The Theory of Large Deviations and Applications to Statistical Mechanics

Lecture Notes for École de Physique Les Houches August 5–8, 2008

updated November 19, 2009

Richard S. Ellis
Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA 01003

rsellis@math.umass.edu
http://www.math.umass.edu/~rsellis

Our Lives Are Large Deviations

Statistically, the probability of any one of us being here is so small that you'd think the mere fact of existing would keep us all in a contented dazzlement of surprise. We are alive against the stupendous odds of genetics, infinitely outnumbered by all the alternates who might, except for luck, be in our places.

Even more astounding is our statistical improbability in physical terms. The normal, predictable state of matter throughout the universe is randomness, a relaxed sort of equilibrium, with atoms and their particles scattered around in an amorphous muddle. We, in brilliant contrast, are completely organized structures, squirming with information at every covalent bond. We make our living by catching electrons at the moment of their excitement by solar photons, swiping the energy released at the instant of each jump and storing it up in intricate loops for ourselves. We violate probability, by our nature. To be able to do this systemically, and in such wild varieties of form, from viruses to whales, is extremely unlikely; to have sustained the effort successfully for the several billion years of our existence, without drifting back into randomness, was nearly a mathematical impossibility.

Lewis Thomas, *The Lives of a Cell*
(New York: Viking Press, 1974), p. 141

"Life, by Any Reasonable Measure, Is Impossible"

Art is a way of saying what it means to be alive, and the most salient feature of existence is the unthinkable odds against it. For every way that there is of being here, there are an infinity of ways of not being here. Historical accident snuffs out whole universes with every clock tick. Statistics declare us ridiculous. Thermodynamics prohibits us. Life, by any reasonable measure, is impossible, and my life—this, here, now—infinitely more so. Art is a way of saying, in the face of all that impossibility, just how worth celebrating it is to be able to say anything at all.

Richard Powers, *Conjunctions*,
quoted in John Leonard, "Mind Painting,"
The New York Review of Books,
11 January 2001, p. 47.

"Something Just Short of Infinity to One"

> This is the kind of question Henry liked to put to himself when he was a schoolboy: what are the chances of this particular fish, from that shoal, off that continental shelf ending up in the pages, no, on this page of this copy of the Daily Mirror? Something just short of infinity to one. Similarly, the grains of sand on a beach, arranged just so. The random ordering of the world, the unimaginable odds against any particular condition, still please him. Even as a child, and especially after Aberfan[1], he never believed in fate or providence, or the future being made by someone in the sky. Instead, at every instant, a trillion trillion futures; the pickiness of pure chance and physical laws seemed like freedom from the scheming of a gloomy god.
>
> Ian McEwan, *Saturday*
> (New York: Nan A. Talese, 2005), pp. 228–229

[1] On 21 October 1966, 144 people, 116 of them children, were killed when thousands of tons of coal waste slid onto the village of Aberfan in South Wales.

Contents

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.

The theory of large deviations has been applied in an astonishingly wide variety of areas including the following: probability and statistical mechanics; nonlinear dynamics, statistics, information theory, engineering, and biology; the study of turbulence, traffic flow, and financial markets; the prediction of hurricanes, earthquakes, avalanches, tsunamis, and volcanic eruptions; the prognosis of epilepsy and heart attacks; the analysis of crowd disasters, opinion dynamics, cultural dynamics, and the evolution of language. As the three quotations given at the start of these notes indicate, large deviations also characterize our lives and life in general and our ability to say anything at all about our miraculous existence, "the most salient feature" of which "is the unthinkable odds against it" (Richard Powers).

These notes 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 [38, §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; e.g., the liquid-gas transition or 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 comprised 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 [38, 63, 89], which cover a number of topics relevant to these notes.

One of my main purposes is to show the utility of the theory of large deviations by applying it to a number of statistical mechanical models. Our applications of the theory include the following.

- A derivation of the form of the Gibbs state for a discrete ideal gas (section 5).
- A probabilistic description of the phase transition in the Curie-Weiss model of a ferromagnet in terms of the breakdown of the law of large numbers for the spin per site (section 9).
- A derivation of the phase-transition structure of the Curie-Weiss-Potts model and the mean-field Blume-Capel model (section 9).
- An analysis of equivalence and nonequivalence of ensembles for a general class of models, including spin models and models of coherent structures in turbulence (section 10).
- A derivation of variational formulas that describe the equilibrium macrostates in models of two-dimensional turbulence (section 11). In terms of these macrostates, coherent vortices of two-dimensional turbulence can be studied.

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.

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 [6]. 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 [65, 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 [54]. 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].

Although the topic of these notes is the theory of large deviations and not the history of science, it is important to appreciate the radical nature of Boltzmann's ideas. His belief in the existence of atoms and his use of probabilistic laws at the microscopic level of atoms and molecules to derive macroscopic properties of matter profoundly challenged the conventional wisdom of 19^{th} century physics: physical laws express absolute truths based not on probabilistic assumptions, but on Newton's laws of motion and precise measurements of observable phenomena.

For his subversive attack on the temple of conventional wisdom, Boltzmann would eventually pay the ultimate price [10, p. 34].

> Boltzmann had never worried about his health, but had sacrificed it to his scientific activity. When however even that vacation in Duino did not bring any relief from his illness, in a moment of deep depression he committed suicide by hanging on 5 September 1906. The next day he should have gone to Vienna to start his lectures.

The irony is that in 1905, the year before Boltzmann's suicide, Einstein applied Boltzmann's insights with great success [65, ch. 11]. In one paper he used Boltzmann's idea to partition the energy of a gas into discrete units in order to explain a phenomenon known as the photoelectric effect. This work would mark the beginning of quantum mechanics and would eventually win Einstein the Nobel Prize. In two other papers also written in 1905 Einstein gave a statistical mechanical explanation based directly on Boltzmann's theory to explain the random motion of a particle suspended in a fluid, a phenomenon known as Brownian motion. This work strongly corroborated the existence of atoms, putting statistical mechanics on a firm theoretical basis. From these papers and two additional 1905 papers on special relativity, the second of which contains the famous formula $E = mc^2$, modern physics was born.

Boltzmann's insights are now part of the canon, but he paid for this with his life. Without his insights, modern physics might never have been born, and unborn, it would not have become our civilization's main conceptual lens for interpreting the universe and our place in it.

We next introduce the theory of large deviations, using nonrigorous but suggestive terminology and notation to motivate the results. Precise formulations are saved for later sections. For a review of the theory emphasizing applications to statistical mechanics, [85] is recommended. Consider a sequence $\{X_j, j \in \mathbb{N}\}$ of independent, identically distributed (i.i.d.) random variables satisfying $E\{\exp(tX_1)\} < \infty$ for all $t \in \mathbb{R}$. For $n \in \mathbb{N}$ we define $S_n = \sum_{j=1}^n X_j$. According to the law of large numbers, $S_n/n \to x_0 = E\{X_1\}$ a.s., and thus for any $\varepsilon > 0$

$$\lim_{n\to\infty} P\{|S_n/n - x_0| \geq \varepsilon\} = 0.$$

The theory of large deviations investigates the exponential rate at which these probabilities converge to 0.

There are two related approaches, both of which are rigorous consequences of Cramér's Theorem 7.3. The first approach focuses on the probability that S_n/n lies in certain subsets of

$\mathbb{R}$. Cramér's Theorem refines the law of large numbers by showing that for any $\varepsilon > 0$ there exists $J(\varepsilon) > 0$ such that

$$P\{|S_n/n - x_0| \geq \varepsilon\} \approx e^{-nJ(\varepsilon)} \to 0.$$

More generally, for appropriate Borel sets A in $\mathbb{R}$ there exists $I(A) \geq 0$ such that

$$P\{S_n/n \in A\} \approx e^{-nI(A)} \tag{1.1}$$

and $I(A) > 0$ if $x_0 = E\{X_1\}$ does not lie in the closure of A. The second approach focuses on the probability that S_n/n is close to a number $x \in \mathbb{R}$, which for now we denote by $P\{S_n/n \in dx\}$. Specifically, we seek a nonnegative function I such that for $x \in \mathbb{R}$

$$P\{S_n/n \in dx\} \approx e^{-nI(x)}; \tag{1.2}$$

in order to be consistent with the law of large numbers, we require that $I(x) = 0$ if and only if $x = x_0$. According to Cramér's Theorem 7.3, $I(x)$ exists and is given by

$$I(x) = \sup_{t \in \mathbb{R}} \{tx - \log E\{e^{tX_1}\}\}.$$

There are obvious consistency conditions linking the two asymptotic formulas (1.1) and (1.2). In order to derive the latter from the former, we set A in (1.1) equal to the open ball $B(x, \varepsilon) = \{y \in \mathbb{R} : |y - x| < \varepsilon$ and require

$$I(x) = \lim_{\varepsilon \to 0} I(B(x, \varepsilon)).$$

Conversely, given a Borel set A in $\mathbb{R}$ we use (1.2) to write

$$P\{S_n/n \in A\} = \int_A P\{S_n/n \in dx\} \approx \int_A e^{-nI(x)}\, dx.$$

We now apply a general asymptotic principle due to Laplace that will be applied numerous times in the sequel. To exponential order the asymptotic behavior of an integral such as $\int_A e^{-nI(x)} dx$ is determined by the largest value of the integrand. Hence it is to be expected that

$$P\{S_n/n \in A\} \approx \exp[-n \inf_{x \in A} I(x)],$$

which when compared with (1.1) implies that $I(A) = \inf_{x \in A} I(x)$.

When suitably interpreted, either formula

$$P\{S_n/n \in A\} \approx e^{-nI(A)} \text{ or } P\{S_n/n \in dx\} \approx e^{-nI(x)}$$

gives the large deviation principle for S_n/n with rate function I. The general concept of a large deviation principle is defined in Definition 6.1.

In order to see an important feature of the formula $P\{S_n/n \in dx\} \approx e^{-nI(x)}$, we consider the special case of tossing a fair die. Let $\Lambda = \{1, 2, 3, 4, 5, 6\}$ and define $P = P_n$ to be the probability measure on the configuration space Λ^n which assigns equal probability $1/6^n$ to each $\omega = (\omega_1, \ldots, \omega_n) \in \Lambda^n$. Then for any subset B of Λ^n

$$P_n\{B\} = \frac{1}{6^n} \cdot \text{card}(B),$$

where card(B) denotes the cardinality of B. The random variable $S_n(\omega)/n = \sum_{j=1}^n \omega_j/n$ maps Λ^n into the closed interval $[1, 6]$. For $x \in [1, 6]$ the formula

$$P\{S_n/n \in dx\} = \frac{1}{6^n} \cdot \text{card}\{\omega \in \Lambda^n : S_n(\omega)/n \in dx\} \approx e^{-nI(x)}$$

shows that $I(x)$ measures the multiplicity of $\omega \in \Lambda^n$ such that $S_n(\omega)/n$ is close to x. In order to anticipate connections with statistical mechanics, we call ω a microstate of the die-tossing game and x a macrostate.

We return to general setup involving the i.i.d. sequence X_j. Another basic concept in the theory of large deviations is the Laplace principle, which can be formally derived from the large deviation formula $P\{S_n/n \in dx\} \approx e^{-nI(x)}$. For any bounded, continuous function f on $\mathbb{R}$ we expect that

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log E\{\exp[nf(S_n/n)] &= \lim_{n\to\infty} \frac{1}{n} \log \int_{\mathbb{R}} \exp[nf(x)]\, P\{S_n/n \in dx\} \\ &= \lim_{n\to\infty} \frac{1}{n} \log \int_{\mathbb{R}} \exp[nf(x)]\, \exp[-nI(x)]\, dx \\ &= \sup_{x\in\mathbb{R}} \{f(x) - I(x)\}. \end{aligned}$$

The last line of this display follows from the observation that the asymptotic behavior of the integral in the second line is determined by the largest value of the integrand. The general concept of a Laplace principle is defined in Definition 6.8. The calculation just given motivates the fact that the Laplace principle is a consequence of the large deviation principle; in fact, the Laplace principle and the large deviation principle are equivalent [Thm. 6.9].

The calculations just given for the i.i.d. sequence X_j indicates the general procedure in classical probability, in which one typically goes from the law of large numbers to the theory of large deviations. Specifically, the law of large numbers implies that certain probabilities converge to 0, and the theory of large deviations calculates the exponential rate of decay of these probabilities.

In applications to statistical mechanics, one typically studies physical systems using a probability distribution known as the canonical ensemble. The form of the canonical ensemble actually makes it easier to derive large deviation estimates directly. From these estimates the law of large numbers and related results follow by considering the set of points at which the rate function equals 0. We will see examples of this procedure in sections 9–11.

We complete the introduction by giving an overview of the contents of each section of these notes.

- **Section 2.** We introduce a basic probabilistic model for random variables having a finite state space.
- **Section 3.** Boltzmann's discovery of the asymptotic behavior of multinomial probabilities in terms of relative entropy is described.
- **Section 4.** We prove a conditional 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?
- **Section 5.** The probabilities of the energy states of a discrete ideal gas are calculated, generalizing the calculation in section 4.

The solutions of the problems in sections 4 and 5 motivate the form of the Gibbs canonical ensemble. This is a probability distribution used to determine the equilibrium properties of statistical mechanical systems; it is discussed in section 9 for a specific model and in section 10 for a general class of models.

- **Section 6.** We introduce the general concepts of a large deviation principle and a Laplace principle together with related results.
- **Section 7.** We prove Cramér's Theorem, which is the large deviation principle for the sample means of i.i.d. random variables.
- **Section 8.** The generalization of Cramér's Theorem known as the Gärtner-Ellis Theorem is presented.

In the remainder of the sections the theory of large deviations is applied to a number of questions in statistical mechanics.

- **Section 9.** The theory of large deviations 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.
- **Section 10.** Our work in the preceding section leads 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.

- **Section 11.** The general procedure developed in the preceding section is used along with Sanov's Theorem to derive variational formulas that describe the equilibrium macrostates in two models of coherent states in two-dimensional turbulence; namely, the Miller-Robert theory and a modification of that theory proposed by Turkington.

Sanov's Theorem, used in section 11 to analyze two models of coherent states in two-dimensional turbulence, generalizes Boltzmann's 1877 calculation. Because Sanov's Theorem plays a vital role in the derivation, this final application of the theory of large deviations brings our focus back home to Boltzmann, whose research in the foundations of statistical mechanics allowed the theory to blossom.

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

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

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

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 $\text{card}(B)$ 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

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

Example 2.1. 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, noninteracting 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.

- $\Omega = \Omega_n = \Lambda^n$, where $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$.
- $\mathcal{F}$ is the set of all subsets of Λ^n.
- 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 relative entropy.

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 these lectures, 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; $y_1 < y_2 < \ldots < y_\alpha$ a set of α real numbers; $\rho_1, \rho_2, \ldots, \rho_\alpha$ a set of α positive real numbers summing to 1; Λ the set $\{y_1, y_2, \ldots, y_\alpha\}$; and P_n 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 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 $\varepsilon > 0$

$$\lim_{n \to \infty} P_n\{L_n \in B(\rho, \varepsilon)\} = 1. \tag{3.1}$$

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

$$\lim_{n\to\infty} P_n\{L_n \in B(\gamma, \varepsilon)\} = 0. \tag{3.2}$$

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 3.1 (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., Prop. 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 convexity inequality is that for $x \geq 0$, $x \log x \geq x - 1$ with equality if and only if $x = 1$.

Lemma 3.2. *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 $\gamma \in \mathcal{P}_\alpha$

$$\frac{\gamma_k}{\rho_k} \log \frac{\gamma_k}{\rho_k} \geq \frac{\gamma_k}{\rho_k} - 1 \tag{3.3}$$

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 $I_\rho(\gamma)$. 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 3.4. 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 3.3 (Boltzmann's Discovery–Formulation 1). *For any* $\gamma \in \mathcal{P}_\alpha$

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

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

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

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

Pseudo-Theorem 3.3 has the following interesting consequence. Let γ be any vector in $\mathcal{P}_\alpha$ which differs from ρ. Since $I_\rho(\gamma) > 0$ [Lem. 3.2], 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 (3.2).

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 3.3 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

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

one expects that to exponential order

$$P_n\{L_n \in A\} \approx \exp[-nI_\rho(A)] \text{ as } n \to \infty. \tag{3.4}$$

As formulated in Corollary 3.5, 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 6.1, and a general form of Sanov's Theorem is stated in Theorem 6.7. The special case of Sanov's Theorem stated next is proved in Theorem 7.5.

Theorem 3.4 (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 $\mathcal{P}_\alpha$*

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

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

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

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 $\overline{B}(\gamma, \varepsilon)$:

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

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

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

The bounds in the last two displays can be proved via combinatorics and Stirling's formula as in the heuristic proof of Pseudo-Theorem 3.3; one can easily adapt the calculations given in [38, §I.4]. The details are omitted. ■

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 we can now derive a rigorous version of the asymptotic formula (3.4). 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.

Corollary 3.5. *Let A be any Borel subset of $\mathcal{P}_\alpha$ satisfying $\overline{A^\circ} = \overline{A}$. Then*

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

Proof. We apply the large deviation upper bound to $\overline{A}$ and the large deviation lower bound to A°. Since $\overline{A} \supset A \supset A^\circ$, it follows that $I_\rho(\overline{A}) \le I_\rho(A) \le I_\rho(A^\circ)$ and that

$$\begin{aligned} -I_\rho(\overline{A}\,) &\ge \limsup_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in \overline{A}\,\} \\ &\ge \limsup_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in A\} \\ &\ge \liminf_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in A\} \\ &\ge \liminf_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in A^\circ\} \\ &\ge -I_\rho(A^\circ). \end{aligned}$$

The continuity of I_ρ on $\mathcal{P}_\alpha$ implies that $I_\rho(A^\circ) = I_\rho(\overline{A^\circ})$. Since by hypothesis $\overline{A^\circ} = \overline{A}$, we conclude that the extreme terms in this display are equal to each other and to $I_\rho(A)$ and thus that

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

The desired limit follows. ■

The next corollary of Theorem 3.4 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 6.4 is extremely useful in applications. For example, we will use it in section 9 to analyze several lattice spin models and in section 10 to motivate the definitions of the sets of equilibrium macrostates for the canonical ensemble and the microcanonical ensemble for a general class of systems [Thms. 10.2(c), 10.5(c)].

Corollary 3.6. *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 $C < \infty$*

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

Proof. Since $I_\rho(\gamma) > I_\rho(\rho) = 0$ for any $\gamma \ne \rho$, 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 3.6 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 (3.1), 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. This limit is the simplest example, and the first of several more complicated but related formulations to be encountered in this paper, of what is commonly called a maximum entropy principle. Following the usual convention in the physical literature, we will continue to use this terminology in referring to such principles even though we are minimizing the relative entropy — equivalently, maximizing $-I_\rho(\gamma)$ — rather than maximizing the physical entropy. When $\rho_k = 1/\alpha$ for each k, the two quantities differ by a minus sign and an additive constant.

Maximum Entropy Principle 3.7. *$\gamma_0 \in \mathcal{P}_\alpha$ is an equilibrium value of L_n with respect to P_n if and only if γ_0 minimizes $I_\rho(\gamma)$ over $\mathcal{P}_\alpha$; this occurs if and only if $\gamma_0 = \rho$.*

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 3.4.

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 occurs, then one could efficiently redesign the system so that the overload does not happen. This conditioned limit theorem is an ideal learning tool because it involves an unexpected application of the large deviation estimates for the empirical vector discussed in section 3.

We use the notation of section 3. Thus let $\alpha \geq 2$ be an integer; $y_1 < y_2 < \ldots < y_\alpha$ a set of α real numbers; $\rho_1, \rho_2, \ldots, \rho_\alpha$ a set of α positive real numbers summing to 1; Λ the set $\{y_1, y_2, \ldots, y_\alpha\}$; and P_n 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 given in Theorem 4.1. This theorem has the bonus of giving insight into a basic construction in statistical mechanics. As we will see in section 5, 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 $P_n\{L_n \in A\}$ for subsets A of $\mathcal{P}_\alpha$.

The conditioned limit theorem that we will consider can be restated in the following form (see step 3 in the proof of Theorem 4.1). Suppose that one is given a particular set Γ for which $P_n\{L_n \in \Gamma\} > 0$ 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

$$\lim_{n\to\infty} P_n\{L_n \in B \,|\, L_n \in \Gamma\} = \lim_{n\to\infty} P_n\{L_n \in B \cap \Gamma\} \cdot \frac{1}{P_n\{L_n \in \Gamma\}} = 1 \qquad (4.1)$$

is valid. In order to motivate the answer, we use the formal, large-deviation notation

$$P_n\{L_n \in A\} \approx \exp[-nI_\rho(A)] \text{ as } n \to \infty \tag{4.2}$$

for Borel subsets A of $\mathcal{P}_\alpha$. This notation was introduced in (3.4). Since to exponential order

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

one should obtain the conditioned limit (4.1) if B satisfies $I_\rho(B \cap \Gamma) = I_\rho(\Gamma)$. If one can determine the point in Γ where the infimum of I_ρ is attained, then one picks B to contain this point. As we explain in Proposition 4.3, such a minimizing point can be determined for an important class of sets Γ. It will lead to a second maximum entropy principle for L_n with respect to the conditional probabilities $P_n\{\cdot \,|\, L_n \in \Gamma\}$.

We first formulate the conditioned limit theorem for the case of tossing a die. The basic probabilistic model introduced in Example 2.1(b) involves the following quantities.

- Ω_n equals Λ^n, where $\Lambda = \{1, 2, 3, 4, 5, 6\}$.
- For each $k \in \{1, \ldots, 6\}$, $\rho_k = 1/6$.
- For each $\omega \in \Omega_n$, $P_n\{\omega\} = 1/6^n$.
- For each $j \in \{1, 2, \ldots, n\}$, $X_j(\omega) = \omega_j$.
- $S_n(\omega) = \sum_{j=1}^n X_j(\omega) = \sum_{j=1}^n \omega_j$.

