemble is not. This is potentially useful if one wants to work out the equilibrium properties of a system in the microcanonical ensemble, a notoriously difficult problem because of the equality constraint appearing in the definition of this ensemble. An overview of Costeniuc *et al.* (2005*b*) is given in Costeniuc *et al.* (2006*b*), and in Costeniuc *et al.* (2006*a*) it is applied to the Curie–Weiss–Potts model.

This completes our presentation of the theory of large deviations. The seed from which the theory blossomed was Boltzmann's discovery of a statistical interpretation of entropy. After examining his discovery and outlining a number of basic results in large deviation theory, in this paper we have brought the theory back to its roots by applying it to several problems in statistical mechanics. It is hoped that the reader will be inspired by this paper to further investigate the symbiotic and mutually invigorating relationship between these two fields.

Acknowledgment

The research of Richard S. Ellis is supported by a grant from the National Science Foundation (NSF-DMS-0604071).

References

Barré J., D. Mukamel, and S. Ruffo (2001). Inequivalence of ensembles in a system with long-range interactions. *Phys. Rev. Lett.* 87:030601.

Barré J., D. Mukamel, and S. Ruffo (2002). Ensemble inequivalence in mean-field models of magnetism. In T. Dauxois, S. Ruffo, E. Arimondo, and M. Wilkens (editors), *Dynamics and Thermodynamics of Systems with Long-Range Interactions*, Lecture Notes in Physics 602, pp. 45–67. New York: Springer.

Blume M. (1966). Theory of the first-order magnetic phase change in UO_2. *Phys. Rev.* 141:517–524.

Blume M., V. J. Emery, and R. B. Griffiths (1971). Ising model for the λ transition and phase separation in He^3–He^4 mixtures. *Phys. Rev. A* 4:1071–1077.

Boltzmann L. (1877). 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, no. 76, 373–435.

Borges E. P. and C. Tsallis (2002). Negative specific heat in a Lennard-Jones-like gas with long-range interactions. *Physica A* 305:148–151.

Boucher C., R. S. Ellis, and B. Turkington (2000). Derivation of maximum entropy principles in two-dimensional turbulence via large deviations. *J. Stat. Phys.* 98:1235–1278.

Caglioti E., P. L. Lions, C. Marchioro, and M. Pulvirenti (1992). A special class of stationary flows for two-dimensional Euler equations: A statistical mechanical description. *Commun. Math. Phys.* 143:501–525.

Capel H. W. (1966). On the possibility of first-order phase transitions in Ising systems of triplet ions with zero-field splitting. *Physica* 32:966–988.

Capel H. W. (1967*a*). On the possibility of first-order phase transitions in Ising systems of triplet ions with zero-field splitting II. *Physica* 33:295–331.

Capel H. W. (1967*b*). On the possibility of first-order phase transitions in Ising systems of triplet ions with zero-field splitting III. *Physica* 37:423–441.

Costeniuc M., R. S. Ellis, and H. Touchette (2005*a*). Complete analysis of phase transitions and ensemble equivalence for the Curie–Weiss–Potts model. *J. Math. Phys.* 46:063301 (25 pages).

Costeniuc M., R. S. Ellis, H. Touchette, and B. Turkington (2005*b*). The generalized canonical ensemble and its universal equivalence with the microcanonical ensemble. *J. Stat. Phys.* 119:1283–1329.

Costeniuc M., R. S. Ellis, and H. Touchette (2006*a*). Nonconcave entropies from generalized canonical ensembles. *Phys. Rev. E* 74:010105(R) (4 pages).

Costeniuc M., R. S. Ellis, H. Touchette, and B. Turkington (2006*b*). Generalized canonical ensembles and ensemble equivalence. *Phys. Rev. E* 73:026105 (8 pages).

Costeniuc M., R. S. Ellis, and P. T.-H. Otto (2007). Multiple critical behavior of probabilistic limit theorems in the neighborhood of a tricritical point. *J. Stat. Phys.* 127:495–552.

Dauxois T., P. Holdsworth, and S. Ruffo (2000). Violation of ensemble equivalence in the antiferromagnetic mean-field XY model. *Eur. Phys. J. B* 16:659.