We define

$$\bar{y} = \sum_{k=1}^{6} k\rho_k = 3.5,$$

which equals $E^{P_n}\{X_1\}$. 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 $1 < z - a < z < \bar{y}$. The conditioned limit for the case of die tossing involves positive numbers $\{\rho_k^*, k = 1, \ldots, 6\}$ summing to 1 and satisfying

$$\rho_k^* = \lim_{n\to\infty} P_n\{X_1 = k \,|\, S_n/n \in [z - a, z]\}. \tag{4.3}$$

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 $\bar{y} < z < z + a < 6$.

The limit (4.3) 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 $\rho^* =$

$(\rho_1^*, \ldots, \rho_6^*)$ 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 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.

Interpretation of (4.3) in terms of a crooked gambling game. You participate in a crooked gambling game being played with a loaded die, which is represented by the event that $S_n/n \in [z-a, z]$ for all n. Then the limit (4.3) represents the actual probabilities of the loaded die. Do you agree with this interpretation?

We formulate the limit in the last display for a general state space $\Lambda = \{y_1, \ldots, y_\alpha\}$ and a given positive vector $\rho = (\rho_1, \ldots, \rho_\alpha) \in \mathcal{P}_\alpha$. As above, define

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

and for $a > 0$ fix a closed interval $[z-a, z] \subset (y_1, \bar{y})$.

Theorem 4.1. *Let $z \in (y_1, \bar{y})$ be given and choose $a > 0$ such that $z - a > y_1$. Then for each $k \in \{1, 2, \ldots, \alpha\}$*

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

The quantity $\rho^{(\beta)} = (\rho_1^{(\beta)}, \ldots, \rho_\alpha^{(\beta)})$ is a probability vector having 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 $\sum_{k=1}^\alpha y_k \rho_k^{(\beta)} = z$.

In this theorem S_n/n is conditioned to lie in the interval $[z-a, z]$, where $a > 0$ and $[z-a, z] \subset (y_1, \bar{y})$. We point out alternative versions of the theorem for different choices of the interval in which S_n/n is conditioned to lie. First, consider the interval $[z, z+a]$, where $a > 0$ and $[z, z+a] \subset (\bar{y}, y_\alpha)$. In this case the theorem holds as stated except that $\beta > 0$. Second, consider any interval containing $\bar{y}$ in its interior. In this case the theorem holds as stated except $\beta = 0$.

In Theorem 5.1 we generalize Theorem 4.1 by proving that for any y_{k_1} and y_{k_2} in Λ

$$\begin{aligned}
&\lim_{n\to\infty} P_n\{X_1 = y_{k_1}, X_2 = y_{k_2} \,|\, S_n/n \in [z-a, z]\} \\
&\quad = \rho_{k_1}^{(\beta)} \rho_{k_2}^{(\beta)} \\
&\quad = \lim_{n\to\infty} P_n\{X_1 = y_{k_1} \,|\, S_n/n \in [z-a, z]\} \cdot \lim_{n\to\infty} P_n\{X_2 = y_{k_2} \,|\, S_n/n \in [z-a, z]\}.
\end{aligned}$$

This limit is somewhat surprising. Indeed, although X_1 and X_2 are independent with respect to the original product measure P_n, this independence is lost when P_n is replaced by the conditional distribution $P_n\{\cdot \mid S_n/n \in [z-a,z]\}$. However, in the limit $n \to \infty$ the independence of X_1 and X_2 is regained. More generally, in Theorem 5.1 we prove an analogous result for the limiting conditional distribution of $X_1, X_2, \ldots, X_\ell$ for any $\ell \geq 3$ when S_n/n is conditioned to lie in $[z-a,z]$. This limiting conditional distribution is the product measure on Ω_ℓ with one-dimensional marginals $\rho^{(\beta)}$.

The proof of Theorem 4.1 involves a number of technicalities including the rigorous formulation of Boltzmann's exponential-order estimates in Theorem 3.4 and Corollary 3.5. The proof is postponed until later in this section. I next give a heuristic proof of the theorem that in my opinion gives more insight that the actual proof. Our goal is to verify that the limit

$$\rho_k^{(\beta)} = \lim_{n\to\infty} P_n\{X_1 = y_k \mid S_n/n \in dz\} \tag{4.4}$$

exists and has the form given in Theorem 4.1. The event $\{S_n/n \in [z-a,z]\}$ has been rewritten using the formal notation $\{S_n/n \in dz\}$. The basic idea is to rewrite this limit in terms of the empirical vector L_n and then use the exponential-order estimate (3.4).

We start by showing that the probability vector $\rho^{(\beta)}$ in Theorem 4.1 is well defined. In order to carry this out, we introduce for $\beta \in \mathbb{R}$

$$c(\beta) = \log E^{P_n}\{e^{\beta X_1}\} = \log\left(\sum_{k=1}^{\alpha} \exp[\beta y_k]\,\rho_k\right). \tag{4.5}$$

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 Lemma 4.2 we prove that $c''(\beta) > 0$ for all β, which implies that c' is strictly increasing on $\mathbb{R}$. From the formula

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

one easily verifies that

$$\lim_{\beta\to-\infty} c'(\beta) = y_1 \;\text{ and }\; \lim_{\beta\to\infty} c'(\beta) = y_\alpha.$$

It follows that c' is a one-to-one function mapping $\mathbb{R}$ onto the open interval (y_1, y_α). Thus there exists a unique $\beta = \beta(z)$ such that

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

We conclude that the probability vector $\rho^{(\beta)}$ in Theorem 4.1 is well defined. Since $y_1 < z < \bar{y}$ and $c'(0) = \bar{y}$, $\beta(z)$ is negative.

For $z \in (y_1, \bar{y})$ we introduce the set

$$\Gamma(z) = \left\{ \gamma \in \mathcal{P}_\alpha : \sum_{k=1}^{\alpha} y_k \gamma_k \in dz \right\}, \tag{4.6}$$

where the formal notation $\sum_{k=1}^{\alpha} y_k \gamma_k \in dz$ represents the constraint $\sum_{k=1}^{\alpha} y_k \gamma_k \in [z-a, z]$. The next step is to show that $\rho^{(\beta)}$ has the property that

$$\lim_{n \to \infty} P_n\{L_n \in d\rho^{(\beta)} \,|\, L_n \in \Gamma(z)\} = 1, \tag{4.7}$$

where the event $\{L_n \in d\rho^{(\beta)}\}$ stands for the event $\{L_n \in B(\rho^{(\beta)}, \varepsilon)\}$ for any $\varepsilon > 0$. We then show how the desired limit (4.4) follows from the limit in the last display.

A key fact used to prove (4.7) is that I_ρ attains its infimum over $\Gamma(z)$ at the unique vector $\rho^{(\beta)}$. The actual proof is postponed until Proposition 4.3. We give the main idea here, first motivating this fact by considering 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

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

It follows that $\gamma_k = \exp[-\lambda - 1] \exp[\beta y_k] \rho_k$. Now pick λ so that $\sum_{k=1}^{\alpha} \gamma_k = 1$ and pick $\beta = \beta(z)$ so that $\sum_{k=1}^{\alpha} y_k \gamma_k = z$. With these choices $\gamma_k = \rho_k^{(\beta)}$, where $\beta = \beta(z)$ is as specified in Theorem 4.1.

We now show that I_ρ attains its infimum over the set

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

at the unique vector $\rho^{(\beta)}$. This proof is based on properties of the relative entropy and does not use calculus. The set $\Gamma(z)$ is defined in (4.6) in terms of the constraint $\sum_{k=1}^{\alpha} y_k \gamma_k \in dz$. For simplicity, in the definition of $\widetilde{\Gamma}(z)$ this constraint is replaced by the equality constraint $\sum_{k=1}^{\alpha} y_k \gamma_k = z$. 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 (4.5). Hence for any $\gamma \in \widetilde{\Gamma}(z)$

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

Since $I_{\rho^{(\beta)}}(\gamma) \geq 0$ with equality if and only if $\gamma = \rho^{(\beta)}$ [Lem. 3.2], it follows that if $\gamma = \rho^{(\beta)}$, then $I_\rho(\rho^{(\beta)}) = \beta z - c(\beta)$ and that if $\gamma \neq \rho^{(\beta)}$, then

$$I_\rho(\gamma) > \beta z - c(\beta) = I_\rho(\rho^{(\beta)}).$$

This completes the proof that I_ρ attains its infimum over $\widetilde{\Gamma}(z)$ at the unique vector $\rho^{(\beta)}$.

We now motivate the limit (4.7) using the formal, large-deviation notation (4.2), which was introduced in (3.4). Sanov's Theorem 6.7 in combination with Theorem 6.3 is the rigorous statement of the large deviation limits for the distribution of L_n; a special case that applies to the current setup is given in Theorem 3.4 and Corollary 3.5. For large n we have by (4.2)

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

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

$$I_\rho(B(\rho^{(\beta)}, \varepsilon) \cap \Gamma(z)) = I_\rho(\Gamma(z)). \tag{4.8}$$

This is indeed the case since, as has just been shown, I_ρ attains its infimum over $\Gamma(z)$ at the unique point $\rho^{(\beta)}$. This gives (4.8) and completes the motivation of the limit (4.7), which states that

$$\lim_{n\to\infty} P_n\{L_n \in d\rho^{(\beta)} \,|\, L_n \in \Gamma(z)\} = 1.$$

Since L_n takes values in the set of probability vectors $\mathcal{P}_\alpha$, L_n lies in $\Gamma(z)$ if and only if $\sum_{k=1}^{\alpha} y_k L_n(y_k) \in dz$. We claim that for each $\omega \in \Omega_n$

$$\sum_{k=1}^{\alpha} y_k L_n(\omega, y_k) = S_n(\omega)/n; \tag{4.9}$$

i.e., that the mean of the empirical vector equals the sample mean S_n/n. It will then follow from this equality and the limit (4.7) that

$$\lim_{n\to\infty} P_n\{L_n \in d\rho^{(\beta)} \,|\, S_n/n \in dz\} = \lim_{n\to\infty} P_n\{L_n \in d\rho^{(\beta)} \,|\, L_n \in \Gamma(z)\} = 1. \tag{4.10}$$

This is just one step away from the desired limit (4.4). The proof of (4.9) is straightforward. For each $\omega \in \Omega_n$

$$\begin{aligned}\sum_{k=1}^{\alpha} y_k L_n(\omega, y_k) &= \sum_{k=1}^{\alpha} y_k \cdot \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j(\omega)}(y_k) \\ &= \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{\alpha} y_k \delta_{X_j(\omega)}(y_k) \\ &= \frac{1}{n} \sum_{j=1}^{n} X_j(\omega) \; = \; S_n/n.\end{aligned}$$

This completes the proof of (4.9) and yields the limit (4.10).

We are now ready to motivate the desired limit (4.4). The limit (4.10) implies that for all large n and all $k \in \{1, 2, \ldots, \alpha\}$, we have with probability close to 1

$$\begin{aligned}\rho_k^{(\beta)} &= E^{P_n}\{\rho_k^{(\beta)} \,|\, S_n/n \in dz\} \;\approx\; E^{P_n}\{L_n(y_k) \,|\, S_n/n \in dz\} \\ &= \frac{1}{n} \sum_{j=1}^{n} E^{P_n}\{\delta_{X_j}(y_k) \,|\, S_n/n \in dz\} \\ &= \frac{1}{n} \sum_{j=1}^{n} P_n\{X_j = y_k \,|\, S_n/n \in dz\} \\ &= P_n\{X_1 = y_k \,|\, S_n/n \in dz\}.\end{aligned}$$

The last line follows by symmetry. This completes the motivation of the desired limit (4.4):

$$\lim_{n\to\infty} P_n\{X_1 = y_k \,|\, S_n/n \in dz\} = \rho_k^{(\beta)}.$$

The heuristic proof of Theorem 4.1 is done.

We now prove Theorem 4.1 with full details. In order to show that $\rho^{(\beta)}$ in Theorem 4.1 is well defined, we need the following lemma concerning properties of the cumulant generating function c defined in (4.5).

Lemma 4.2. *For $\beta \in \mathbb{R}$ the cumulant generating function $c(\beta) = \log(\sum_{k=1}^{\alpha} \exp[\beta y_k]\rho_k)$ has the following properties.*

(a) *$c''(\beta) > 0$ for all β; i.e., c is strictly convex on $\mathbb{R}$.*

(b) *$c'(0) = \sum_{k=1}^{\alpha} y_k \rho_k = \bar{y}$.*

(c) *$c'(\beta) \to y_1$ as $\beta \to -\infty$ and $c'(\beta) \to y_\alpha$ as $\beta \to \infty$.*

(d) *c' is a one-to-one function mapping $\mathbb{R}$ onto the open interval (y_1, y_α), which is the interior of the smallest interval containing the set $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$.*

Proof. (a) We define

$$\langle y \rangle_\beta = \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \sum_{k=1}^{\alpha} y_k \, \exp[\beta y_k]\, \rho_k$$

and

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

and calculate

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

and

$$\begin{aligned} c''(\beta) &= \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j} \cdot \sum_{k=1}^{\alpha} y_k^2 \, \exp[\beta y_k]\, \rho_k - \langle y \rangle_\beta^2 \\ &= \langle (y - \langle y \rangle_\beta)^2 \rangle_\beta \, > \, 0. \end{aligned}$$

The last line gives part (a). This calculation shows that $c'(\beta)$ equals the mean of the probability vector $\exp[\beta y_k]\rho_k / \sum_{j=1}^{\alpha} \exp[\beta y_j]\, \rho_j$, and $c''(\beta)$ equals the variance of this probability vector.

(b) This follows from the formula for $c'(\beta)$ in part (a).

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

$$\lim_{\beta \to -\infty} c'(\beta) = \lim_{\beta \to -\infty} \frac{1}{\sum_{j=1}^{\alpha} \exp[\beta(y_j - y_1)]\, \rho_j} \cdot \sum_{k=1}^{\alpha} y_k \, \exp[\beta(y_k - y_1]\, \rho_k = y_1.$$

One similarly proves that $\lim_{\beta \to \infty} c'(\beta) = y_\alpha$.

(d) According to part (a), $c'(\beta)$ is a strictly increasing function of β. It follows from the limits in part (c) that c' is a one-to-one function mapping $\mathbb{R}$ onto the open interval (y_1, y_α). This completes the proof of the lemma. ■

We now prove Theorem 4.1.

Proof of Theorem 4.1. The probability vector $\rho^{(\beta)}$ in Theorem 4.1 is well defined because of Lemma 4.2, which shows that there exists a unique $\beta = \beta(z)$ satisfying

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

as claimed. Since $y_1 < z < \bar{y}$ and $c'(0) = \bar{y}$, $\beta(z)$ is negative.

The proof of Theorem 4.1 proceeds in four steps.

Step 1. 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 $\rho^{(\beta)}$. Step 1 is to prove that for every $\varepsilon > 0$

$$\lim_{n\to\infty} P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, L_n \in \Gamma(z)\} = 1. \tag{4.11}$$

The proof of this limit depends on the next proposition.

Proposition 4.3. *Let $z \in (y_1, \bar{y})$ be given. Then I_ρ attains its infimum over*

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

at the unique point $\rho^{(\beta)} = (\rho_1^{(\beta)}, \ldots, \rho_\alpha^{(\beta)})$ defined in Theorem 4.1: *for each $k = 1, \ldots, \alpha$*

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

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

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 (4.5). 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

$$I_\rho(\rho^{(\beta)}) = I_{\rho^{(\beta)}}(\rho^{(\beta)}) + \beta \sum_{k=1}^{\alpha} y_k \rho_k^{(\beta)} - c(\beta) = \beta z - c(\beta). \tag{4.12}$$

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 $\gamma = \rho^{(\beta)}$ [Lem. 3.2], 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 is complete. ■

We now prove the limit (4.11) by showing that

$$\lim_{n \to \infty} P_n\{L_n \in [B(\rho^{(\beta)}, \varepsilon)]^c \,|\, L_n \in \Gamma(z)\} = 0. \tag{4.13}$$

The key is to use Corollary 3.5, 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 3.5, it follows that

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

According to Proposition 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 display this yields the desired limit (4.13), showing in fact that the convergence to 0 is exponentially fast. This completes the proof of the limit (4.11). Step 1 in the proof of Theorem 4.1 is done.

Step 2. Step 2 is to prove that

$$\lim_{n \to \infty} P_n\{L_n \in B(\rho^{(\beta)}, \varepsilon) \,|\, S_n/n \in [z - a, z]\} = 1. \tag{4.14}$$

Since for each $\omega \in \Omega_n$

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

it follows that $\{\omega \in \Omega_n : L_n(\omega) \in \Gamma(z)\} = \{\omega \in \Omega_n : S_n(\omega)/n \in [z - a, z]\}$. Hence the limit (4.11) yields the limit (4.14). Step 2 in the proof of Theorem 4.1 is done.

Step 3. Step 3 is to prove that for any continuous function f mapping $\mathcal{P}_\alpha$ into $\mathbb{R}$

$$\lim_{n\to\infty} E^{P_n}\{f(L_n) \,|\, S_n/n \in [z-a,z]\} = f(\rho^{(\beta)}). \tag{4.15}$$

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

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

By (4.14) the probability in the last line of the display converges to 0 as $n \to \infty$. Since $\delta > 0$ is arbitrary, the limit (4.15) is proved. Step 3 in the proof of Theorem 4.1 is done.

Step 4. Step 4 is to prove that for each $k \in \{1,\dots,\alpha\}$

$$\lim_{n\to\infty} P_n\{X_1 = y_k \,|\, S_n/n \in [z-a,z]\} = \rho_k^{(\beta)}. \tag{4.16}$$

By symmetry

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

Now define f_k to be the continuous function on $\mathcal{P}_\alpha$ that maps γ to γ_k. The limit (4.15) yields

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

This is the limit (4.16). The proof of Theorem 4.1 is complete. ■

We next comment on the relationship between Proposition 4.3 and a basic result in the theory of large deviations known as the contraction principle. In Proposition 4.3 we prove that for $z \in (y_1, \bar{y})$ I_ρ attains its infimum over

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

at the unique point $\rho^{(\beta)} = (\rho_1^{(\beta)}, \ldots, \rho_\alpha^{(\beta)})$. For each $k = 1, \ldots, \alpha$

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

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

$$\widetilde{\Gamma}(z) = \{\gamma \in \mathcal{P}_\alpha : \sum_{k=1}^{\alpha} y_k \gamma_k = z\};$$

Since $\rho^{(\beta)}$ is an element of $\widetilde{\Gamma}(z)$, it follows that the infimum of I_ρ over $\Gamma(z)$ coincides with the infimum of I_ρ over $\widetilde{\Gamma}(z)$, and by (4.12) this infimum coincides with the quantity $\beta z - c(\beta)$. For $y \in \mathbb{R}$ we define

$$I(y) = \sup_{\beta \in \mathbb{R}} \{\beta y - c(\beta)\}.$$

Since c is strictly convex [Lem. 4.2)(a)] and $\beta = \beta(z)$ is the unique value of β satisfying

$$\sum_{k=1}^{\alpha} y_k \rho_k^{(\beta)} = c'(\beta) = z,$$

it follows that $I(z) = \beta z - c(\beta)$. We conclude that

$$\begin{aligned} &\inf\{I_\rho(\gamma) : \gamma \in \mathcal{P}_\alpha : \sum_{k=1}^{\alpha} y_k \gamma_k \in [z-a, z]\} \qquad (4.17) \\ &= \inf\{I_\rho(\gamma) : \gamma \in \mathcal{P}_\alpha : \sum_{k=1}^{\alpha} y_k \gamma_k = z\} = I(z). \end{aligned}$$

This is a special case of an equality stated in (6.6), which is a consequence of a general contraction principle given in Theorem 6.12. It is not difficult to show that the same formula is valid for any $z \in [y_1, y_\alpha]$.

The equality stated in (4.17) plays a central role in the theory because it leads to a special case of Cramér's Theorem for the sequence $\{X_j, j \in \mathbb{N}\}$ of i.i.d. random variables considered in this section. These random variables take values in $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$ and have common distribution $\rho = \sum_{k=1}^{\alpha} \rho_k \delta_{y_k}$. For $n \in \mathbb{N}$ define the sample mean $S_n/n = \sum_{j=1}^{n} X_j/n$, which takes values in $[y_1, y_\alpha]$. Using the large deviation principle for L_n given in Theorem 3.4, we show that the sequence S_n/n satisfies the large deviation principle on $[y_1, y_\alpha]$ with rate function $I(y)$.

We start by defining the continuous function ψ mapping $\mathcal{P}_\alpha$ into $[y_1, y_\alpha]$ by $\psi(\gamma) = \sum_{k=1}^{\alpha} \gamma_k y_k$. The key idea in deriving Cramér's Theorem is to use ψ to relate S_n/n to L_n as follows:

$$\frac{S_n}{n} = \frac{1}{n} \sum_{k=1}^{\alpha} y_k \cdot \text{card}\{j \in \{1, \ldots, n\} : X_j = y_k\} = \sum_{k=1}^{\alpha} y_k L_n(y_k) = \psi(L_n).$$

In terms of ψ, (4.17) states that for $z \in [y_1, y_\alpha]$

$$\inf\{I_\rho(\gamma) : \gamma \in \mathcal{P}_\alpha, \psi(\gamma) = z\} = I(z) = \sup_{\beta \in \mathbb{R}}\{\beta z - c(\beta)\}.$$

Now let K be any closed subset of $[y_1, y_\alpha]$. Since ψ is continuous, $\psi^{-1}(K)$ is a closed subset of $\mathcal{P}_\alpha$. Hence by the large deviation upper bound in part (a) of Theorem 3.4

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log P_n\{S_n/n \in K\} \\
&\quad = \limsup_{n\to\infty} \frac{1}{n} \log P_n\{\psi(L_n) \in K\} \\
&\quad = \limsup_{n\to\infty} \frac{1}{n} \log P_n\{L_n \in \psi^{-1}(K)\} \\
&\quad = -\inf\{I_\rho(\gamma) : \gamma \in \psi^{-1}(K)\} = -\inf\{I_\rho(\gamma) : \psi(\gamma) \in K\} \\
&\quad = -\inf_{z\in K} (\inf\{I_\rho(\gamma) : \gamma \in \mathcal{P}_\alpha, \psi(\gamma) = z\}) \\
&\quad = -\inf_{z\in K} I(z) = -I(K).
\end{aligned}$$

A similar calculation shows that if U is any open subset of $[y_1, y_\alpha]$, then

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n\{S_n/n \in U\} \geq -\inf_{z\in U} I(z) = -I(U).$$

This gives the following special case of Cramér's Theorem. More general formulations are given in section 7.

Theorem 4.4. *The sequence of sample means S_n/n satisfies the large deviation principle on $[y_1, y_\alpha]$ with rate function $I(z) = \sup_{\beta\in\mathbb{R}}\{\beta z - c(\beta)\}$ in the following sense.*

(a) *For any closed subset K of $[y_1, y_\alpha]$*

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n\{S_n/n \in K\} \leq -I(K).$$

(b) *For any open subset U of $[y_1, y_\alpha]$*

$$\liminf_{n\to\infty} \frac{1}{n} \log P_n\{S_n/n \in U\} \geq -I(U).$$

We return to Proposition 4.3, combining it with part (a) of Theorem 4.1 to give the second maximum entropy principle in these notes.

Maximum Entropy Principle 4.5. *Conditioned on the event $S_n/n \in [z-a, z]$, the asymptotically most likely configuration of L_n is $\rho^{(\beta)}$, which is the unique $\gamma \in \mathcal{P}_\alpha$ that minimizes $I_\rho(\gamma)$ subject to the constraint that $\gamma \in \Gamma(z)$. In statistical mechanical terminology, $\rho^{(\beta)}$ is the equilibrium macrostate of L_n with respect to the conditional probabilities $P_n\{\cdot \mid S_n/n \in [z-a, z]\}$.*

As shown in (4.15), for any continuous function f mapping $\mathcal{P}_\alpha$ into $\mathbb{R}$

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

This limit is another expression of the Maximum Entropy Principle 4.5.

With some additional work one can generalize part (a) of Theorem 4.1 by proving that with respect to the conditional probabilities $P_n\{\cdot \mid S_n/n \in [z-a, a]\}$, L_n satisfies the large deviation principle on $\mathcal{P}_\alpha$ with rate function

$$I(\gamma) = \begin{cases} I_\rho(\gamma) - I_\rho(\Gamma(z)) & \text{if } \gamma \in \Gamma(z) \\ \infty & \text{if } \gamma \in \mathcal{P}_\alpha \setminus \Gamma(z). \end{cases}$$

This large deviation principle is closely related to the large deviation principle for statistical mechanical models with respect to the microcanonical ensemble, which will be considered in Theorem 10.5.

In the next section we will show how calculations analogous to those used to motivate Theorem 4.1 can be used to derive the form of the Gibbs state for the discrete ideal gas.

5 Canonical Ensemble for Models in Statistical Mechanics

The discussion in the preceding section concerning a loaded die applies with minor changes to the discrete ideal gas, introduced in part (c) of Examples 2.1. We continue to use the notation of the preceding sections. Thus let $\alpha \geq 2$ be an integer; $y_1 < y_2 < \ldots < y_\alpha$ a set of α real numbers; $\rho_1, \rho_2, \ldots, \rho_\alpha$ a set of α positive real numbers summing to 1; Λ the set $\{y_1, y_2, \ldots, y_\alpha\}$; and P_n 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.

The discrete ideal gas consists of n identical, noninteracting particles, each having α possible energy levels $y_1, y_2, \ldots, y_\alpha$. For $\omega \in \Omega_n$ we write $H_n(\omega)$ in place of $S_n(\omega) = \sum_{j=1}^{n} \omega_j$; $H_n(\omega)$ denotes the total energy in the configuration ω. In the absence of further information, one assigns the equal probabilities $\rho_k = 1/\alpha$ to each of the y_k's. Defining $\bar{y} = \sum_{k=1}^{\alpha} y_k \rho_k$, suppose that the energy per particle, H_n/n, is conditioned to lie in an interval $[z - a, z]$, where a is a small positive number and $y_1 \leq z - a < z < \bar{y}$. According Theorem 4.1, for each $k \in \{1, \ldots, \alpha\}$

$$\begin{aligned} &\lim_{n\to\infty} P_n\{X_1 = y_k \,|\, H_n/n \in [z-a, z]\} \\ &\quad = \rho_k^{(\beta)} = \frac{1}{\sum_{j=1}^{\alpha} \exp[-\beta y_j]\, \rho_j} \cdot \exp[-\beta y_k]\, \rho_k, \end{aligned}$$

where $\beta = \beta(z) \in \mathbb{R}$ is the unique value of β satisfying $\sum_{k=1}^{\alpha} y_k \rho_k^{(\beta)} = z$. In order to be consistent with conventions in statistical mechanics, we have replaced β by $-\beta$ in the definition of $\rho_k^{(\beta)}$. Thus in the last display $\beta = \beta(z)$ is positive.

Let $\ell \geq 2$ be a positive integer. The limit in the last display leads to a natural question. Conditioned on $H_n/n \in [z - a, z]$, as $n \to \infty$ what is the limiting conditional distribution of the random variables $X_1, \ldots, X_\ell$, which represent the energy levels of the first ℓ particles? Although $X_1, \ldots, X_\ell$ are independent with respect to the original product measure P_n, this independence is lost when P_n is replaced by the conditional distribution $P_n\{\cdot \,|\, H_n/n \in [z - a, z]\}$. Hence the answer given in the next theorem is somewhat surprising: with respect to $P_n\{\cdot \,|\, H_n/n \in [z - a, z]\}$, the limiting distribution is the product measure on Ω_ℓ with one-dimensional marginals $\rho^{(\beta)}$. In other words, in the limit $n \to \infty$ the independence of $X_1, \ldots, X_\ell$ is regained. The theorem leads to, and in a sense motivates, the form of the canonical ensemble of the discrete ideal gas. We will end the section by discussing Gibbs states for this and other statistical mechanical models. As in Theorem 4.1, a theorem analogous to the following would hold if $[z - a, z] \subset [y_1, \bar{y})$ were replaced by $[z, z + a] \subset (\bar{y}, y_\alpha]$.

Theorem 5.1. *Let $\ell \in \mathbb{N}$, $\ell \geq 2$, $y_{k_1}, \ldots, y_{k_\ell} \in \Lambda$, and $[z - a, z] \subset [y_1, \bar{y})$ be given. Then*

$$\lim_{n\to\infty} P_n\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell} \,|\, H_n/n \in [z-a, z]\} = \prod_{j=1}^{\ell} \rho_{k_j}^{(\beta)}. \tag{5.1}$$

Comments on the Proof. We consider $\ell = 2$; arbitrary $\ell \in \mathbb{N}$ can be handled similarly. For $\omega \in \Omega_n$ and $i, j \in \{1, \ldots, \alpha\}$ define

$$\begin{aligned} L_{n,2}(\{y_i, y_j\}) &= L_{n,2}(\omega, \{y_i, y_j\}) \\ &= \frac{1}{n}\left(\sum_{j=1}^{n-1} \delta_{X_j(\omega), X_{j+1}(\omega)}\{y_i, y_j\} + \delta_{X_n(\omega), X_1(\omega)}\{y_i, y_j\}\right). \end{aligned}$$

This counts the relative frequency with which the pair $\{y_i, y_j\}$ appears in the configuration $(\omega_1, \ldots, \omega_n, \omega_1)$. We then define the empirical pair vector

$$L_{n,2} = \{L_{n,2}(\{y_i, y_j\}), i, j = 1, \ldots, \alpha\}.$$

This takes values in the set $\mathcal{P}_{\alpha,2}$ consisting of all $\tau = \{\tau_{i,j}, i, j = 1, \ldots, \alpha\}$ satisfying $\tau_{i,j} \geq 0$ and $\sum_{i,j=1}^{\alpha} \tau_{i,j} = 1$. Suppose one can show that $\tau^* = \{\rho_i^{(\beta)} \rho_j^{(\beta)}, i, j = 1, \ldots, \alpha\}$ has the property that for every $\varepsilon > 0$

$$\lim_{n \to \infty} P_n\{L_{n,2} \in B(\tau^*, \varepsilon) \,|\, H_n/n \in [z - a, z]\} = 1. \tag{5.2}$$

Then as in Theorem 4.1, it will follow that

$$\lim_{n \to \infty} P_n\{X_1 = y_i, X_2 = y_j \,|\, H_n/n \in [z - a, z]\} = \rho_i^{(\beta)} \rho_j^{(\beta)}.$$

Like the analogous limit in part (a) of Theorem 4.1, (5.2) can be proved by showing that the sequence $\{L_{n,2}, n \in \mathbb{N}\}$ satisfies the large deviation principle on $\mathcal{P}_{\alpha,2}$ [38, §I.5] and that the rate function attains its infimum over an appropriately defined, closed convex subset of $\mathcal{P}_{\alpha,2}$ at the unique point τ^* [cf. (4.8)]. The proof of the theorem for general $\ell \in \mathbb{N}$, $\ell \geq 2$, follows from the large deviation principle given in Theorem IX.4.3 in [38]. The details are omitted. ■

The quantity appearing on the right side of (5.1) defines a probability measure $P_{\ell,\beta}$ on Ω_ℓ that equals the product measure with one-dimensional marginals $\rho^{(\beta)}$. In the notation of Theorem 5.1,

$$P_{\ell,\beta}\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell}\} = \prod_{j=1}^{\ell} \rho_{k_j}^{(\beta)}.$$

$P_{\ell,\beta}$ can be written in terms of the total energy $H_\ell(\omega) = \sum_{j=1}^{\ell} \omega_j$: for $\omega \in \Omega_\ell$

$$P_{\ell,\beta}\{\omega\} = \prod_{j=1}^{\ell} \rho^{(\beta)}\{\omega_j\} = \frac{1}{Z_\ell(\beta)} \cdot \exp[-\beta H_\ell(\omega)]\, P_\ell\{\omega\},$$

where $P_\ell\{\omega\} = \prod_{j=1}^{\ell} \rho\{\omega_j\} = 1/\alpha^\ell$,

$$Z_\ell(\beta) = \sum_{\omega \in \Omega_\ell} \exp[-\beta H_\ell(\omega)]\, P_\ell\{\omega\} = \left(\sum_{k=1}^{\alpha} \exp[-\beta y_k]\, \rho_k\right)^\ell,$$

and $\beta = \beta(z) \in \mathbb{R}$ is the unique value of β satisfying $\sum_{k=1}^{\alpha} y_k \rho_k^{(\beta)} = z$.

Theorem 5.1 can be motivated by a non-large deviation calculation that we present using a formal notation [72]. Since $\bar{y} = \sum_{k=1}^{\alpha} y_k \rho_k = E^{P_n}\{X_1\}$, by the weak law of large numbers $P_n\{H_n/n \sim \bar{y}\} \approx 1$ for large n. Since the conditioning is on a set of probability close to 1, one expects that

$$\begin{aligned}
&\lim_{n\to\infty} P_n\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell} \,|\, H_n/n \sim \bar{y}\} \\
&\quad = \lim_{n\to\infty} P_n\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell}\} \\
&\quad = \prod_{j=1}^{\ell} \rho_{k_j} = P_\ell\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell}\}.
\end{aligned}$$

Now take $z \neq \bar{y}$ and for any $\beta > 0$ let $P_{n,\beta}$ denote the product measure on Ω_n with one-dimensional marginals $\rho^{(\beta)}$. A short calculation shows that for any $\beta > 0$

$$\begin{aligned}
&P_n\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell} \,|\, H_n/n \sim z\} \\
&\quad = P_{n,\beta}\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell} \,|\, H_n/n \sim z\}.
\end{aligned}$$

If one picks $\beta = \beta(z)$ such that $z = \sum_{k=1}^{\alpha} y_k \rho_k^{(\beta(z))} = E^{P_{n,\beta(z)}}\{X_1\}$, then by the weak law of large numbers $P_{n,\beta(z)}\{H_n/n \sim z\} \approx 1$, and since the conditioning is on a set of probability close to 1, again one expects that

$$\begin{aligned}
&\lim_{n\to\infty} P_n\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell} \,|\, H_n/n \sim z\} \\
&\quad = \lim_{n\to\infty} P_{n,\beta(z)}\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell} \,|\, H_n/n \sim z\} \\
&\quad = \lim_{n\to\infty} P_{n,\beta(z)}\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell}\} \\
&\quad = \prod_{j=1}^{\ell} \rho_{k_j}^{(\beta(z))} = P_{\ell,\beta(z)}\{X_1 = y_{k_1}, \ldots, X_\ell = y_{k_\ell}\}.
\end{aligned}$$

This is consistent with Theorem 5.1.

For any subset B of Ω_ℓ, (5.1) implies that

$$\lim_{n\to\infty} P_n\{(X_1, \ldots, X_\ell) \in B \,|\, H_n/n \in [z-a, z]\} = P_{\ell,\beta}\{B\}. \tag{5.3}$$

Since $\sum_{\omega\in\Omega_\ell} [H_\ell(\omega)/\ell]\, P_{\ell,\beta}\{\omega\} = \sum_{k=1}^{\alpha} y_k \rho_k^{(\beta)}$, the constraint on $\beta = \beta(z)$ can be expressed as a constraint on $P_{\ell,\beta}$:

$$\text{choose } \beta = \beta(z) \text{ so that } \sum_{\omega\in\Omega_\ell} [H_\ell(\omega)/\ell]\, P_{\ell,\beta}\{\omega\} = z. \tag{5.4}$$

The conditional probability on the left side of (5.3) is known as the microcanonical ensemble, and the probability on the right side of (5.3) as the canonical ensemble. This limit expresses

the equivalence of the two ensembles provided β is chosen in accordance with (5.4). Since the canonical ensemble has a much simpler form than the microcanonical ensemble, one usually prefers to work with the former. The parameter β appearing in the canonical ensemble can be interpreted as being proportional to the inverse temperature. In section 10 we will discuss related issues involving the equivalence of ensembles in a much broader setting, showing that for models in which interactions are present, in general the microcanonical formulation gives rise to a richer set of equilibrium properties than the canonical formulation.

This discussion motivates the definition of the canonical ensemble for a wide class of statistical mechanical models that are defined in terms of an energy function. We will write the energy function, or Hamiltonian, and the corresponding canonical ensemble as H_n and $P_{n,\beta}$ rather than as H_ℓ and $P_{\ell,\beta}$, as we did in the preceding paragraph. The notation of section 2 is used. Thus P_n is the product measure on the set of subsets of $\Omega_n = \Lambda^n$ with one-dimensional marginals ρ. Noninteracting systems such as the discrete ideal gas have Hamiltonians of the form $H_n(\omega) = \sum_{j=1}^n H_{n,j}(\omega_j)$, where each $H_{n,j}$ is a function only of ω_j. In the next definition we do not restrict to this case.

Definition 5.2. *Let H_n be a function mapping Ω_n into $\mathbb{R}$; $H_n(\omega)$ defines the energy of the configuration ω and is known as a Hamiltonian. Let β be a parameter proportional to the inverse temperature. Then the canonical ensemble is the probability measure*

$$P_{n,\beta}\{\omega\} = \frac{1}{Z_n(\beta)} \cdot \exp[-\beta H_n(\omega)]\, P_n\{\omega\} \text{ for } \omega \in \Omega_n,$$

where $Z_n(\beta)$ is the normalization factor that makes $P_{n,\beta}$ a probability measure. That is,

$$Z_n(\beta) = \sum_{\omega \in \Omega_n} \exp[-\beta H_n(\omega)]\, P_n\{\omega\}.$$

We call $Z_n(\beta)$ the partition function. For $B \subset \Omega_n$ we define

$$P_{n,\beta}\{B\} = \sum_{\omega \in B} P_{n,\beta}\{\omega\}.$$

One can also characterize the canonical ensemble in terms of a maximum entropy principle [75, p. 6]. Given $n \in \mathbb{N}$ and a Hamiltonian H_n, let $B_n \subset \mathbb{R}$ denote the smallest closed interval containing the range of $\{H_n(\omega)/n, \omega \in \Omega_n\}$. For each $z \in B_n^\circ$, the interior of B_n, define $C_n(z)$ to be the set of probability measures Q on Ω_n satisfying the energy constraint $\sum_{\omega \in \Omega_n} [H_n(\omega)/n]\, Q\{\omega\} = z$.

Maximum Entropy Principle 5.3. *Let $n \in \mathbb{N}$ and a Hamiltonian $H_n : \Omega_n \mapsto \mathbb{R}$ be given. The following conclusions hold.*

(a) *For each $z \in B_n^\circ$ there exists a unique $\beta = \beta(z) \in \mathbb{R}$ such that $P_{n,\beta} \in C_n(z)$.*

(b) *The relative entropy I_{P_n} attains its infimum over $C_n(z)$ at the unique measure $P_{n,\beta}$, and $I_{P_n}(P_{n,\beta}) = nI_\rho(\rho^\beta)$.*

Part (a) can be proved like part (a) of Theorem 4.1 while part (b) can be proved like Proposition 4.3. We leave the details to the reader.

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 Generalities: Large Deviation Principle and Laplace Principle

In Theorem 3.4 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 9.2 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 9.1, analyzing the phase-transition structure of 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 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 $\{-1, 1\}^{\mathbb{Z}^d}$ [40, §11]. In section 11, our analysis of models of two-dimensional turbulence involves a large deviation principle on the space of probability measures on $T^2 \times \mathcal{Y}$, where T^2 is the unit torus in $\mathbb{R}^2$ and $\mathcal{Y}$ is a compact subset of $\mathbb{R}$.

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 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 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 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 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 $\int_{\mathcal{Y}} f d\Pi_n \rightarrow \int_{\mathcal{Y}} f d\Pi$ 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 $m(\Pi, \Pi_n) \rightarrow 0$ and $\mathcal{P}(\mathcal{Y})$ is a Polish space with respect to m [53, §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 $x_n \rightarrow x$, then $\liminf_{n\rightarrow\infty} I(x_n) \geq I(x)$. 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 6.1 (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 $\mathcal{X}$*

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

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

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

We explore several consequences of this definition. It is reassuring that a large deviation principle has a unique rate function. The following result is proved right before Definition 6.10 as a consequence of Theorem 6.9.

Theorem 6.2. *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. The proof is analogous to the proof of Corollary 3.5.

Theorem 6.3. *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*

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

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 $I(\overline{A}) \le I(A) \le I(A^\circ)$ and

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

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 3.6.

Theorem 6.4. *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 $C < \infty$*

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

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 $I(x_n) \le 1$ 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

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

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. ■

In the next section we will prove Cramér's Theorem, which is the large deviation principle for the sample means of i.i.d. random variables. Here is a statement of the theorem. 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 $\mathbb{R}^d$.

Theorem 6.5 (Cramér's Theorem). *Let $\{X_j, j \in \mathbb{N}\}$ 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 $x_0 = E\{X_1\}$.

For application in subsection 9.1, we next state a special case of Cramér's Theorem, for which the rate function can be given explicitly.

Corollary 6.6. *In the basic probability model of section 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*

$$I(x) = \tfrac{1}{2}(1-x)\log(1-x) + \tfrac{1}{2}(1+x)\log(1+x). \tag{6.1}$$

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 easily verifies 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 (6.1). 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 (6.1). ■

Corollary 6.6 is easy to motivate using the formal notation of Pseudo-Theorem 3.3. For any $x \in [-1, 1]$, $S_n(\omega)/n \sim x$ if and only if approximately $\frac{n}{2}(1-x)$ of the ω_j's equal -1 and approximately $\frac{n}{2}(1+x)$ 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

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

For application in section 11, we 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 $\{X_j, j \in \mathbb{N}\}$ a sequence of i.i.d. random variables mapping Ω into $\mathcal{Y}$ and having the common distribution ρ. For $n \in \mathbb{N}$, $\omega \in \Omega$, 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 6.7 (Sanov's Theorem). *The sequence L_n satisfies the large deviation principle on $\mathcal{P}(\mathcal{Y})$ with rate function given by the relative entropy with respect to ρ. For $\gamma \in \mathcal{P}(\mathcal{Y})$ this quantity is defined by*

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

This theorem is proved, for example, in [25, §6.2] and in [36, Ch. 2]. If the support of ρ is a finite set $\Lambda \subset \mathbb{R}$, then Theorem 6.7 reduces to Theorem 3.4.

The concept of a Laplace principle will be useful in the analysis of statistical mechanical models. As we will see in section 10, where a general class of statistical mechanical models are studied, the Laplace principle gives a variational formula for the canonical free energy [Thm. 10.2(a)].

Definition 6.8 (Laplace 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 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 display 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 in [88]. In fact, we have the following stronger result, which shows that the large deviation principle and the Laplace principle are equivalent.

Theorem 6.9. *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 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

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

As we will see in the proof of Theorem 6.9, showing that the Laplace principle implies the large deviation principle involves approximating φ_A by suitable bounded, continuous functions.