Dauxois T., V. Latora, A. Rapisarda, S. Ruffo, and A. Torcini (2002). The Hamiltonian mean field model: From dynamics to statistical mechanics and back. In T. Dauxois, S. Ruffo, E. Arimondo, and M. Wilkens (editors), *Dynamics and Thermodynamics of Systems with Long-Range Interactions*, Lecture Notes in Physics 602, pp. 458–487. New York: Springer.

Dembo A. and O. Zeitouni (1998). *Large Deviations Techniques and Applications*, second edition. New York: Springer.

DiBattista M., A. Majda, and B. Turkington (1998). Prototype geophysical vortex structures via large-scale statistical theory. *Geophys. Astrophys. Fluid Dyn.* 89:235–283.

DiBattista M., A. Majda, and M. Grote (2000). Meta-stability of equilibrium statistical structures for prototype geophysical flows with damping and driving. *Physica D* 151:271–304.

Dobrushin R. L. and S. B. Shlosman (1994). Large and moderate deviations in the Ising Model. *Adv. Soviet Math.* 20:91–219.

Donsker M. D. and S. R. S. Varadhan (1983). Asymptotic evaluation of certain Markov process expectations for large time, IV. *Commun. Pure Appl. Math.* 36:183–212.

Dupuis P. and R. S. Ellis (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. New York: Wiley.

Ellis R. S. (1995). An overview of the theory of large deviations and applications to statistical mechanics. *Scand. Actuarial J.* No. 1, 97–142.

Ellis R. S. (2006). *Entropy, Large Deviations, and Statistical Mechanics*. New York: Springer, 1985. Reprinted in *Classics of Mathematics* series.

Ellis R. S. (2008). The theory of large deviations and applications to statistical mechanics. Lecture notes for École de Physique Les Houches, August 5–8, 2008. Posted at http://www.math.umass.edu/∼rsellis/pdf-files/Les-Houches-lectures.pdf.

Ellis R. S. and C. M. Newman (1978*a*). Limit theorems for sums of dependent random variables occurring in statistical mechanics. *Z. Wahrsch. verw. Geb.* 44:117–139.

Ellis R. S. and C. M. Newman (1978*b*). The statistics of Curie–Weiss models. *J. Stat. Phys.* 19:149–161.

Ellis R. S. and K. Wang (1990). Limit theorems for the empirical vector of the Curie–Weiss–Potts model. *Stoch. Proc. Appl.* 35:59–79.

Ellis R. S. and K. Wang (1992). Limit theorems for maximum likelihood estimators in the Curie–Weiss–Potts model. *Stoch. Proc. Appl.* 40:251–288.

Ellis R. S., C. M. Newman, and J. S. Rosen (1980). Limit theorems for sums of dependent random variables occurring in statistical mechanics, II: Conditioning, multiple phases, and metastability. *Z. Wahrsch. verw. Geb.* 51:153–169.

Ellis R. S., K. Haven, and B. Turkington (2000). Large deviation principles and complete equivalence and nonequivalence results for pure and mixed ensembles. *J. Stat. Phys.* 101:999–1064.

Ellis R. S., K. Haven, and B. Turkington (2002). Nonequivalent statistical equilibrium ensembles and refined stability theorems for most probable flows. *Nonlinearity* 15:239–255.

Ellis R. S., R. Jordan, P. Otto, and B. Turkington (2004*a*). A statistical approach to the asymptotic behavior of a generalized class of nonlinear Schrödinger equations. *Commun. Math. Phys.*, 244:187–208.

Ellis R. S., H. Touchette, and B. Turkington (2004*b*). Thermodynamic versus statistical nonequivalence of ensembles for the mean-field Blume–Emery–Griffiths model. *Physica A* 335:518–538.

Ellis R. S., P. Otto, and H. Touchette (2005). Analysis of phase transitions in the mean-field Blume–Emery–Griffiths model. *Ann. Appl. Probab.* 15:2203–2254.

Ellis R. S., J. Machta, and P. T. Otto (2008). Asymptotic behavior of the magnetization near critical and tricritical points via Ginzburg–Landau polynomials. *J. Stat. Phys.* 133:101–129.