Proof of Theorem 6.9. We first prove that the large deviation principle with rate function I implies the Laplace principle with the same rate function. Let f be a bounded, continuous

function on $\mathcal{X}$. There exists $M \in (0,\infty)$ such that $-M \le f(x) \le M$ for all $x \in \mathcal{X}$. For N a positive integer and $j \in \{1,2,\ldots,N\}$, consider the closed sets

$$F_{N,j} \doteq \left\{ x \in \mathcal{X} : -M + \frac{2(j-1)M}{N} \le f(x) \le -M + \frac{2jM}{N} \right\},$$

whose union over j equals $\mathcal{X}$. The large deviation upper bound yields

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log E^{P_n}\{\exp[nf(Y_n)]\} \\
&\quad \le \limsup_{n\to\infty} \frac{1}{n} \log \left(\sum_{j=1}^{N} \int_{F_{N,j}} \exp[nf(x)]\, P_n\{Y_n \in dx\} \right) \\
&\quad \le \max_{j\in\{1,2,\ldots,N\}} \left\{ -M + \frac{2jM}{N} - I(F_{N,j}) \right\} \\
&\quad \le \max_{j\in\{1,2,\ldots,N\}} \sup_{x\in F_{N,j}} \{f(x) - I(x)\} + \frac{2M}{N} \\
&\quad = \sup_{x\in\mathcal{X}} \{f(x) - I(x)\} + \frac{2M}{N}.
\end{aligned}$$

Sending $N \to \infty$, we obtain upper Laplace limit:

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

In order to prove the corresponding lower Laplace limit, let x be an arbitrary point in $\mathcal{X}$ and ε an arbitrary positive number. We apply the large deviation lower bound to the open set

$$G = \{y \in \mathcal{X} : f(y) > f(x) - \varepsilon\},$$

obtaining

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log E^{P_n}\{\exp[nf(Y_n)]\} &\ge \liminf_{n\to\infty} \frac{1}{n} \log E^{P_n}\{1_G(Y_n)\, \exp[nf(Y_n)]\} \\
&\ge f(x) - \varepsilon + \liminf_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in G\} \\
&\ge f(x) - \varepsilon - I(G) \\
&\ge f(x) - I(x) - \varepsilon.
\end{aligned}$$

Since $x \in \mathcal{X}$ is arbitrary, it follows that

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

Sending $\varepsilon \to 0$ yields the lower Laplace limit. This completes the proof that the large deviation principle with rate function I yields the Laplace principle with the same rate function.

We now prove the converse, first showing that the Laplace principle with rate function I implies the large deviation upper bound. Given a closed set F, let $m(x, F)$ denote the distance from x to F with respect to the metric m on $\mathcal{X}$. For $j \in \mathbb{N}$ define

$$f_j(x) = -j(m(x, F) \wedge 1). \tag{6.2}$$

Then f_j is a bounded, continuous function and

$$f_j \downarrow \varphi_F(x) = \begin{cases} 0 & \text{if } x \in F \\ -\infty & \text{if } x \in F^c. \end{cases}$$

It follows that

$$\frac{1}{n} \log P_n\{Y_n \in F\} = \frac{1}{n} \log E^{P_n}\{\exp[n\varphi_F(Y_n)]\} \leq \frac{1}{n} \log E^{P_n}\{\exp[n\, f_j(Y_n)]\},$$

which by the Laplace principle implies that

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in F\} \leq \lim_{n\to\infty} \frac{1}{n} \log E^{P_n}\{\exp[nf_j(Y_n)]\} = \sup_{x\in\mathcal{X}}\{f_j(x) - I(x)\}.$$

The proof is completed by showing that

$$\lim_{j\to\infty} \sup_{x\in\mathcal{X}}\{f_j(x) - I(x)\} = -I(F). \tag{6.3}$$

Half of this is easy. Since $f_j \geq \varphi_F$,

$$\sup_{x\in\mathcal{X}}\{f_j(x) - I(x)\} \geq \sup_{x\in\mathcal{X}}\{\varphi_F(x) - I(x)\} = -\inf_{x\in\mathcal{X}}\{I(x) - \varphi_F(x)\} = -\inf_{x\in F} I(x) = -I(F),$$

and thus

$$\liminf_{j\to\infty} \sup_{x\in\mathcal{X}}\{f_j(x) - I(x)\} \geq -I(F).$$

The final step in the proof of the large deviation upper bound is to show that

$$\limsup_{j\to\infty} \sup_{x\in\mathcal{X}}\{f_j(x) - I(x)\} \leq -I(F).$$

The proof is given in [36, pp. 10–11].

We complete the proof of the theorem by showing that the Laplace principle with rate function I implies the large deviation lower bound. Let G be an open set. If $I(G) = \infty$, then there is nothing to prove, so we may assume that $I(G) < \infty$. Let x be any point in G such that $I(x) < \infty$ and choose a real number $M > I(x)$. There exists $\delta > 0$ such that $B(x, \delta) \doteq \{y \in \mathcal{X} : m(y, x) < \delta\}$ is a subset of G. In terms of M, x, and δ, we define

$$f(y) \doteq -M\left(\frac{d(y, x)}{\delta} \wedge 1\right). \tag{6.4}$$

This function is bounded and continuous and satisfies $f(x) = 0, f(y) = -M$ for $y \in B(x,\delta)^c$ and $0 \geq f(z) \geq -M$ for all $z \in \mathcal{X}$. We then have

$$\begin{aligned} &E^{P_n}\{\exp[nf(Y_n)]\} \\ &\quad \leq e^{-nM} P_n\{Y_n \in B(x,\delta)^c\} + P_n\{Y_n \in B(x,\delta)\} \leq e^{-nM} + P_n\{Y_n \in B(x,\delta)\}, \end{aligned}$$

and therefore by the Laplace principle

$$\begin{aligned} \max\left(\liminf_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in B(x,\delta)\}, -M\right) &\geq \lim_{n\to\infty} \frac{1}{n} \log E^{P_n}\{\exp[nf(Y_n)]\} \\ &= \sup_{y\in\mathcal{X}}\{f(y) - I(y)\} \\ &\geq f(x) - I(x) \\ &= -I(x). \end{aligned}$$

Since $M > I(x)$ and $B(x,\delta) \subset G$, it follows that

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

and thus

$$\liminf_{n\to\infty} \frac{1}{n} \log P\{Y_n \in G\} \geq -\inf\{I(x) : x \in G, I(x) < \infty\} = -I(G).$$

This proves the large deviation lower bound. The proof of Theorem 6.9 is complete. ■

We next prove Theorem 6.2, which states that in a large deviation principle the rate function is unique.

Proof of Theorem 6.2. By Theorem 6.9 it suffices to prove that if Y_n satisfies the equivalent Laplace principle with rate function I and with rate function J, then $I(x) = J(x)$ for all $x \in \mathcal{X}$. Let ξ be an arbitrary point in $\mathcal{X}$. For $j \in \mathbb{N}$ we evaluate the Laplace principle for the bounded continuous function $f_j(x) = -j(m(x,\xi) \wedge 1)$, then use the limit (6.3) for the closed set $F = \{\xi\}$. We obtain

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

and

$$\lim_{j\to\infty} \lim_{n\to\infty} \frac{1}{n} \log E^{P_n}\{\exp[nf(Y_n)]\} = \lim_{j\to\infty} \sup_{x\in\mathcal{X}}\{f_j(x) - J(x)\} = -J(\xi)$$

Thus $I(\xi) = J(\xi)$, as claimed. ■

The concept of exponential tightness will be used in the proof of Theorem 6.11. This concept is defined next.

Definition 6.10. *The sequence Y_n is said to be exponentially tight if for every $M < \infty$ there exists a compact subset K_M such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in K_M^c\} \leq -M. \tag{6.5}$$

The next theorem shows that if Y_n is exponentially tight, then the large deviation upper bound for all compact sets implies the bound for all closed sets. This is a useful observation because one can often prove the bound for compact sets by covering them with a finite class of sets such as balls or halfspaces for which the proof is easier to obtain. We will see an example of this in the proof of Cramér's Theorem in the next section.

Theorem 6.11. *Assume that Y_n is exponentially tight and that Y_n satisfies the large deviation upper bound for any compact subset of $\mathcal{X}$. Then Y_n satisfies the large deviation upper bound for any closed subset of $\mathcal{X}$.*

Proof. We give the proof under the assumption that F is a closed set for which $I(F) < \infty$, omitting the minor modifications necessary to handle the case in which $I(F) = \infty$. Choose $M < \infty$ such that $M > I(F)$ and let K_M be the compact set satisfying (6.5) in the definition of exponential tightness. Since $F \subset (F \cap K_M) \cup K_M^c$ and $F \cap K_M$ is compact, it follows that

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in F\} \\
&\quad \leq \limsup_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in (F \cap K_M) \cup K_M^c\} \\
&\quad \leq \limsup_{n\to\infty} \frac{1}{n} \log (P_n\{Y_n \in F \cap K_M\} + P_n\{Y_n \in K_M^c\}) \\
&\quad = \max\left\{ \limsup_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in F \cap K_M\}, \limsup_{n\to\infty} \frac{1}{n} \log P_n\{Y_n \in K_M^c\} \right\} \\
&\quad \leq \max\{-I(F \cap K_M), -M\} \\
&\quad \leq \max\{-I(F), -M\} = -I(F).
\end{aligned}$$

This completes the proof. ■

We end this section by presenting three 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. An application involving the relative entropy is given after the statement of the theorem.

Theorem 6.12 (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 $[0,\infty]$. Since ψ is continuous, for any $y \in \mathcal{Y}$ for which $J(y) < \infty$, the set $\psi^{-1}(y)$ is a nonempty, closed subset of $\mathcal{X}$, and since I has compact level sets in $\mathcal{X}$, I attains its infimum over $\psi^{-1}(y)$. It follows that there exists $x \in \mathcal{X}$ such that $\psi(x) = y$ and $I(x) = J(y)$. In order to prove that J has compact level sets in $\mathcal{Y}$, consider any sequence y_n in $\mathcal{Y}$ satisfying $J(y_n) \leq M$ for some $M < \infty$. Then there exists $x_n \in \mathcal{X}$ such that $\psi(x_n) = y_n$ and $I(x_n) = J(y_n) \leq M$. Since I has compact level sets in $\mathcal{X}$, there exists a subsequence $x_{n'}$ and a point $x \in \mathcal{X}$ such that $x_{n'} \to x$ and $I(x) \leq M$. The continuity of ψ and the definition of J imply that

$$y_{n'} = \psi(x_{n'}) \to y = \psi(x) \text{ and } J(y) \leq I(x) \leq M.$$

We conclude that J has compact level sets in $\mathcal{Y}$.

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

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

Similarly, if G is an open subset of $\mathcal{Y}$, then

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

This completes the proof. ■

In Theorem 4.4 we use the contraction principle to derive Cramér's Theorem from Sanov's Theorem for i.i.d. random variables taking values in a finite subset of $\mathbb{R}$. We now use the contraction principle in a different but related way, namely, to relate the rate functions in Sanov's Theorem and in Cramér's Theorem. For simplicity, we restrict to the case of a probability measure on $\mathbb{R}$; much more general versions are available. For example, in [32, Thm. 5.2] it is shown to hold in the case of random variables taking values in a Banach space. Let ρ be a probability measure on $\mathbb{R}$ having compact support K, $\{X_j, j \in \mathbb{N}\}$ a sequence of i.i.d. random variables having common distribution ρ, and $L_n = n^{-1}\sum_{j=1}^n \delta_{X_j}$ the corresponding empirical measures. Since K is compact, the function ψ mapping $\gamma \in \mathcal{P}(K)$ to $\int_K x\gamma(dx)$ is bounded and continuous, and

$$\psi(L_n) = \int_K x\, L_n(dx) = \frac{1}{n}\sum_{j=1}^n \int_K x\, \delta_{X_j}(dx) = \frac{1}{n}\sum_{j=1}^n X_j = \frac{S_n}{n}.$$

Since L_n satisfies the large deviation principle on $\mathcal{P}(\mathcal{K})$ with rate function given by the relative entropy I_ρ [Thm. 6.7], the contraction principle implies that S_n/n satisfies the large deviation on $\mathcal{K}$ with rate function

$$J(y) = \inf\left\{ I_\rho(\gamma) : \gamma \in \mathcal{P}(\mathcal{K}), \int_K x\,\gamma(dx) = y \right\}$$

Since a rate function in a large deviation principle is unique [Thm. 6.2], J must equal the rate function I in Cramér's Theorem [Thm. 6.5]. We conclude that for all $y \in \mathbb{R}$

$$I(y) = \sup_{t\in\mathbb{R}}\{ty - c(t)\} = \inf\left\{ I_\rho(\gamma) : \gamma \in \mathcal{P}(\mathcal{K}), \int_K x\,\gamma(dx) = y \right\}. \tag{6.6}$$

For the basic probability model presented in section 2, a special case of this formula is given in (4.17).

We emphasize that in order to apply the contraction principle, one needs the hypothesis that ρ has compact support. It is satisfying to know that (6.6) is valid without this extra hypothesis [38, Thm. VIII.3.1].

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 9 and 10 when we prove the large deviation principle for statistical mechanical models with respect to the canonical ensemble [Thms. 9.1, 9.3, 9.5, and 10.2].

Theorem 6.13. *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

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

It is easy to motivate this theorem. Using the formal notation

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

and defining

$$Z_n = \int_{\mathcal{X}} \exp[-n\,\psi(Y_n)]\,dP_n,$$

we have

$$\begin{aligned} P_{n,\psi}\{Y_n \in dx\} &= \frac{1}{Z_n} \cdot \exp[-n\psi(x)]\,P_n\{Y_n \in dx\} \\ &\asymp \frac{1}{Z_n} \cdot \exp[-n(\psi(x) + I(x))]\,dx. \end{aligned} \tag{6.7}$$

By the Laplace principle

$$\lim_{n\to\infty}\frac{1}{n}\log Z_n = \sup_{y\in\mathcal{X}}\{-\psi(y)-I(y)\} = -\inf_{y\in\mathcal{X}}\{\psi(y)+I(y)\}.$$

Substituting $Z_n(\beta) \asymp \exp[-n\inf_{y\in\mathcal{X}}\{\psi(y)+I(y)\}]$ into (6.7) yields

$$\begin{aligned} P_{n,\psi}\{Y_n \in dx\} &\asymp \exp\left[-n\left(\psi(x)+I(x)-\inf_{y\in\mathcal{X}}\{I(y)+\psi(y)\}\right)\right]\,dx \\ &= \exp[-nI_\psi(x)]\,dx. \end{aligned}$$

This gives the desired large deviation principle with rate function I_ψ. The constant $\inf_{y\in\mathcal{X}}\{I(y)+\psi(y)\}$ must appear in the formula for the rate function I_ψ in order to guarantee that the infimum of I_ψ over $\mathcal{X}$ equals 0.

Proof of Theorem 6.13. 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 $[0,\infty]$, 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 [Thm. 6.9]. 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

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

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

According to our next result, if random variables X_n are superexponentially close to random variables Y_n that satisfy the large deviation principle, then X_n satisfies the large deviation principle with the same rate function. A proof based on the equivalent Laplace principle is given in Theorem 1.3.3 in [36].

Theorem 6.14. *Assume that Y_n satisfies the large deviation principle on $\mathcal{X}$ with rate function I and denote by $m(\cdot,\cdot)$ the metric on $\mathcal{X}$ (e.g., $m(x,y)=|x-y|$ if $\mathcal{X}=\mathbb{R}$). Assume also that X_n is superexponentially close to Y_n in the following sense: for each $\delta>0$*

$$\limsup_{n\to\infty}\frac{1}{n}\log P_n\{m(Y_n,X_n)>\delta\} = -\infty. \tag{6.8}$$

Then X_n satisfies the large deviation principle on $\mathcal{X}$ with the same rate function I. The condition (6.8) *is satisfied if*

$$\lim_{n\to\infty} \sup_{\omega\in\Omega_n} m(X_n(\omega), Y_n(\omega)) = 0.$$

This completes our discussion of the large deviation principle, the Laplace principle, and related general results. In the next section we present several basic results in the theory of convex functions and then prove Cramér's Theorem.

7 Cramér's Theorem

Cramér's Theorem is the large deviation principle for sums of i.i.d. random vectors taking values in $\mathbb{R}^d$. A special case of Cramér's Theorem was proved in Theorem 4.4 for i.i.d. random variables taking values in a finite subset of $\mathbb{R}$. In this section Cramér's Theorem is proved for i.i.d. random vectors taking values in $\mathbb{R}^d$. We then give an application and state an infinite dimensional version.

Let $\{X_j, j \in \mathbb{N}\}$ be a sequence of i.i.d. random vectors defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in $\mathbb{R}^d$. We are interested in the large deviation principle for the sample means S_n/n, where $S_n = \sum_{j=1}^n X_j$. The basic assumption is that the moment generating function $E\{\exp\langle t, X_1\rangle\}$ is finite for all $t \in \mathbb{R}^d$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. We define for $t \in \mathbb{R}^d$ the cumulant generating function

$$c(t) = \log E\{\exp\langle t, X_1\rangle\}, \tag{7.1}$$

which is finite, convex, and differentiable, and for $x \in \mathbb{R}^d$ we define the Legendre-Fenchel transform

$$I(x) = \sup_{t \in \mathbb{R}^d} \{\langle t, x\rangle - c(t)\}. \tag{7.2}$$

The function $c(t)$ is finite for all $t \in \mathbb{R}^d$, and the differentiability follows from the dominated convergence theorem. The convexity of c can be proved either by a calculation similar to that used to prove part (a) of Lemma 4.2 or by applying Hölder's inequality with $p = 1/\lambda$ and $q = 1/(1-\lambda)$ [38, Prop. VII.1.1].

The basic theory of convex functions and the Legendre-Fenchel transform is developed in chapter VI of [38]. Here are some relevant definitions. A subset D of $\mathbb{R}^d$ is said to be convex if for all x and y in D and all $\lambda \in (0,1)$, we have $\lambda x + (1-\lambda)y \in D$. A function f mapping $\mathbb{R}^d$ into $\mathbb{R} \cup \{\infty\}$ is said to be convex on $\mathbb{R}^d$ if $f(x) < \infty$ for some $x \in \mathbb{R}^d$ and for all x and y in $\mathbb{R}^d$ and all $\lambda \in (0,1)$

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

Given a function f mapping $\mathbb{R}^d$ into $\mathbb{R} \cup \{\infty\}$, the domain of f, written dom f, is defined to be the set of $x \in \mathbb{R}^d$ for which $f(x) < \infty$. One easily proves that if f is a convex function on $\mathbb{R}^d$, then dom f is a convex subset of $\mathbb{R}^d$. In addition, a function f mapping $\mathbb{R}^d$ into $\mathbb{R} \cup \{\infty\}$ is said to be lower semicontinuous if whenever $x_n \to x \in \mathbb{R}$, we have $\liminf_{n\to\infty} f(x_n) \geq f(x)$.

We continue by introducing several more definitions for a convex function f on $\mathbb{R}^d$. If y is a point in $\mathbb{R}^d$, then $z \in \mathbb{R}^d$ is said to be a subgradient of f at y if

$$f(x) \geq f(y) + \langle z, x - y\rangle \text{ for all } x \in \mathbb{R}^d.$$

Thus, if $d = 1$, then a subgradient z is the slope of a supporting line to the graph of f at y while if $d \geq 2$, then a subgradient z has the property that $[z, -1]$ is the slope of a supporting hyperplane

to the graph of f at y, where $[z, -1]$ denotes the vector in $\mathbb{R}^{d+1}$ whose first d coordinates agree with those of z and whose last component equals -1. We then define the subdifferential of f at y to be the set

$$\partial f(y) = \{z \in \mathbb{R}^d : z \text{ is a subgradient of } f \text{ at } y\}.$$

The following basic fact, proved in [79, p. 242], will be used in the proof of Cramer's Theorem.

Theorem 7.1. *Let f be a convex function on $\mathbb{R}^d$. Then f is differentiable at y if and only if f has a unique subgradient at y. If unique, this subgradient is $\nabla f(y)$.*

Given f a convex function on $\mathbb{R}^d$, the Legendre-Fenchel transform of f is defined for $y \in \mathbb{R}^d$ by

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle x, y\rangle - f(x)\}.$$

In particular, the function I defined in (7.2) equals c^*. In general, f^* is convex and lower semicontinuous on $\mathbb{R}^d$ [38, Lem. VI.5.2]. It follows that the mapping of f to f^* maps the class of convex functions into itself. We now ask whether this map is onto; i.e., given f convex on $\mathbb{R}^d$, whether there exists a convex function g on $\mathbb{R}^d$ such that $f = g^*$. The answer is no because if $f = g^*$, then f must be lower semicontinuous. Hence, if f is not lower semicontinuous, then the representation $f = g^*$ cannot hold. As part (e) of the next theorem shows, the mapping of f to f^* is one-to-one and onto as a mapping of the class of functions that are both convex and lower semicontinuous on $\mathbb{R}^d$; in fact, if f is convex and lower semicontinuous, then $f = (f^*)^* = f^{**}$.

In order to prove part (a), we need a definition. Let D be a convex subset of $\mathbb{R}^d$. The relative interior of D of $\mathbb{R}^d$, denoted by $\text{ri}(D)$, is defined as the interior of D when considered as a subset of the smallest affine set that contains D. Clearly, if the smallest affine set that contains D is $\mathbb{R}^d$, then the relative interior of D equals the interior of D. This is the case if, for example, $d = 1$ and D is a nonempty interval. If D is a single point, then the relative interior of D equals D.

Theorem 7.2. *Let f be a convex, lower semicontinuous function on $\mathbb{R}^d$. Then the following conclusions hold.*

(a) *f^* is convex and lower semicontinuous on $\mathbb{R}^d$.*

(b) *$f^*(y) \geq \langle x, y\rangle - f(x)$ for all x and y in $\mathbb{R}^d$.*

(c) *$f^*(y) = \langle x, y\rangle - f(x)$ if and only if $y \in \partial f(x)$.*

(d) *$y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$.*

(e) *$f = f^{**}$; i.e., for all $x \in \mathbb{R}^d$*

$$f(x) = \sup_{y \in \mathbb{R}^d} \{\langle x, y\rangle - f^*(y)\}.$$

Proof. (a) Since f is a convex function, $\mathrm{dom}\, f$ is a convex subset of $\mathbb{R}^d$. We need the fact, proved in [79, Thm. 23.4], that for any $x_0 \in \mathrm{ri}(\mathrm{dom}\, f)$, $\partial f(x_0)$ is nonempty. For any $y \in \partial f(x_0)$, we have $f(x) \geq f(x_0) + \langle y, x - x_0 \rangle$ for all x, which implies that

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - f(x)\} \leq \langle x_0, y \rangle - f(x_0) < \infty.$$

In addition, since $f(x) < \infty$ for some $x \in \mathbb{R}^d$, for all $y \in \mathbb{R}^d$ we have $f^*(y) \geq \langle x, y \rangle - f(x) > -\infty$. It follows that f^* maps $\mathbb{R}^d$ into $\mathbb{R} \cup \{\infty\}$ and that $f^*(y) < \infty$ for some y.

To prove that f^* is convex, take any y_1 and y_2 in $\mathbb{R}^d$ and any $\lambda \in (0, 1)$. Writing $\alpha_x(y)$ for the affine function $\langle x, y \rangle - f(x)$, we have

$$\begin{aligned} f^*(\lambda y_1 + (1-\lambda)y_2) &= \sup_{x \in \mathbb{R}^d} \{\lambda \alpha_x(y_1) + (1-\lambda)\alpha_x(y_2)\} \\ &\leq \lambda \sup_{x \in \mathbb{R}^d} \alpha_x(y_1) + (1-\lambda) \sup_{x \in \mathbb{R}^d} \alpha_x(y_2) \\ &= \lambda f^*(y_1) + (1-\lambda) f^*(y_2). \end{aligned}$$

To prove that f^* is lower semicontinuous, let y_n be a sequence converging to some y. Then for any $x \in \mathbb{R}^d$

$$\liminf_{n \to \infty} f^*(y_n) \geq \liminf_{n \to \infty} (\langle x, y_n \rangle - f(x)) = \langle x, y \rangle - f(x).$$

Since x is arbitrary, we conclude that

$$\liminf_{n \to \infty} f^*(y_n) \geq \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - f(x)\} = f^*(y).$$

This completes the proof of part (a).

(b) This is an immediate consequence of the definition of f^*.

(c) The subgradient inequality defining the condition $y \in \partial f(x)$ can be written as

$$\langle y, x \rangle - f(x) \geq \langle y, w \rangle - f(w) \text{ for all } w \in \mathbb{R}^d.$$

This is the same as saying that the function $\langle y, w \rangle - f(w)$ attains its supremum at $w = x$. Since by definition this supremum is $f^*(y)$, we have proved part (c).

(d) By part (c) applied to f and f^{**}, the equation $\langle y, x \rangle = f^*(y) + f^{**}(x)$ holds if and only if $x \in \partial f^*(y)$ for some $y \in \mathbb{R}^d$. By part (e), this equation is the same as $\langle y, x \rangle = f^*(y) + f(x)$, and again by part (c) this holds if and only if $y \in \partial f(x)$. Hence $y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$.

(e) For any x and y in $\mathbb{R}^d$, $f(x) \geq \langle x, y \rangle - f^*(y)$. This implies that

$$f(x) \geq \sup_{y \in \mathbb{R}^d} \{\langle x, y \rangle - f^*(y)\} = f^{**}(x).$$

Let x_0 be any point in $\mathbb{R}^d$. We prove that $f^{**}(x_0) \geq f(x_0)$ using the fact, proved in [38, Lem. VI.5.5], that if α_0 is any real number satisfying $\alpha_0 < f(x_0)$, then there exists $y_0 \in \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$

$$\langle y_0, x\rangle - f(x) \leq \langle y_0, x_0\rangle - \alpha_0. \tag{7.3}$$

This inequality implies that $f^*(y_0) \leq \langle y_0, x_0\rangle - \alpha_0$ or that

$$\alpha_0 \leq \langle y_0, x_0\rangle - f^*(y_0) \leq \sup_{y\in\mathbb{R}^d}\{\langle y, x_0\rangle - f^*(y)\} = f^{**}(x_0).$$

Letting α_0 converge to $f(x_0)$, we conclude that $f(x_0) \leq f^{**}(x_0)$. The proof of the theorem is complete.

The proof that there exists $y_0 \in \mathbb{R}^d$ such that (7.3) holds for all $x \in \mathbb{R}^d$ is the most difficult step in the proof of the theorem. In order to understand the geometric content of this inequality, we define the epigraph of f to be the set of $(x, \alpha) \in \mathbb{R}^d \times \mathbb{R}$ satisfying $\alpha \geq f(x)$; i.e., the set of all (x, α) lying on or above the graph of f. Since f is convex and lower semicontinuous, the epigraph of f is a closed convex set in $\mathbb{R}^{d+1}$. The content of (7.3) is that if a point (x_0, α_0) does not belong to the epigraph of f, then it can be separated from the epigraph of f by a non-vertical hyperplane in $\mathbb{R}^{d+1}$. ■

We use the theory of convex functions to prove Cramér's Theorem, which we first consider in the case $d = 1$. Let α be a real number exceeding the mean value $E\{X_1\}$. Assuming that ρ has an absolutely continuous component and that certain other conditions hold, Cramér obtained in his 1938 paper [22] an asymptotic expansion for the probability $P\{S_n/n \in [\alpha, \infty)\}$, which implies the large deviation limit

$$\lim_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in [\alpha, \infty)\} = -I(\alpha) = -I([\alpha, \infty)).$$

In the modern theory of large deviations the following generalization of this limit is known as Cramér's Theorem.

Theorem 7.3 (Cramér's Theorem). *Let $\{X_j.j \in \mathbb{N}\}$ 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 $t \in \mathbb{R}^d$. Define $c(t) = \log E\{\exp\langle t, X_1\rangle\}$ for $t \in \mathbb{R}^d$. 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 $x_0 = E\{X_1\}$.

Infinite-dimensional generalizations of Cramér's Theorem have been proved by many authors, including [1] and [32, §5]. The book [25] presents Cramér's Theorem first in the setting of $\mathbb{R}^d$ and then in the setting of a complete, separable metric space. At the end of this section we will derive from Cramér's Theorem the large deviation principle for the empirical vectors stated

in Theorem 3.4. This is a special case of Sanov's Theorem 6.7. We will also indicate how to prove a general version of Sanov's Theorem from an infinite-dimensional version of Cramér's Theorem.

Before proving Cramér's Theorem, it is worthwhile to motivate the form of the rate function I. Assuming that the sequence S_n/n satisfies the large deviation principle on $\mathbb{R}^d$ with some convex, lower semicontinuous rate function J, we prove that $J = I$. Since for each $t \in \mathbb{R}^d$

$$\begin{aligned} c(t) &= \log E\{\exp\langle t, X_1\rangle\} = \frac{1}{n}\log E\{\exp\langle t, S_n\rangle\} \\ &= \frac{1}{n}\log \int_{\mathbb{R}^d} \exp[n\langle t, x\rangle]\, P\{S_n/n \in dx\}, \end{aligned}$$

it follows that

$$c(t) = \lim_{n\to\infty} \frac{1}{n}\log \int_{\mathbb{R}^d} \exp[n\langle t, x\rangle]\, P\{S_n/n \in dx\}.$$

We now use the hypothesis that S_n/n satisfies the large deviation principle on $\mathbb{R}^d$ with some convex, lower semicontinuous rate function J. Although the function mapping $x \mapsto \langle t, x\rangle$ is not bounded, a straightforward extension of Theorem 6.9 allows us to apply the Laplace principle to evaluate the last limit, yielding

$$c(t) = \sup_{x\in\mathbb{R}^d}\{\langle t, x\rangle - J(x)\} = J^*(t).$$

Because J is convex and lower semicontinuous, part (e) of Theorem 7.2 implies that $c^* = J^{**} = J$. Since by definition $I = c^*$, we obtain the desired formula; namely, for each $x \in \mathbb{R}^d$

$$J(x) = \sup_{t\in\mathbb{R}^d}\{\langle t, x\rangle - c(t)\} = I(x).$$

This completes the motivation of the form of the rate function in Cramér's Theorem.

We now turn to the proof of Cramér's Theorem. The main tool used in the proof of the large deviation upper bound is Chebyshev's inequality, introduced by Chernoff in [15], while the main tool used in the proof of the large deviation lower bound is a change of measure, introduced by Cramér in his 1938 paper [22]. These same tools for proving the large deviation bounds in Cramér's Theorem continue to be used in modern developments of the theory.

Proof of Theorem 7.3. We first show that I is a rate function, then prove part (b) followed by the proofs of the large deviation upper bound and lower bound.

I is a rate function. Since I is defined as a Legendre-Fenchel transform, it is automatically convex and lower semicontinuous. By part (a) of Theorem 6.4, the infimum of I over $\mathbb{R}^d$ equals 0, and so I maps $\mathbb{R}^d$ into the extended nonnegative real numbers $[0, \infty]$. We now consider a level set $K_L = \{x \in \mathbb{R}^d : I(x) \le L\}$, where L is any nonnegative real number. This set is closed since I is lower semicontinuous. If x is in K_L, then for any $t \in \mathbb{R}^d$

$$\langle t, x\rangle \le c(t) + I(x) \le c(t) + L.$$

Fix any positive number R. The finite, convex, continuous function c is bounded on the ball of radius R with center 0, and so there exists a number $\Gamma < \infty$ such that

$$\sup_{\|t\| \le R} \langle t, x \rangle = R\|x\| \le \sup_{\|t\| \le R} c(t) + L \le \Gamma < \infty.$$

This implies that K_L is bounded and thus that the level sets of I are compact. The sketch of the proof that I is a rate function is complete.

Part (b). We have already remarked that I is convex and lower semicontinuous. Since I is a rate function, I attains its infimum of 0 at some point $x_0 \in \mathbb{R}^d$ [Thm. 6.4(a)]. We next show that I attains its infimum of 0 at the unique point $E\{X_1\}$. By the definition of the subdifferential, I attains its infimum at x_0 if and only if 0 is in $\partial I(x_0)$. By parts (d) and (c) of Theorem 7.2, 0 is in $\partial I(x_0)$ if and only if x_0 is in $\partial c(0)$, and for any $x_0 \in \partial c(0)$, $I(x_0) = \langle 0, x_0 \rangle - c(0) = 0$. Since c is differentiable, $\partial c(0) = \{\nabla c(0)\} = \{E\{X_1\}\}$ [Thm. 7.1]. Hence it follows that I attains its infimum of 0 at the unique point $E\{X_1\}$.

Large deviation upper bound. We first prove this bound in the case $d = 1$. Our aim is to prove that for any nonempty closed subset F of $\mathbb{R}$

$$\limsup_{n \to \infty} \frac{1}{n} \log P\{S_n/n \in F\} \le -I(F).$$

Let $x_0 = E\{X_1\}$. We first show this for the closed intervals $[\alpha, \infty)$, where $\alpha > x_0$. For any $t > 0$ Chebyshev's inequality implies

$$\begin{aligned} P\{S_n/n \in [\alpha, \infty)\} &= P\{tS_n \ge nt\alpha\} \\ &\le \exp[-nt\alpha]\, E\{\exp[tS_n]\} \\ &= \exp[-nt\alpha] \prod_{i=1}^{n} E\{\exp[tX_i]\} \\ &= \exp[-nt\alpha]\ (E\{\exp[tX_1]\})^n \\ &= \exp[-nt\alpha + n \log E\{\exp[tX_1]\}] \\ &= \exp[-n(t\alpha - c(t))]. \end{aligned}$$

It follows that for any $t > 0$

$$\limsup_{n \to \infty} \frac{1}{n} \log P\{S_n/n \in [\alpha, \infty)\} \le -(t\alpha - c(t)),$$

and thus that

$$\limsup_{n \to \infty} \frac{1}{n} \log P\{S_n/n \in [\alpha, \infty)\} \le -\sup_{t>0} \{t\alpha - c(t)\}. \tag{7.4}$$

The next lemma will allow us to rewrite the right-hand side of this inequality as in the statement of Cramér's Theorem.

Lemma 7.4. *If $\alpha > x_0$, then*

$$\sup_{t>0}\{t\alpha - c(t)\} = I(\alpha) = I([\alpha, \infty).$$

Proof. Since $c(t)$ is continuous at $t = 0$,

$$I(\alpha) = \sup_{t\in\mathbb{R}}\{t\alpha - c(t)\} = \sup_{t\neq 0}\{t\alpha - c(t)\}.$$

Since c is differentiable and convex, we have $c'(0) \geq c(t)/t$ for any $t < 0$. Therefore, for any $t < 0$

$$t\alpha - c(t) = t(\alpha - c(t)/t) \leq t(\alpha - c'(0)) < 0 = 0 \cdot \alpha - c(0).$$

The second inequality holds since $\alpha > x_0 = E\{X_1\} = c'(0)$ and $t < 0$. From this display we see that the supremum in the formula for $I(\alpha)$ cannot occur for $t < 0$. It follows that

$$I(\alpha) = \sup_{t>0}\{t\alpha - c(t)\}.$$

$I(x)$ is nonnegative, convex function satisfying $I(x_0) = 0$. Thus $I(x)$ is nonincreasing for $x \leq x_0$ and is nondecreasing for $x \geq x_0$. This implies that $I(\alpha) = \inf\{I(x) : x \geq \alpha\} = I([\alpha, \infty)$. The proof of the lemma is complete. ■

Inequality (7.4) and the lemma imply that if F is the closed interval $[\alpha, \infty)$, then the large deviation upper bound holds:

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in F\} \leq -I(\alpha) = -I(F).$$

A similar proof yields the large deviation upper bound if $F = (-\infty, \alpha]$ and $\alpha < x_0$.

Now let F be an arbitrary nonempty closed set in $\mathbb{R}$. If $x_0 \in F$, then $I(F)$ equals 0 and the large deviation upper bound holds automatically since $\log P\{S_n/n \in F\}$ is always nonpositive. If $x_0 \notin F$, then let (α_1, α_2) be the largest open interval containing x_0 and having empty intersection with F. F is a subset of $(-\infty, \alpha_1] \cup [\alpha_2, \infty)$, and by the first part of the proof

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in F\} \\
&\quad \leq \limsup_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in (-\infty, \alpha_1] \cup [\alpha_2, \infty)\} \\
&\quad \leq \limsup_{n\to\infty} \frac{1}{n} \log(P\{S_n/n \in [-\infty, \alpha_1]\} + P\{S_n/n \in [\alpha_2, \infty)\}) \\
&\quad = \max\left\{\limsup_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in [-\infty, \alpha_1]\}, \limsup_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in [\alpha_2, \infty)\}\right\} \\
&\quad \leq \max\{-I(\alpha_1), -I(\alpha_2)\} \\
&\quad = -\min\{(I(\alpha_1), I(\alpha_2)\}\,.
\end{aligned}$$

If $\alpha_1 = -\infty$ or $\alpha_2 = \infty$, then the corresponding term is missing. $I(x)$ is nonnegative, convex function satisfying $I(x_0) = 0$. Thus $I(x)$ is nonincreasing for $x \leq x_0$ and is nondecreasing for $x \geq x_0$. It follows that $I(F) = \min\{(I(\alpha_1), I(\alpha_2)\}$. Hence from the last display we conclude that

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in F\} \leq -I(F).$$

This completes the proof of the large deviation upper bound for $d = 1$.

We now prove the large deviation upper bound for $d > 1$. Using the hypothesis that $c(t) < \infty$ for all $t \in \mathbb{R}^d$, it is straightforward to prove that S_n/n is exponentially tight and that the compact set K_M appearing in the definition 6.10 of exponential tightness can be taken to be the hypercube $K_M = [-b, b]^d$, where b depends on M. The details are omitted. By Theorem 6.11 the upper bound will follow for any closed set if we can prove it for any compact set K. If $I(K) = 0$, then the upper bound holds automatically since $\log P_n\{S_n/n \in K\}$ is always nonpositive. Details will now be given under the assumption that $I(K) < \infty$. The minor modifications necessary to prove the upper bound when $I(K) = \infty$ are omitted.

The technique of the proof exhibits a remarkable interplay among analysis, geometry, and probability and readily extends to the much more general setting of the Gärtner-Ellis Theorem, which we will consider in the next section [Thm. 8.1]. We start by picking $\varepsilon > 0$ to satisfy $\varepsilon < I(K)$ and by defining for $t \in \mathbb{R}^d$ the open halfspace

$$H_t = \{x \in \mathbb{R}^d : \langle t, x\rangle - c(t) > I(K) - \varepsilon\};$$

if $t = 0$, then $H_0 = \emptyset$ since $I(K) - \varepsilon > 0$. Since for all $x \in K$ we have $I(x) > I(K) - \varepsilon$, it follows that

$$\begin{aligned} K &\subset \{x \in \mathbb{R}^d : I(x) > I(K) - \varepsilon\} \\ &= \left\{x \in \mathbb{R}^d : \sup_{t\in\mathbb{R}^d}\{\langle t, x\rangle - c(t)\} > I(K) - \varepsilon\right\} \\ &= \bigcup_{t\in\mathbb{R}^d} \{x \in \mathbb{R}^d : \langle t, x\rangle - c(t) > I(K) - \varepsilon\} \\ &= \bigcup_{t\in\mathbb{R}^d} H_t. \end{aligned}$$

Since K is compact, there exists $r \in \mathbb{N}$ and nonzero $t_1, \ldots, t_r \in \mathbb{R}^d$ such that $K \subset \cup_{i=1}^r H_{t_i}$.

Thus by Chebyshev's inequality

$$\begin{aligned} P\{S_n/n \in K\} &\leq \sum_{i=1}^{r} P\{S_n/n \in H_{t_i}\} \\ &= \sum_{i=1}^{r} P\{\langle t_i, S_n\rangle > n[c(t_i) + I(K) - \varepsilon]\} \\ &\leq \sum_{i=1}^{r} \exp[-n(c(t_i) + I(K) - \varepsilon)]\, E\{\exp[\langle t_i, S_n\rangle]\} \\ &= \sum_{i=1}^{r} \exp[-n(c(t_i) + I(K) - \varepsilon)]\, \exp[n\, c(t_i)] \\ &= r \exp[-n[I(K) - \varepsilon], \end{aligned} \tag{7.5}$$

from which it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in K\} \leq -I(K) + \varepsilon.$$

Sending $\varepsilon \to 0$ completes the proof of the large deviation upper bound for $d > 1$. This argument generalizes the proof for $d = 1$, in which we covered an arbitrary closed set F by the intervals $(-\infty, \alpha_1] \cup [\alpha_2, \infty)$.

Large deviation lower bound. In contrast to the large deviation upper bound, which is proved by a global estimate involving Chebyshev's inequality, the large deviation lower bound is proved by a local estimate, the heart of which involves a change of measure. The proof is somewhat more technical than the proof of the large deviation upper bound. We denote the common distribution of the random vectors X_j by $\rho(dx) = P\{X_j \in dx\}$. In general

$$c(t) = \log E\{\exp\langle t, X_1\rangle\} = \log \int_{\mathbb{R}^d} \exp\langle t, x\rangle\, \rho(dx)$$

is a finite, convex, differentiable function on $\mathbb{R}^d$. We first prove the lower bound under the highly restrictive assumption that the support of ρ is all of $\mathbb{R}^d$ or more generally that the smallest convex set containing the support of ρ is all of $\mathbb{R}^d$. In this case, for each $z \in \mathbb{R}^d$ there exists $t \in \mathbb{R}^d$ such that $\nabla c(t) = z$ [38, Thms. VIII.3.3, VIII.4.3].

For $z \in \mathbb{R}^d$ and $\varepsilon > 0$, we denote by $B(z, \varepsilon)$ the open ball with center z and radius ε. Let G be an open subset of $\mathbb{R}^d$. Then for any point $z_0 \in G$ there exists $\varepsilon > 0$ such that $B(z_0, \varepsilon) \subset G$, and so

$$P\{S_n/n \in G\} \geq P\{S_n/n \in B(z_0, \varepsilon)\} = \int_{\{\sum_{j=1}^n x_j/n \in B(z_0,\varepsilon)\}} \prod_{j=1}^{n} \rho(dx_j).$$

We first assume that G contains the point $\int_{\mathbb{R}^d} x\rho(dx) = E\{X_1\}$ and let $z_0 = \int_{\mathbb{R}^d} x\rho(dx)$. In this case the weak law of large numbers implies that

$$\lim_{n\to\infty} \int_{\{\sum_{j=1}^n x_j/n \in B(z_0,\varepsilon)\}} \prod_{j=1}^n \rho(dx_j) = 1.$$

Since $I(z_0) = 0$, we obtain the large deviation lower bound:

$$\liminf_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in G\} \geq 0 = -I(z_0) = -I(G).$$

Of course, in general G does not contain the point $\int_{\mathbb{R}^d} x\rho(dx)$, and the argument in the preceding paragraph breaks down. In this case we let z_0 be an arbitrary point in G and introduce a change of measure, replacing ρ by a new measure ρ_{t_0} whose mean equals z_0. The exponential price that must be paid for introducing this new measure is of the order of $\exp[-nI(z_0)]$. Putting the various estimates together will yield the desired large deviation lower bound.