Ellis R. S., J. Machta, and P. T. Otto (2009). Asymptotic behavior of the finite-size magnetization as a function of the approach to criticality. Submitted for publication.

Ethier S. N. and T. G. Kurtz (1986). *Markov Processes: Characterization and Convergence*. New York: Wiley.

Everdell W. R. (1997). *The First Moderns.* Chicago: The University of Chicago Press.

Eyink G. L. and H. Spohn (1993). Negative-temperature states and large-scale, long-lived vortices in two-dimensional turbulence. *J. Stat. Phys.* 70:833–886.

Föllmer H. and S. Orey (1987). Large deviations for the empirical field of a Gibbs measure. *Ann. Probab.* 16:961–977.

Gross D. H. E. (1997). Microcanonical thermodynamics and statistical fragmentation of dissipative systems: The topological structure of the n-body phase space. *Phys. Rep.* 279:119–202.

Hertel P. and W. Thirring (1971). A soluble model for a system with negative specific heat. *Ann. Phys. (NY)* 63:520.

Kiessling M. K.-H. and J. L. Lebowitz (1997). The micro-canonical point vortex ensemble: Beyond equivalence. *Lett. Math. Phys.* 42:43–56.

Kiessling M. K.-H. and T. Neukirch (2003). Negative specific heat of a magnetically self-confined plasma torus. *Proc. Natl. Acad. Sci. USA* 100:1510–1514.

Lanford O. E. (1973). Entropy and equilibrium states in classical statistical mechanics. In A. Lenard (editor), *Statistical Mechanics and Mathematical Problems*, Lecture Notes in Physics 20, pp. 1–113. Berlin: Springer.

Latora V., A. Rapisarda, and C. Tsallis (2001). Non-Gaussian equilibrium in a long-range Hamiltonian system. *Phys. Rev. E* 64:056134.

Lindley D. (2001). *Boltzmann's Atom: The Great Debate That Launched a Revolution in Physics*. New York: Free Press.

Lynden-Bell D. and R. Wood (1968). The gravo-thermal catastrophe in isothermal spheres and the onset of red-giant structure for stellar systems. *Mon. Not. R. Astron. Soc.* 138:495.

Olla S. (1988) Large deviations for Gibbs random fields. *Probab. Theor. Rel. Fields* 77:343–359.

Parisi G. (1988). *Statistical Field Theory.* Redwood City, CA: Addison-Wesley.

Robert R. (1991). A maximum-entropy principle for two-dimensional perfect fluid dynamics. *J. Stat. Phys.* 65:531–553.

Robert R. and J. Sommeria (1991). Statistical equilibrium states for two-dimensional flows. *J. Fluid Mech.* 229:291–310.

Rockafellar R. T. (1970) *Convex Analysis*. Princeton, NJ: Princeton University Press.

Smith R. A. and T. M. O'Neil (1990). Nonaxisymmetric thermal equilibria of a cylindrically bounded guiding center plasma or discrete vortex system. *Phys. Fluids B* 2:2961–2975.

Thirring W. (1970) Systems with negative specific heat. *Z. Phys.* 235:339–352.

Touchette H. (2009). The large deviation approach to statistical mechanics. *Phys. Rep.* 478:1–69.

Touchette H., R. S. Ellis, and B. Turkington (2004). An introduction to the thermodynamic and macrostate levels of nonequivalent ensembles. *Physica A* 340:138–146.

Turkington B., A. Majda, K. Haven, and M. DiBattista (2001). Statistical equilibrium predictions of jets and spots on Jupiter. *Proc. Natl. Acad. Sci. USA* 98:12346–12350.

Varadhan S. R. S. (1966). Asymptotic properties and differential equations. *Commun. Pure Appl. Math.* 19:261–286.

Wightman A. S. (1979). Convexity and the notion of equilibrium state in thermodynamics and statistical mechanics. Introduction to R. B. Israel, *Convexity in the Theory of Lattice Gases.* Princeton: Princeton University Press.

Wu F. Y. (1982). The Potts model. *Rev. Mod. Phys.* 54:235–268.