Given $z_0 \in G$, we choose $t_0 \in \mathbb{R}^d$ such that $\nabla c(t_0) = z_0$. We then introduce the change of measure given by the exponential family

$$\rho_{t_0}(dx) = \frac{1}{\int_{\mathbb{R}^d} e^{\langle t_0,x\rangle}\, \rho(dx)} \cdot e^{\langle t_0,x\rangle}\, \rho(dx) = \frac{1}{e^{c(t_0)}} \cdot e^{\langle t_0,x\rangle}\, \rho(dx).$$

Similar exponential families arise in Theorems 4.1 and 5.1. By the definition of $c(t_0)$, ρ_{t_0} is a probability measure, and the mean of ρ_{t_0} is z_0. Indeed

$$\int_{\mathbb{R}^d} x\, \rho_{t_0}(dx) = \frac{1}{e^{c(t_0)}} \cdot \int_{\mathbb{R}^d} x\, e^{\langle t_0,x\rangle}\, \rho(dx) = \nabla c(t_0) = z_0.$$

Furthermore, since $c(t)$ is finite, convex, and continuous on $\mathbb{R}^d$, part (c) of Theorem 7.2 implies that

$$I(z_0) = \sup_{t\in\mathbb{R}^d} \{\langle t, z_0\rangle - c(t)\} = \langle t_0, z_0\rangle - c(t_0).$$

We thus obtain the lower bound

$$\begin{aligned}
&P_n\{S_n/n \in G\} \\
&\quad \geq P_n\{S_n/n \in B(z_0,\varepsilon)\} \\
&\quad = \int_{\{\sum_{j=1}^n x_j/n \in B(z_0,\varepsilon)\}} \prod_{j=1}^n \rho(dx_j) \\
&\quad = \int_{\{\sum_{j=1}^n x_j/n \in B(z_0,\varepsilon)\}} \left(\prod_{j=1}^n \frac{d\rho}{d\rho_{t_0}}(x_j)\right) \prod_{j=1}^n \rho_{t_0}(dx_j) \\
&\quad = \int_{\{\sum_{j=1}^n x_j/n \in B(z_0,\varepsilon)\}} \exp\Big[-n\left(\Big\langle t_0, \textstyle\sum_{j=1}^n x_j/n\Big\rangle - c(t_0)\right)\Big] \prod_{j=1}^n \rho_{t_0}(dx_j) \\
&\quad \geq \exp[-n(\langle t_0, z_0\rangle - c(t_0)) - n\|t_0\|\varepsilon] \int_{\{\sum_{j=1}^n x_j/n \in B(z_0,\varepsilon)\}} \prod_{j=1}^n \rho_{t_0}(dx_j) \\
&\quad = \exp[-nI(z_0) - n\|t_0\|\varepsilon] \int_{\{\sum_{j=1}^n x_j/n \in B(z_0,\varepsilon)\}} \prod_{j=1}^n \rho_{t_0}(dx_j).
\end{aligned}$$

Since the mean of the probability measure ρ_{t_0} equals z_0, the weak law of large numbers for i.i.d. random vectors with common distribution ρ_{t_0} implies that

$$\lim_{n\to\infty} \int_{\{\sum_{j=1}^n x_j/n \in B(z_0,\varepsilon)\}} \prod_{j=1}^n \rho_{t_0}(dx_j) = 1.$$

Hence it follows that

$$\liminf_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in G\} \geq -I(z_0) - \|t_0\|\varepsilon.$$

We now send $\varepsilon \to 0$, and since z_0 is an arbitrary point in G, we can replace $-I(z_0)$ by $-\inf_{z_0 \in G} I(z_0) = -I(G)$. This completes the proof of the large deviation lower bound when the support of ρ is all of $\mathbb{R}^d$ or more generally when the smallest convex set containing the support of ρ is all of $\mathbb{R}^d$.

When this hypothesis does not hold, then the range of $\nabla c(t)$ for $t \in \mathbb{R}^d$ is no longer all of $\mathbb{R}^d$, and the argument just given breaks down. To handle the case of general ρ, we find a set A with the properties that $I(G) = I(G \cap A)$ and that A is a subset of the range of $\nabla c(t)$. If we can do this, then the proof just given, specialized to arbitrary $z_0 \in G \cap A$, yields

$$\liminf_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in G\} \geq -I(z_0).$$

Since z_0 is an arbitrary point in $G \cap A$, we can replace $-I(z_0)$ by

$$-\inf_{z_0 \in G\cap A} I(z_0) = -I(G \cap A) = -I(G),$$

and we are done.

By definition, the domain of I, dom I, is the set of $x \in \mathbb{R}^d$ for which $I(x) < \infty$ and the relative interior of dom I, ri(dom I), is the interior of dom I when considered as a subset of the smallest affine set that contains dom I. Using several properties of convex sets and convex functions, we will show that the desired set A equals ri(dom I). In order to see this, we first note that since $I(x)$ equals ∞ for $x \notin$ dom I, $I(G)$ equals ∞ if $G \cap$ dom I is empty. In this case the large deviation lower bound is valid. If $G \cap$ dom I is nonempty, then $I(G)$ equals $I(G \cap \text{dom } I)$. The set $G \cap$ ri(dom I) is also nonempty [79, p, 46], and by the continuity property of I expressed in [38, Thm. VI.3.2]

$$I(G) = I(G \cap \text{dom } I) = I(G \cap \text{ri}(\text{dom } I)).$$

This is the first required property of $A = \text{ri}(\text{dom } I)$. The second desired property of this set — namely, that ri(dom I) is a subset of the range of $\nabla c(t)$ for $t \in \mathbb{R}^d$ — is a consequence of [38, Thm. VI.5.7], which is based on additional duality properties involving $c(t)$ and its Legendre-Fenchel transform $I(x)$ [79]. This completes the proof of the large deviation lower bound and thus the proof of Cramér's Theorem. ■

As we have just seen, a key step in the proof of Cramér's Theorem is the fact that ri(dom I) is a subset of the range of $\nabla c(t)$ for $t \in \mathbb{R}^d$. This fact can be directly verified in the case of the basic probability model introduced in section 2. Using the notation introduced there,

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

According to part (d) of Lemma 4.2, the range of $c'(t)$ for $t \in \mathbb{R}$ is the open interval (y_1, y_α), which can be shown to equal the interior of the domain of I [38, Thm. VIII.3.3]. It follows that in this case the interior of the domain of I, and hence the relative interior, coincides with the range of $c'(t)$.

We now apply Cramér's Theorem to derive the special case of Sanov's Theorem given in Theorem 3.4. The latter states the large deviation principle for the empirical vectors of i.i.d. random variables having a finite state space. Let $\alpha \geq 2$ be an integer; $y_1 < y_2 < \ldots < y_\alpha$ a set of α real numbers; $\rho_1, \rho_2, \ldots, \rho_\alpha$ a set of α positive real numbers summing to 1; and $\{X_j, j \in \mathbb{N}\}$ a sequence of i.i.d. random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, taking values in $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$, and having distribution $\rho = \sum_{k=1}^{\alpha} \rho_k \delta_{y_k}$. As in Theorem 3.4, we may take $\{X_j, j = 1, \ldots, n\}$ to be the coordinate functions on Λ^n and impose on this space the product measure P_n with one dimensional marginals ρ. For $\omega \in \Omega$ and $y \in \Lambda$ we consider

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

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

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

Here is a restatement of the special case of Sanov's Theorem first given in Theorem 3.4.

Theorem 7.5. *The sequence of empirical vectors L_n satisfies the large deviation principle on $\mathcal{P}_\alpha$ with rate function the relative entropy*

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

Since L_n equals the sample mean of the i.i.d. random variables

$$Y_j(\omega) = (\delta_{X_j(\omega)}\{y_1\}, \ldots, \delta_{X_j(\omega)}\{y_\alpha\}),$$

Theorem 7.5 should follow from Cramér's Theorem, as we now verify. According to Cramér's Theorem, for $\gamma \in \mathbb{R}^\alpha$ the rate function is given by

$$I(\gamma) = \sup_{t \in \mathbb{R}^\alpha}\{\langle \gamma, t\rangle - c(t)\}, \text{ where } c(t) = \log E\{\exp\langle t, Y_1\rangle\} = \log\left(\sum_{k=1}^{\alpha} e^{t_k}\rho_k\right).$$

In the next proposition we show that for $\gamma \in \mathcal{P}_\alpha$, $I(\gamma)$ equals the relative entropy $I_\rho(\gamma)$ and that for $\gamma \in \mathbb{R}^\alpha \setminus \mathcal{P}_\alpha$, $I(\gamma)$ equals ∞.

Proposition 7.6. *For $\gamma \in \mathcal{P}_\alpha$*

$$I(\gamma) = \sup_{t \in \mathbb{R}^\alpha}\{\langle \gamma, t\rangle - \log(\textstyle\sum_{k=1}^{\alpha} e^{t_k}\rho_k)\} = \begin{cases} I_\rho(\gamma) & \text{for } \gamma \in \mathcal{P}_\alpha \\ \infty & \text{for } \gamma \in \mathbb{R}^\alpha \setminus \mathcal{P}_\alpha. \end{cases}$$

Sketch of Proof. Let G be any open set having empty intersection with $\mathcal{P}_\alpha$. Since $P\{L_n \in G\} = 0$, the large deviation lower bound implies that

$$-\infty = \liminf_{n\to\infty} \frac{1}{n} \log P\{L_n \in G\} \geq -I(G).$$

It follows that $I(\gamma) = \infty$ for $\gamma \in \mathbb{R}^\alpha \setminus \mathcal{P}_\alpha$. The exercise of proving this directly from the definition of $I(\gamma)$ as a Legendre-Fenchel transform is left to the reader.

Defining $\mathcal{P}_\alpha^\circ$ to be the set of $\gamma \in \mathcal{P}_\alpha$ having all positive components, we next prove that $I(\gamma) = I_\rho(\gamma)$ for $\gamma \in \mathcal{P}_\alpha^\circ$. Let $\mathbb{R}_+^\alpha$ denote the positive orthant of $\mathbb{R}^\alpha$. Since $-\log$ is strictly convex on $(0, \infty)$, Jensen's inequality implies that for any $\gamma \in \mathcal{P}_\alpha^\circ$

$$\sup_{s \in \mathbb{R}_+^\alpha} \left\{ \sum_{k=1}^\alpha \gamma_k \log s_k - \log \sum_{k=1}^\alpha \gamma_k s_k \right\} \leq 0$$

with equality if and only if $s_k = \text{const}$. For $\gamma \in \mathcal{P}_\alpha^\circ$, as t runs through $\mathbb{R}^\alpha$, the vector s having components $s_k = e^{t_k}\rho_k/\gamma_k$ runs through $\mathbb{R}_+^\alpha$. Hence

$$\begin{aligned}
I(\gamma) &= \sup_{t \in \mathbb{R}^\alpha} \left\{ \langle \gamma, t \rangle - \log \left(\sum_{k=1}^\alpha e^{t_k} \rho_k \right) \right\} \\
&= \sup_{s \in \mathbb{R}_+^\alpha} \left\{ \sum_{k=1}^\alpha \gamma_k \log \frac{\gamma_k s_k}{\rho_k} \; - \; \log \left(\sum_{k=1}^\alpha \gamma_k s_k \right) \right\} \\
&= \sum_{k=1}^\alpha \gamma_k \log \frac{\gamma_k}{\rho_k} \; + \; \sup_{s \in \mathbb{R}_+^\alpha} \left\{ \sum_{k=1}^\alpha \gamma_k \log s_k - \log \sum_{k=1}^\alpha \gamma_k s_k \right\} \\
&= \sum_{k=1}^\alpha \gamma_k \log \frac{\gamma_k}{\rho_k} \\
&= I_\rho(\gamma).
\end{aligned}$$

This completes the proof that $I(\gamma) = I_\rho(\gamma)$ for $\gamma \in \mathcal{P}_\alpha^\circ$. In order to prove this equality for all $\gamma \in \mathcal{P}_\alpha$, we use the continuity of I_ρ on $\mathcal{P}_\alpha$ and the continuity property of I on $\mathcal{P}_\alpha$ stated in [38, Thm. VI.3.2]. The proof of the proposition is complete. ■

In Theorem 5.2 in [32] the following infinite dimensional version of Cramér's Theorem is proved.

Theorem 7.7. *Let $\mathcal{X}$ be a Banach space with dual space $\mathcal{X}^*$ and $\{X_j, j \in \mathbb{N}\}$ a sequence of i.i.d. random vectores taking values in $\mathcal{X}$ and having common distribution ρ. Assume that $E\{\exp(t\|X_1\|)\} < \infty$ for every $t > 0$. Then the sequence of sample means S_n/n satisfies the large deviation principle on $\mathcal{X}$ with rate function*

$$I(x) = \sup_{\theta \in \mathcal{X}^*} \left\{ \langle \theta, x \rangle - \log \int_{\mathcal{X}} \exp\langle \theta, y \rangle \, \rho(dy) \right\}.$$

The rate function I is convex and lower semicontinuous and attains its infimum of 0 at the unique point $x_0 = E\{X_1\} = \int_{\mathcal{X}} x\rho(dx)$.

We now return to the setting of Sanov's Theorem, considering the empirical measures L_n of a sequence $\{X_j, j \in \mathbb{N}\}$ of i.i.d. random variables taking values in $\mathbb{R}^d$ [Thm. 6.7] and more

generally in a complete, separable metric space $\mathcal{Y}$. Let ρ denote the common distribution of X_j. Then

$$L_n = \frac{1}{n} \sum_{j=1}^{n} \delta_{X_j}$$

takes values in the complete, separable metric space $\mathcal{P}(\mathcal{Y})$ of probability measures on $\mathcal{Y}$. Since L_n is the sample mean of the i.i.d. random variables δ_{X_j}, it is reasonable to conjecture that Sanov's Theorem can be derived as a consequence of a suitable infinite-dimensional version of Cramér's Theorem. While Theorem 7.7 cannot be applied because $\mathcal{P}(\mathcal{Y})$ is not a Banach space, the derivation of Sanov's Theorem from a suitable infinite-dimensional version of Cramér's Theorem is carried out in [26, Thm. 3.2.17]. This reference first proves that L_n satisfies the large deviation principle on $\mathcal{P}(\mathcal{Y})$ with rate function $I(\gamma)$ given by the Legendre-Fenchel transform

$$I(\gamma) = \sup_{f \in \mathcal{C}(\mathcal{Y})} \left\{ \int_{\mathcal{Y}} f \, d\gamma - \log \int_{\mathcal{Y}} e^f d\rho \right\},$$

where $\mathcal{C}(\mathcal{Y})$ denotes the set of bounded, continuous functions mapping $\mathcal{Y}$ into $\mathbb{R}$. The proof of Sanov's Theorem is completed by showing that $I(\gamma)$ equals the relative entropy $I_\rho(\gamma)$. A special case of this identification of the relative entropy with a Legendre-Fenchel transform is given in Proposition 7.6. An independent derivation of Sanov's Theorem for i.i.d. random vectors taking values in a complete, separable space is given in [36, Ch. 2], which applies ideas from stochastic optimal control theory.

In the next section we present a generalization of Cramér's Theorem that does not require the underlying random variables to be independent. Both in Cramér's Theorem and in this generalization the rate functions are defined by Legendre-Fenchel transforms and so are always convex. This convexity is not a general feature. Indeed, at the end of the next section we present two examples of large deviation principles in which the rate function is not convex.

8 Gärtner-Ellis Theorem

For each $n \in \mathbb{N}$ let $(\Omega_n, \mathcal{F}_n, P_n)$ be a probability space and let Y_n be a random vector mapping Ω_n into $\mathbb{R}^d$. In 1977 Gärtner proved an important generalization of Cramer's Theorem, assuming only that the limit

$$c(t) = \lim_{n\to\infty} \frac{1}{n} \log E^{P_n}\{\exp[n\langle t, Y_n\rangle]\} \tag{8.1}$$

exists and is finite for every $t \in \mathbb{R}^d$ and that $c(t)$ is a differentiable function of $t \in \mathbb{R}^d$ [58]. Gärtner's result is that Y_n satisfies the large deviation principle on $\mathbb{R}^d$ with rate function equal to the Legendre-Fenchel transform

$$I(x) = \sup_{t\in\mathbb{R}^d}\{\langle t, x\rangle - c(t)\}.$$

Using ideas from convex analysis, I generalized Gärtner's result by relaxing the condition that $c(t)$ exist and be finite for every $t \in \mathbb{R}^d$ [39]. The theorem is now known in the literature as the Gärtner-Ellis Theorem [25, §2.3, §2.5].

Gärtner's result contains Cramér's Theorem as a special case. In order to see this, let Y_n equal $\sum_{j=1}^n X_j/n$, where X_j is a sequence of i.i.d. random vectors satisfying $E\{\exp\langle t, X_1\rangle\} < \infty$ for every $t \in \mathbb{R}^d$. In this case the limit $c(t)$ in (8.1) equals $\log E\{\exp\langle t, X_1\rangle\}$, which is a differentiable function of $t \in \mathbb{R}^d$. The corresponding rate function is the same as in Cramer's Theorem.

We next state the Gärtner-Ellis Theorem under the hypotheses of [58] and in a form that is different from but equivalent to Gärtner's result in that paper. This is followed by comments on the generalization proved in [39]. In this theorem the differentiability of $c(t)$ for all $t \in \mathbb{R}^d$ is a sufficient condition for the large deviation lower bound; the large deviation upper bound is always valid. However, as we mention just before Example 8.3 in the context of the Ising model in statistical mechanics, the differentiability of $c(t)$ is not a necessary condition for the validity of the lower bound.

Theorem 8.1. *For each $n \in \mathbb{N}$ let $(\Omega_n, \mathcal{F}_n, P_n)$ be a probability space and let Y_n be a random vector mapping Ω_n into $\mathbb{R}^d$. We assume that the limit*

$$c(t) = \lim_{n\to\infty} \frac{1}{n} \log E^{P_n}\{\exp[n\langle t, Y_n\rangle]\}$$

exists and is finite for every $t \in \mathbb{R}^d$. For $x \in \mathbb{R}^d$ we define

$$I(x) = \sup_{t\in\mathbb{R}^d}\{\langle t, x\rangle - c(t)\}.$$

The following conclusions hold.

(a) *I is a rate function. Furthermore, I is convex and lower semicontinuous.*

(b) *The large deviation upper bound is valid. Namely, for every closed subset F of $\mathbb{R}^d$*

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

(c) *Assume in addition that $c(t)$ is differentiable for all $t \in \mathbb{R}^d$. Then the large deviation lower bound is valid. Namely, for every open subset G of $\mathbb{R}^d$*

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

Hence, if $c(t)$ is differentiable for all $t \in \mathbb{R}^d$, then Y_n satisfies the large deviation principle on $\mathbb{R}^d$ with rate function I.

The theorem is proved by suitably generalizing the proof of Cramér's Theorem (see [38, Ch. 7]). In the case of the large deviation upper bound, the generalization is easy to see. As in the proof of Cramér's Theorem, the assumption that the limit function $c(t)$ is finite for every $t \in \mathbb{R}^d$ implies that Y_n is exponentially tight. Hence by Theorem 6.11, the upper bound will follow for any closed set if we can prove it for any compact set K. If $I(K) = 0$, then the upper bound holds automatically since $\log P_n\{Y_n \in K\}$ is always nonpositive. In order to handle the case when $I(K) < \infty$, we argue as in the proof of Cramér's Theorem. Given $\varepsilon > 0$ satisfying $\varepsilon < I(K)$, there exists $r \in \mathbb{N}$ and nonzero $t_1, \ldots, t_r \in \mathbb{R}^d$ such that $K \subset \cup_{i=1}^r H_{t_i}$, where H_{t_i} denotes the open halfspace

$$H_{t_i} = \{x \in \mathbb{R}^d : \langle t_i, x\rangle - c(t_i) > I(K) - \varepsilon\}.$$

As in the display (7.5), Chebyshev's inequality yields

$$\begin{aligned} P\{Y_n \in K\} &\leq \sum_{i=1}^r P\{Y_n \in H_{t_i}\} \\ &= \sum_{i=1}^r P\{n\langle t_i, Y_n\rangle > n[c(t_i) + I(K) - \varepsilon]\} \\ &\leq \sum_{i=1}^r \exp[-n(c(t_i) + I(K) - \varepsilon)]\, E\{\exp[n\langle t_i, Y_n\rangle]\} \\ &= \sum_{i=1}^r \exp[-n(c(t_i) + I(K) - \varepsilon)]\, \exp[n\, c_n(t_i)], \end{aligned}$$

where

$$c_n(t) = \frac{1}{n} \log E^{P_n}\{\exp[n\langle t, Y_n\rangle]\}.$$

Since $c_n(t_i) \to c(t_i)$, there exists $N \in \mathbb{N}$ such that $c_n(t_i) < c(t_i) + \varepsilon$ for all $n \geq N$ and all $i = 1, \ldots, r$. Thus for all $n \geq N$

$$P_n\{Y_n \in K\} \leq r \exp[-n(I(K) - 2\varepsilon)],$$

from which it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \log P\{Y_n \in K\} \leq -I(K) + 2\varepsilon.$$

Sending $\varepsilon \to 0$, we complete the proof of the large deviation upper bound in the Gärtner-Ellis Theorem when $I(K) < \infty$. The minor modifications necessary to prove the upper bound when $I(K) = \infty$ are omitted.

The proof of the large deviation lower bound requires a new idea. We recall that the proof of the lower bound in Cramér's Theorem invoked the weak law of large numbers with respect to the change of measure given by the product measure with one-dimensional marginals ρ_{t_0}. In the proof of the lower bound in the Gärtner-Ellis Theorem again one uses a change of measure, but the weak law of large numbers with respect to a product measure is not available. The innovation is to replace the weak law of large numbers by an order-1 estimate based on the large deviation upper bound in the Gärtner-Ellis Theorem. Details are given in [38, §VII.3].

In the extension of Gärtner's result that I proved in [39], it is assumed that for all $t \in \mathbb{R}^d$, $c(t)$ exists as an extended real number in $(-\infty, \infty]$. Then the large deviation upper bound as stated in part (b) of Theorem 8.1 is valid with rate function

$$I(x) = \sup_{t\in\mathbb{R}^d} \{\langle t, x\rangle - c(t)\}.$$

We denote by $\mathcal{D}$ the set of $t \in \mathbb{R}^d$ for which $c(t)$ is finite. If c is differentiable on the interior of $\mathcal{D}$, then c is called steep if $\|\nabla c(t_n)\| \to \infty$ for any sequence t_n in the interior of $\mathcal{D}$ that converges to a boundary point of $\mathcal{D}$. For example, if c is lower semicontinuous, $\mathcal{D}$ is open, and c is differentiable on $\mathcal{D}$, then c is steep. In the extension of Gärtner's result, it is proved that if c is differentiable on the interior of $\mathcal{D}$ and is steep, then the large deviation lower bound as stated in part (c) of Theorem 8.1 is valid with rate function

$$I(x) = \sup_{t\in\mathbb{R}^d} \{\langle t, x\rangle - c(t)\}.$$

We next give an application of the Gärtner-Ellis Theorem to finite-state Markov chains. Let $\alpha \geq 2$ be an integer, $y_1 < y_2 < \ldots < y_\alpha$ be a set of α real numbers, and $\{X_j, j \in \mathbb{N}\}$ a Markov chain taking values in $\Lambda = \{y_1, y_2, \ldots, y_\alpha\}$. We denote by $\rho \in \mathbb{R}^\alpha$ the initial distribution $\rho_k = P\{X_1 = y_k\}$ and by $\pi(k, \ell)$ the transition probabilities $P\{X_{j+1} = y_\ell | X_j = y_k\}$ for $j \in \mathbb{N}$ and $1 \leq k, \ell \leq \alpha$. Under the assumption that the matrix $\pi = \{\pi(k, \ell)\}$ is irreducible and aperiodic, we have the following two-part theorem. Part (a) is the large deviation principle for the sample means $Y_n = \sum_{j=1}^{\alpha} X_j/n$, and part (b) is the large deviation principle for the empirical vectors $L_n = \sum_{j=1}^{\alpha} \delta_{X_j}/n$. The hypothesis that π is irreducible holds if, for example, π is a positive matrix. Part (b) is a special case of a result proved in [31].

Theorem 8.2. *We assume that the transition probability matrix π of the Markov chain X_j is aperiodic and irreducible. The following conclusions hold.*

(a) *For $t \in \mathbb{R}$ let $B(t)$ be the matrix with entries $[B(t)]_{k,\ell} = \exp(tx_k)\pi(k,\ell)$. Then for all $t \in \mathbb{R}$, $B(t)$ has a unique largest positive eigenvalue $\lambda(t)$ which is differentiable for all $t \in \mathbb{R}$. Furthermore, for any choice of the initial distribution ρ, the sample means Y_n satisfy the large deviation principle on $\mathbb{R}$ with rate function*

$$I(x) = \sup_{t \in \mathbb{R}} \{tx - \log \lambda(t)\}.$$

(b) *We denote by $\mathcal{P}_\alpha$ the set of probability vectors in $\mathbb{R}^\alpha$. Then for any choice of the initial distribution ρ, the empirical vectors L_n satisfy the large deviation principle on $\mathcal{P}_\alpha$ with rate function*

$$I_\pi(\gamma) = -\inf_{u>0} \sum_{k=1}^{\alpha} \gamma_k \log \frac{(\pi u)_k}{u_k}.$$

In this formula u is any positive vector in $\mathbb{R}^\alpha$, and $(\pi u)_k = \sum_{\ell=1}^{\alpha} \pi(k,\ell) u_\ell$.

Sketch of Proof. (a) That $B(t)$ has a unique largest positive eigenvalue $\lambda(t)$ is a consequence of the Perron-Frobenius Theorem [80, Thm. 1.1]. The differentiability of $\lambda(t)$ is a consequence of the implicit function theorem and the fact that $\lambda(t)$ is a simple root of the characteristic equation for $B(t)$. For $t \in \mathbb{R}$ we calculate

$$\begin{aligned}
c(t) &= \lim_{n\to\infty} \frac{1}{n} \log E\{\exp[ntY_n]\} \\
&= \lim_{n\to\infty} \frac{1}{n} \log E\{\exp\langle t \textstyle\sum_{j=1}^n X_j \rangle\} \\
&= \lim_{n\to\infty} \frac{1}{n} \log \sum_{k_1,k_2,\ldots,k_n=1}^{\alpha} \exp(t \textstyle\sum_{j=1}^n x_{k_j})\, \rho_{k_1}\, \pi(k_1,k_2)\, \pi(k_2,k_3) \cdots \pi(k_{n-1},k_n) \\
&= \lim_{n\to\infty} \frac{1}{n} \log \sum_{k_1,k_2,\ldots,k_n=1}^{\alpha} \rho_{k_1}\, (B(t))_{k_1,k_2}\, (B(t))_{k_2,k_3} \cdots (B(t))_{k_{n-1},k_n}\, e^{tx_{k_n}} \\
&= \lim_{n\to\infty} \frac{1}{n} \log \sum_{k_1,k_n=1}^{\alpha} \rho_{k_1} [B(t)^{n-1}]_{k_1,k_n}\, e^{tx_{k_n}}.
\end{aligned}$$

Using a standard limit theorem for irreducible, aperiodic Markov chains [56, p. 356], one proves that for each $1 \le k, \ell \le \alpha$

$$\lim_{n\to\infty} \frac{1}{n} \log [B(t)^n]_{k,\ell} = \log \lambda(t).$$

Details are given in [38, Lem. IX.4.1]. It follows that $c(t) = \log \lambda(t)$. Since $\lambda(t)$ and thus $c(t)$ are differentiable for all t, the Gärtner-Ellis Theorem implies that Y_n satisfies the large deviation principle on $\mathbb{R}$ with the indicated rate function I.

(b) We refer the reader to [39, Thm. III.1], where this large deviation principle for the empirical vectors L_n is proved. The basic idea is that as in part (a) the rate function is given by a Legendre-Fenchel transform in $\mathbb{R}^\alpha$. One then shows that this Legendre-Fenchel transform equals I_π on $\mathcal{P}_\alpha$ and equals ∞ on $\mathbb{R}^\alpha \setminus \mathcal{P}_\alpha$. ■

We end this section by examining several features of the Gärtner-Ellis Theorem. Since in that theorem the rate function is always convex, a natural question is whether there exist large deviation principles having nonconvex rate functions. Two such examples are given next. Additional examples appear in [29].

One of the hypotheses of the Gärtner-Ellis Theorem is the differentiability of the limit function $c(t)$. An interesting problem is to investigate the validity of the large deviation principle when this condition is violated. Unfortunately the situation is complicated, and a general theory has not yet been discovered. In Example 8.3 the differentiability of the limit function $c(t)$ does not hold, and the rate function is not given by a Legendre-Fenchel transform. In another example arising in the Ising model in statistical mechanics, the same hypothesis of the Gärtner-Ellis Theorem is not valid for all sufficiently large values of the inverse temperature defining the model. However, the rate function in the large deviation principle for the spin per site is defined by the identical Legendre-Fenchel transform appearing in the statement of the Gärtner-Ellis Theorem [40, Thm. 11.1].

The first example involves an extreme case of dependent random variables.

Example 8.3. We define a random varible X_1 by the probability distribution $P\{X_1 = 1\} = P\{X_1 = -1\} = \frac{1}{2}$. For each integer $j \geq 2$ we define random variables $X_j = X_1$, and for $n \in \mathbb{N}$ we set

$$Y_n = \frac{1}{n} \sum_{j=1}^{n} X_j.$$

Let us first try to apply the Gärtner-Ellis Theorem to the sequence Y_n. For each $t \in \mathbb{R}$ and $x \in \mathbb{R}$ we calculate

$$c(t) = \lim_{n \to \infty} \frac{1}{n} \log E\{\exp(ntY_n)\} = \lim_{n \to \infty} \frac{1}{n} \log \left(\frac{1}{2} \left[e^{nt} + e^{-nt} \right] \right) = |t|$$

and

$$I(x) = \sup_{t \in \mathbb{R}} \{tx - c(t)\} = \begin{cases} 0 & \text{if } |x| \leq 1 \\ \infty & \text{if } |x| > 1. \end{cases}$$

Since $c(t) = |t|$ is not differentiable at t = 0, the Gärtner-Ellis Theorem is not applicable. In fact, Y_n satisfies the large deviation principle on $\mathbb{R}$ with the rate function

$$J(x) = \begin{cases} 0 & \text{if } x \in \{1, -1\} \\ \infty & \text{if } x \in \mathbb{R} \setminus \{1, -1\}. \end{cases}$$

This is easily checked since W_n has the distribution $P\{W_n = 1\} = P\{W_n = -1\} = \frac{1}{2}$. The function I is the largest convex function less than or equal to the rate function J, and as one easily verifies, $I = J^{**}$. This completes the first example. ■

The second example generalizes Cramer's Theorem to the setting of a random walk with an interface.

Example 8.4. We define the sets

$$\Lambda^{(1)} = \{x \in \mathbb{R}^d : x_1 \leq 0\},\ \Lambda^{(2)} = \{x \in \mathbb{R}^d : x_1 > 0\},\ \Gamma = \{x \in \mathbb{R}^d : x_1 = 0\},$$

where x_1 denotes the first component of $x \in \mathbb{R}^d$. We define a random walk model for which the distribution of the next step depends on the halfspace $\Lambda^{(1)}$ or $\Lambda^{(2)}$ in which the random walk is currently located. To this end let $\rho^{(1)}$ and $\rho^{(2)}$ be two distinct probability measures on $\mathbb{R}^d$. Although it is not necessary, for simplicity we assume that the support of each measure is all of $\mathbb{R}^d$. Let $\{X_j^{(1)}, j \in \mathbb{N}\}$ and $\{X_j^{(2)}, j \in \mathbb{N}\}$ be independent sequences of i.i.d. random vectors with probability distributions $P\{X_j^{(1)} \in dx\} = \rho^{(1)}(dx)$ and $P\{X_j^{(2)} \in dx\} = \rho^{(2)}(dx)$. We consider the stochastic process $\{S_n, n \in \mathbb{N} \cup \{0\}\}$, where $S_0 = 0$ and S_{n+1} is defined recursively from S_n by the formula

$$S_{n+1} = S_n + 1_{\{S_n \in \Lambda^{(1)}\}} \cdot X_n^{(1)} + 1_{\{S_n \in \Lambda^{(2)}\}} \cdot X_n^{(2)}.$$

For $i = 1, 2$, $1_{\{S_n \in \Lambda^{(i)}\}}$ denotes the indicator function of the set $\{S_n \in \Lambda^{(i)}\}$. Because of the abrupt change in distribution across the surface Γ, we call this random walk a model with discontinuous statistics. In [34] we show that S_n/n satisfies the large deviation principle on $\mathbb{R}^d$. The rate function is given by an explicit formula that takes a complicated form along the interface Γ. We will not give the definition of the rate function here, but merely note that in general it is a nonconvex function on $\mathbb{R}^d$ which is convex in each of the halfspaces $\Lambda^{(1)}$ and $\Lambda^{(2)}$. If the measures $\rho^{(l)}$ and $\rho^{(2)}$ coincide, then the main theorem of [34] reduces to Cramér's Theorem.

The large deviation phenomena investigated in [34] are an example of the fascinating problems that arise in the study of other Markov processes with discontinuous statistics. The main theorem of [34] is generalized in [36, Ch. 6] to a large deviation principle for the entire path of the random walk. In [37] a large deviation upper bound is proved for a general class of Markov processes with discontinuous statistics. An important group of processes with discontinuous statistics arises in the study of queueing systems. The large deviation principle for a general class of such systems is proved in [35]. This completes the second example. ■

In the next section we begin our study of statistical mechanical models by considering the Curie-Weiss spin model and other mean-field models.

9 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.

9.1 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 [38, §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

$$H_n(\omega) = -\frac{1}{2n} \sum_{i,j=1}^n \omega_i \omega_j = -\frac{n}{2} \left(\frac{S_n(\omega)}{n} \right)^2, \tag{9.1}$$

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

$$\begin{aligned} P_{n,\beta}\{\omega\} &= \frac{1}{Z_n(\beta)} \cdot \exp[-\beta H_n(\omega)]\, P_n\{\omega\} \\ &= \frac{1}{Z_n(\beta)} \cdot \exp\left[\frac{n\beta}{2} \left(\frac{S_n(\omega)}{n} \right)^2 \right] P_n\{\omega\}, \end{aligned} \tag{9.2}$$

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 (9.4)]. 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.

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 6.6, with respect to the product measures P_n, S_n/n satisfies the large deviation principle on $[-1,1]$ with rate function

$$I(x) = \tfrac{1}{2}(1-x)\log(1-x) + \tfrac{1}{2}(1+x)\log(1+x). \tag{9.3}$$

Because of the form of $P_{n,\beta}$ given in (9.2), the large deviation principle for S_n/n with respect to $P_{n,\beta}$ is an immediate consequence of Theorem 6.13 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 9.1. *With respect to the canonical ensemble $P_{n,\beta}$ defined in* (9.2)*, 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 (9.3)*.*

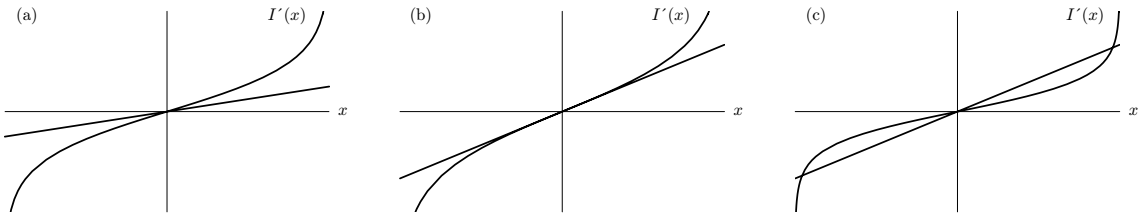

Figure 1: Solutions of $I'(x^*) = \beta x^*$: (a) $\beta < 1$, (b) $\beta = 1$, (c) $\beta > 1$

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 $\inf_{y\in[-1,1]}\{I(y) - \frac{1}{2}\beta y^2\}$ [38, §IV.4]. Global minimum points x^* satisfy

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

The second equation is equivalent to the mean field equation $x^* = (I')^{-1}(\beta x^*) = \tanh(\beta x^*)$ [38, §V.9], [75, §3.2]. Figure 1 motivates the next theorem, which is a consequence of the following easily verified properties of I:

- $I''(0) = 1$.
- I' is convex on $[0, 1]$ and $\lim_{x\to 1} I'(x) = \infty$.
- I' is concave on $[-1, 0]$ and $\lim_{x\to -1} I'(x) = -\infty$.

Theorem 9.2. *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 $\beta \to 1^+$, $m(\beta) \to 1$ as $\beta \to \infty$.*

According to part (b) of Theorem 6.4, if A is any closed subset of $[-1, 1]$ such that $A \cap \mathcal{E}_\beta = \emptyset$, then $I_\beta(A) > 0$ and for some $C < \infty$

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

In combination with Theorem 9.2, we are led to the following weak limits:

$$P_{n,\beta}\{S_n/n \in dx\} \Longrightarrow \begin{cases} \delta_0 & \text{if } 0 < \beta \le 1 \\ \frac{1}{2}\delta_{m(\beta)} + \frac{1}{2}\delta_{-m(\beta)} & \text{if } \beta > 1. \end{cases} \tag{9.4}$$

We call $m(\beta)$ the spontaneous magnetization for the Curie-Weiss model and $\beta_c = 1$ the critical inverse temperature [38, §IV.4]. It is worth remarking that it is much easier to derive the weak limits in (9.4) 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 (9.4) 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 9.2. For $0 < \beta \leq 1$, 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 2 global minimum points at $\pm m(\beta)$, a structure that is consistent with the facts that $-\frac{1}{2}\beta x^2$ has 2 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 $\beta \leq 1$ to $\{\pm\sqrt{3(\beta-1)}\}$ for $\beta > 1$. This behavior is consistent with the continuous phase transition described in Theorem 9.2 and suggests that as $\beta \to 1^+$, $m(\beta) \sim \sqrt{3(\beta-1)} \to 0$.

Before leaving 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 (9.1) 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 6.6 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 the physical importance of $\mathcal{E}_\beta$. This set is the support of the $n \to \infty$ limit of the distributions $P_{n,\beta}\{S_n/n \in dx\}$. As we will see in the next section, an analogous fact is true for a large class of statistical mechanical models [Thm. 10.3].

As shown in (9.4), the large deviation analysis of the Curie-Weiss model yields the limiting behavior of the $P_{n,\beta}$-distributions of S_n/n. For $0 < \beta \leq 1$ 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-Wiess, referring the reader to [38, §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 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 [38, Thm. V.9.4]:

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

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 [38, Thm. V.9.5]

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

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 [48] with $k = 1$.

The results discussed in this subsection have been extensively generalized to a number of models, including the Curie-Weiss-Potts model [18, 51], the mean-field Blume-Capel model [17, 49], and the Ising and related models [30, 57, 73], 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 [25, 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 9.2.

9.2 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 is 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 [90]. 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 $\rho = \frac{1}{q}\sum_{i=1}^{q} \delta_{y_i}$ 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 inverse temperature $\beta > 0$ is defined by the canonical ensemble

$$P_{n,\beta}\{\omega\} = \frac{1}{Z_n(\beta)} \cdot \exp[-\beta H_n(\omega)]\, P_n\{\omega\}, \tag{9.5}$$

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

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

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 k^{th} 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

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

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 3.4, with respect to the product measures P_n, L_n satisfies the large deviation principle on $\mathcal{P}_\alpha$ with rate function the relative entropy I_ρ. Because of the form of $P_{n,\beta}$ given in the last display, the large deviation principle for L_n with respect to $P_{n,\beta}$ is an immediate consequence of Theorem 6.13 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 9.3. *With respect to the canonical ensemble $P_{n,\beta}$ defined in* (9.5)*, 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 k^{th} 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 $\beta = \beta_c$ 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 9.4. *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 k^{th} 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 [51] 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 [51, 52]. Having completed our discussion of the Curie-Weiss-Potts model, we next consider a third mean-field model that exhibits different features.

9.3 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 [4, 12, 13, 14]. This mean-field model is one of the simplest models that exhibit 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 [5].

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 inverse temperature $\beta > 0$ and 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, obtaining

$$P_{n,\beta,K}\{\omega\} = \frac{1}{\widetilde{Z}_n(\beta, K)} \cdot \exp\left[n\beta K\left(\frac{S_n(\omega)}{n}\right)^2\right] (P_\beta)_n\{\omega\}. \tag{9.6}$$

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

$$\rho_\beta\{\omega_j\} = \frac{1}{Z(\beta)} \cdot \exp(-\beta\omega_j^2)\, \rho\{\omega_j\}, \tag{9.7}$$

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

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

Comparing (9.6) with (9.2), 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 9.1. By Cramér's Theorem [Thm. 7.3], with respect to the product measures $(P_\beta)_n$, S_n/n satisfies the large deviation principle with rate function

$$J_\beta(x) = \sup_{t\in\mathbb{R}}\{tx - c_\beta(t)\}, \tag{9.8}$$

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 6.13 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 9.5. *With respect to the canonical ensemble $P_{n,\beta,K}$ defined in* (9.6)*, 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 (9.8)*.*

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 (9.3). 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 in [49], 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 9.6 *For $0 < \beta \leq \beta_c$, we define*

$$K(\beta) = 1/[2\beta c''_\beta(0)] = (e^\beta + 2)/(4\beta). \tag{9.9}$$

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

(a) *For $0 < K \leq K(\beta)$, $\mathcal{E}_{\beta,K} = \{0\}$.*

(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 in [49], 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 9.7 *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* [49] *and defined implicitly for $\beta > \beta_c$ on page* 2231 *of* [49].

(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. The sets that describe the phase-transition structure of the model are shown in Figure 2.

The mean-field Blume-Capel model admits another sequence of macroscopic variables. As we show in [49, §2], the Hamiltonian can also be written in terms of a sequence of empirical vectors. Applying Sanov's Theorem with respect to the product measure P_n, we obtain a large deviation principle for the empirical vectors with respect to the canonical ensemble $P_{n,\beta,K}$. The zeroes of the rate function of that large deviation principle are the set of canonical equilibrium macrostates for the empirical vectors in the mean-field Blume-Capel model. This set of equilibrium macrostates exhibits the same phase transition structure as the set of canonical equilibrium macrostates for the spin per site in Theorems 9.6 and 9.7 [49, Thms. 3.1, 3.2]. In Theorem 3.13 in [49] we describe the one-to-one correspondence relating the set of canonical equilibrium macrostates for the empirical vectors (denoted by $\mathcal{E}_{\beta,K}$ in [49]) and the set of canonical equilibrium macrostates for the spin per site (denoted by $\widetilde{\mathcal{E}}_{\beta,K}$ in [49]). At the end of section 10 we return to the set of canonical equilibrium macrostates for the empirical vectors when we discuss issues related to the equivalence and nonequivalence of ensembles.

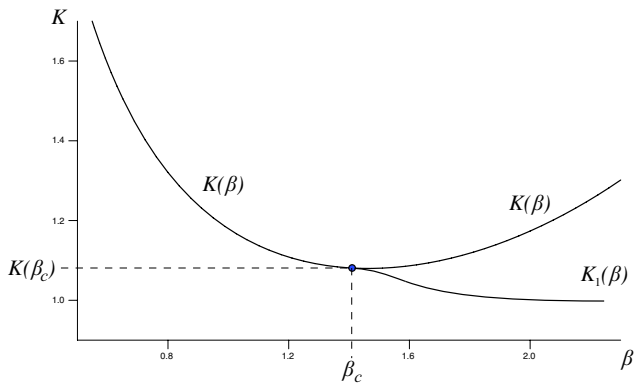

Figure 2: The sets that describe the phase-transition structure of the mean-field Blume-Capel model

This completes our analysis of the mean-field Blume-Capel model. Probabilistic limit theorems for this model are proved in [17, 45, 49]. 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.

10 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 re-evaluate 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 [66] and the work of Thirring [83], 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 non-involutive; i.e., performing it twice does not give back the original function but gives back its concave envelope [49, 84]. 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 [41, 59]. 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 [41, 59, 83]. 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 [18, 19], the mean-field

Blume-Capel model [2, 3, 49, 50], mean-field versions of the Hamiltonian model [24, 64], and the XY model [23]

- Gravitational systems [59, 60, 66, 83]
- Models of coherent structures in turbulence [11, 41, 42, 55, 61, 78]
- Models of plasmas [62, 81];
- Model of the Lennard-Jones gas [7]

Many of these models can be analyzed by the methods to be introduced in this section, which summarize the results in [41]. Further developments in the theory are given in [20]. The reader is referred to these two paper 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 the canonical formulation, a phenomenon that occurs especially in the negative temperature regimes of the vorticity dynamics models [27, 28, 55, 61]. 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 [87].

10.1 Large Deviation Analysis

The general class of models to be considered include both spin models and models of coherent structures in turbulence, and for these two sets of models several of the definitions take slightly different forms. The models to be considered are defined in terms of the following quantities. After presenting the general setup, we will verify that it applies to the Curie-Weiss model. The large deviation analysis of that model, treated in subsection 9.1, inspired the general approach presented here.

- A sequence of probability spaces $(\Omega_n, \mathcal{F}_n, P_n)$ indexed by $n \in \mathbb{N}$, 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 $n \in \mathbb{N}$ 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; we will see such a model in the next section. In this case one replaces H_n in the second bullet 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. Item 3 takes one form for spin models and a different form for models of coherent structures in turbulence. Items 1, 2, and 4 are the same for these two sets of models.

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. The precise description for the two sets of models is as follows.

   Spin models. 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.,

   $$\lim_{n\to\infty} \sup_{\omega\in\Omega_n} |H_n(\omega)/a_n - \tilde{H}(Y_n(\omega))| = 0. \tag{10.1}$$

   Models of coherent structures in turbulence. As $n \to \infty$

   $$H_n(\omega) = \tilde{H}(Y_n(\omega)) + \mathrm{o}(1) \quad \text{uniformly for } \omega \in \Omega_n;$$

   i.e.,

   $$\lim_{n\to\infty} \sup_{\omega\in\Omega_n} |H_n(\omega) - \tilde{H}(Y_n(\omega))| = 0. \tag{10.2}$$

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 $\mathcal{X}$

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

We now verify that this general setup applies to the Curie-Weiss model.

Example 10.1.

- n spins $\omega_i \in \{-1, 1\}$.
- Microstates: $\omega = (\omega_1, \omega_2, \ldots, \omega_n) \in \Omega_n = \{-1, 1\}^n$.
- Prior measures:
$$P_n(\omega) = \frac{1}{2^n} \text{ for each } \omega \in \Omega_n.$$
- Scaling constants: $a_n = n$.
- Hamiltonians:
$$H_n(\omega) = H_n(\omega) = -\frac{1}{2n} \sum_{i,j=1}^{n} \omega_i\omega_j = -\frac{n}{2} \left(\frac{1}{n} \sum_{j=1}^{n} \omega_j \right)^2.$$
- Macroscopic variables:
$$Y_n(\omega) = \frac{1}{n} S_n(\omega) = \frac{1}{n} \sum_{j=1}^{n} \omega_j.$$
- Y_n maps Ω_n into $[-1, 1]$, which is the space of macrostates.
- Energy representation function:
$$H_n(\omega) = -\tfrac{1}{2}(Y_n(\omega))^2 = \tilde{H}(Y_n(\omega)), \text{ where } \tilde{H}(x) = -\tfrac{1}{2}x^2 \text{ for } x \in [-1, 1].$$
Thus (10.1) holds with equality for all ω without the error term $o(a_n)$.
- Large deviation principle with respect to P_n:
$$P_n\{Y_n \in dx\} \asymp e^{-nI(x)}.$$
The version of Cramér's Theorem given in Corollary 6.6 gives the rate function
$$I(x) = \tfrac{1}{2}(1 - x)\log(1 - x) + \tfrac{1}{2}(1 + x)\log(1 + x).$$
This completes the example. ■

Here is a partial list of statistical mechanical models to which the large deviation formalism has been applied. Further details are given in [20, Ex. 2.1].

- The Miller-Robert model of fluid turbulence based on the two dimensional Euler equations [8]. This will be discussed in section 11.
- A model of geophysical flows based on equations describing barotropic, quasi-geostrophic turbulence [42].
- A model of soliton turbulence based on a class of generalized nonlinear Schrödinger equations [43]
- Lattice spin models including the Curie-Weiss model [38, §IV.4], the Curie-Weiss-Potts model [18], the mean-field Blume-Capel model [49], and the Ising model [57, 73]. The large deviation analysis of these models illustrate the three levels of the Donsker-Varadhan theory of large deviations, which are explained in Chapter 1 of [38].
  - Level 1. As we saw in subsection 9.1, 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 6.6. Similar comments apply to the mean-field Blume-Capel model considered in subsection 9.3.
  - Level 2. As we saw in subsection 9.2, for the Curie-Weiss-Potts model [18] 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 3.4.
  - 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 [57, 73]. This is related to level 3 of the Donsker-Varadhan theory, which is formulated for a general class of Markov chains and Markov processes [33]. A special case is treated in [38, Ch. 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 [40, §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 zeroes 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 [41, §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

$$s(u) = -\inf\{I(x) : x \in \mathcal{X}, \tilde{H}(x) = u\}. \tag{10.3}$$

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 [41, Prop. 3.1(a)]. The microcanonical ensemble takes two different forms depending on whether we consider spin models or models of coherent structures in turbulence. For each $u \in \operatorname{dom} s$, $r > 0$, $n \in \mathbb{N}$, and $A \in \mathcal{F}_n$ the microcanonical ensemble for spin models is defined to be the conditioned measure

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

For models of coherent structures in turbulence we work with

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

As shown in [41, 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 $n \in \mathbb{N}$, $\beta \in \mathbb{R}$, 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

$$P_{n,\beta}\{A\} = \frac{1}{Z_n(\beta)} \cdot \int_A \exp[-\beta H_n]\, dP_n. \tag{10.4}$$

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 are 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 in general is not the case.

The large deviation analysis of the canonical ensemble for spin models is summarized in the next theorem, Theorem 10.2. Additional information is given in Theorem 10.3. The modifications in these two theorems necessary for analyzing the canonical ensemble for models of coherent structures in turbulence are indicated in Theorem 10.4.

Part (a) of Theorem 10.2 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 canonical ensemble. Part (b) is the analogue of Theorem 9.1 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 6.4.

Theorem 10.2 (Canonical ensemble for spin models). *For the general spin model we assume that there exists a space of macrostates $\mathcal{X}$, macroscopic variables Y_n, and a Hamiltonian representation function $\tilde{H}$ satisfying*

$$\lim_{n\to\infty} \sup_{\omega\in\Omega_n} |H_n(\omega)/a_n - \tilde{H}(Y_n(\omega))| = 0, \tag{10.5}$$

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 $\varphi(\beta) = -\lim_{n\to\infty} \frac{1}{a_n} \log Z_n(\beta)$ 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* (10.4)*, 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 $C < \infty$

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

Proof. (a) Once we take into account the error between H_n and $a_n\tilde{H}(Y_n)$ expressed in (10.5), the proof of (a) follows from the Laplace principle. Here are the details. By (10.5)

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

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 6.13 with $\psi = \tilde{H}$.

(c) This is proved in Theorem 6.4. The second display 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. ■

The property of the $P_{n,\beta}$-distributions of Y_n expressed in part (c) of Theorem 10.2 has a refinement that arises in our study of the Curie-Weiss model. From Theorem 9.2 we recall that $\mathcal{E}_\beta = \{0\}$ for $0 < \beta \leq 1$ and $\mathcal{E}_\beta = \{\pm m(\beta)\}$ for $\beta > 1$, where $m(\beta)$ is the spontaneous magnetization. According to (9.4), for all $\beta > 0$ the weak limit of $P_{n,\beta}\{S_n/n \in dx\}$ is concentrated on $\mathcal{E}_\beta$. While in the case of the general model treated in the present section one should not expect such a precise formulation, the next theorem gives considerable information, relating weak limits of subsequences of $P_{n,\beta}\{Y_n \in dx\}$ to the set of equilibrium macrostates $\mathcal{E}_\beta$. For example, if one knows that $\mathcal{E}_\beta$ consists of a unique point $\tilde{x}$, then it follows that the entire sequence $P_{n,\beta}\{Y_n \in dx\}$ converges weakly to $\delta_{\tilde{x}}$. This situation corresponds to the absence of a phase transition. The next theorem is proved in [41, Thm. 2.5].

Theorem 10.3 (Canonical ensemble for spin models). *We fix $\beta \in \mathbb{R}$ and use the notation of Theorem* 10.2. *If $\mathcal{E}_\beta$ consists of a unique point $\tilde{x}$, then $P_{n,\beta}\{Y_n \in dx\}$ converges weakly to $\delta_{\tilde{x}}$. If $\mathcal{E}_\beta$ does not consist of a unique point, then any subsequence of $P_{n,\beta}\{Y_n \in dx\}$ has a subsubsequence converging weakly to a probability measure Π_β on $\mathcal{X}$ that is concentrated on $\mathcal{E}_\beta$; i.e., $\Pi_\beta\{(\mathcal{E}_\beta)^c\} = 0$.*

In order to carry out the large deviation analysis of the canonical ensemble for models of coherent structures in turbulence, in Theorems 10.2 and 10.3 one must make two changes: replace the limit (10.5) by

$$\lim_{n\to\infty} \sup_{\omega\in\Omega_n} |H_n(\omega) - \tilde{H}(Y_n(\omega))| = 0, \tag{10.6}$$

where H_n is the Hamiltonian, and replace $Z_n(\beta)$ and $P_{n,\beta}$ by $Z_n(a_n\beta)$ and $P_{n,a_n\beta}$. For easy reference, this is summarized in the next theorem.

Theorem 10.4 (Canonical ensemble for models of coherent structures in turbulence). *For the general model of coherent structures in turbulence we assume that there exists a space of macrostates $\mathcal{X}$, macroscopic variables Y_n, and a Hamiltonian representation function $\tilde{H}$ satisfying* (10.6). *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. Then for each $\beta \in \mathbb{R}$ all the conclusions of Theorems* 10.2 *and* 10.3 *are valid provided that $Z_n(\beta)$ and $P_{n,\beta}$ are replaced by $Z_n(a_n\beta)$ and $P_{n,a_n\beta}$.*

Before carrying out the large deviation analysis of the microcanonical ensemble, we recall the relevant definitions. For $u \in \mathbb{R}$ the microcanonical entropy is defined by

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

For each $u \in \operatorname{dom} s$, $r > 0$, $n \in \mathbb{N}$, and set $A \in \mathcal{F}_n$ the microcanonical ensemble for spin models is defined by

$$P_n^{u,r}\{A\} = P_n\{A \mid H_n/a_n \in [u-r, u+r]\}, \tag{10.7}$$

while the microcanonical ensemble for models of coherent states in turbulence is defined by

$$P_n^{u,r}\{A\} = P_n\{A \mid H_n \in [u-r, u+r]\}, \tag{10.8}$$

In order to simplify the discussion we will work with the microcanonical ensemble for spin models. The treatment of the microcanonical ensemble for models of coherent states in turbulence is analogous. We start our analysis of the microcanonical ensemble by pointing out that $-s$ is the rate function in the large deviation principles, with respect to the prior measures P_n, of 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 [Thm. 6.12]. 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 Theorem 6.14 or can be derived as in the proof of part (a) of Theorem 10.2 by using the equivalent Laplace principle. We summarize this large deviation principle by the notation

$$P_n\{H_n/a_n \in [u-r, u+r]\} \approx \exp[a_n s(u)] \text{ as } n \to \infty, r \to 0. \tag{10.9}$$

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

$$P_n^{u,r}\{Y_n \in B(x,\alpha)\} \approx \exp[-a_n I^u(x)] \text{ as } n \to \infty, r \to 0, \alpha \to 0. \tag{10.10}$$

The calculation that we present shows both the interpretive 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 to 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 to 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 (10.9) 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 (10.10) motivates the correct formula for the rate function [41, Thm. 3.2]:

$$I^u(x) = \begin{cases} I(x) + s(u) & \text{if } \tilde{H}(x) = u, \\ \infty & \text{if } \tilde{H}(x) \neq u. \end{cases} \tag{10.11}$$

We record the facts in the next theorem, which addresses both spin models and models of coherent structures in turbulence. The theorem is proved in [41, §3]. For both of these classes of models the large deviation principle in part (b) takes the same form. 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 [41, Thm. 3.5], is similar to the proof of the analogous property of $\mathcal{E}_\beta$ given in part (c) of Theorem 10.2 and is omitted. A microcanonical analogue of Theorem 10.3 is given in [41, Thm. 3.6].

Theorem 10.5 (Microcanonical ensemble). *Both for the general spin model and for the general model of coherent structures in turbulence we assume that there exists a space of macrostates $\mathcal{X}$, macroscopic variables Y_n, and a Hamiltonian representation function $\tilde{H}$ satisfying* (10.1) *in the case of spin models and* (10.2) *in the case of models of coherent structures in turbulence. 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, $\tilde{H}(Y_n)$ satisfies the large deviation principle with scaling constants a_n and rate function $-s$. With respect to P_n, for the general spin model H_n/a_n also satisfies the large deviation principle with scaling constants a_n and rate function $-s$. For the general model of coherent structures in turbulence the same conclusion holds for H_n.*

(b) *We consider the microcanonical ensemble $P_n^{u,r}$ defined in* (10.7) *for spin models and defined in* (10.8) *for models of coherent structures in turbulence. 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 rate function I^u defined in* (10.11). *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 $C_r < \infty$

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

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

10.2 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

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

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 10.6, 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 $\mathcal{E}^u \subsetneq \mathcal{E}_\beta$.

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 occur by examining concavity and support 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 $u \in \mathbb{R}$:

$$\text{minimize } I(x) \text{ over } \mathcal{X} \text{ subject to the constraint } \tilde{H}(x) = u. \tag{10.12}$$

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

$$\text{minimize } I(x) + \beta\tilde{H}(x) \text{ over } x \in \mathcal{X}. \tag{10.13}$$

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 (10.12) lie among the critical points of $I + \beta\tilde{H}$. However, it does not give, as we do in Theorem 10.7, necessary and sufficient conditions for the solutions of (10.12) to coincide with the solutions of the unconstrained minimization problem (10.13). These necessary and sufficient conditions are expressed in terms of support and concavity properties of the microcanonical entropy $s(u)$. In Theorem 10.6 we give a simplified formulation that focuses on concavity properties of $s(u)$.

Before we explain this, we reiterate a number of properties of $\varphi(\beta)$ and $s(u)$ that emphasize the fundamental nature of these two thermodynamic functions. Properties 1, 2, and 3 show a complete symmetry between the canonical and microcanonical ensembles, a state of affairs that is spoiled by property 4. We use the definitions for spin models.

1. Both $\varphi(\beta)$ and $s(u)$ are given by limits and by variational formulas.

   - $\varphi(\beta)$ expresses the asymptotics of the partition function

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

     via the definition

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

     In addition, $\varphi(\beta)$ is given by the variational formula [Thm. 10.2(a)]

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

- $s(u)$ is defined by the variational formula

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

In addition $s(u)$ expresses the asymptotics of $P_n\{H_n \in du\}$, which satisfies the large deviation principle with rate function $-s(u)$ [Thm. 10.5(a)]; i.e., $P_n\{H_n \in du\} \asymp \exp[a_n s(u)]$. Furthermore, for $u \in \operatorname{dom} s$ we have the limit [41, Prop. 3.1(c)]

$$s(u) = \lim_{r \to 0} \lim_{n \to \infty} \frac{1}{a_n} \log P_n\{H_n/a_n \in [u-r, u+r]\}.$$

2. Both $\varphi(\beta)$ and $s(u)$ are respectively the normalization constants in the rate functions I_β and I^u in the large deviation principles for Y_n with respect to the canonical ensemble and with respect to the microcanonical ensemble:

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

and

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

3. The 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 $\mathcal{E}_\beta$ consists of all $x \in \mathcal{X}$ at which the infimum is attained in

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

- Thus $\mathcal{E}^u$ consists of all $x \in \mathcal{X}$ at which the infimum is attained in

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

4. $\varphi(\beta)$ and $s(u)$ are related via the Legendre-Fenchel transform

$$\varphi(\beta) = \inf_{u \in \mathbb{R}}\{\beta u - s(u)\}. \tag{10.14}$$

As do the two formulas for $\varphi(\beta)$ is item 1, this Legendre-Fenchel transform shows that $\varphi(\beta)$ is always concave, even if $s(u)$ is not. Unless $s(u)$ is concave on $\mathbb{R}$, the dual formula $s(u) = \inf_{\beta \in \mathbb{R}}\{\beta u - \varphi(\beta)\}$ is not valid.

- Proof 1 of (10.14) using variational formulas:

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

- Proof 2 of (10.14) using asymptotic properties:

$$\begin{aligned}\varphi(\beta) &= -\lim_{n\to\infty}\frac{1}{a_n}\log Z_n(\beta)\\ &= -\lim_{n\to\infty}\frac{1}{a_n}\log\int_{\Omega_n}\exp[-\beta H_n]\,dP_n\\ &= -\lim_{n\to\infty}\frac{1}{a_n}\log\int_{\mathbb{R}}\exp[-a_n\beta u]\,P_n\{H_n/a_n\in du\}\\ &= -\sup_{u\in\mathbb{R}}\{-\beta u+s(u)\}\\ &= \inf_{u\in\mathbb{R}}\{\beta u-s(u)\}.\end{aligned}$$

  To derive the next-to-last line we use the fact that with respect to P_n, H_n/a_n satisfies the large deviation principle, and therefore the equivalent Laplace principle, with rate function $-s(u)$ [Thm. 10.5(a)]. Since by (10.1) H_n/a_n is bounded, there exists a compact set K such that $P_n\{H_n/a_n\in K\}=1$ for all n. Hence using the Laplace principle is justified since the function mapping $u\mapsto\beta u$ is bounded and continuous on K.

The complete symmetry between the two ensembles as indicated by properties 1, 2, and 3 is spoiled by property 4. 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 $\mathbb{R}$. In fact, the concavity of s on $\mathbb{R}$ depends on the nature 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 bullets at the beginning of this section. 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 Theorems 10.6 and 10.7.

In Theorem 10.6 we state the main theorem relating concavity properties of the microcanonical entropy with the equivalence and nonequivalence of the microcanonical and canonical ensembles. Before stating the theorem, we illustrate it by considering the microcanonical entropy shown in Figure 3.

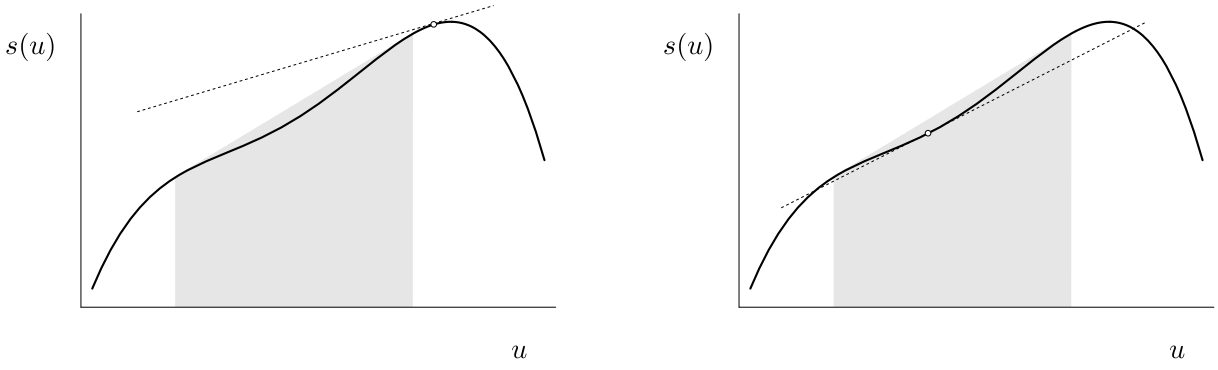

Figure 3: Concavity properties of s determine ensemble equivalence and nonequivalence.

We denote by $[u_l, u_h]$ the projection of the shaded region in Figure 3 onto the u axis.

- For $u < u_l$ and $u > u_h$, s is strictly concave (strictly supporting line), and full equivalence of ensembles holds: $\exists \beta$ such that $\mathcal{E}^u = \mathcal{E}_\beta$.
- For $u = u_l$ and $u = u_h$, s is concave but not strictly concave (nonstrictly supporting line), and partial equivalence of ensembles holds: $\exists \beta$ such that$\mathcal{E}^u \subsetneq \mathcal{E}_\beta$.
- For $u_l < u < u_h$, s is not concave (no supporting line), and nonequivalence of ensembles holds: $\forall \beta, \mathcal{E}^u \cap \mathcal{E}_\beta = \emptyset$.

In order to make this discussion precise, we need several definitions. A function f on $\mathbb{R}$ is said to be concave on $\mathbb{R}$, or 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 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 [79, p. 308]

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

As in the case of convex functions [38, Thm. 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 $u \in \mathbb{R}$ [20, Prop. A.2]. In particular, if for some u, $f(u) \neq f^{**}(u)$, then $f(u) < f^{**}(u)$. We call f the concave, u.s.c. hull of f.

Let $f \not\equiv -\infty$ be a function mapping $\mathbb{R}$ into $\mathbb{R} \cup \{-\infty\}$. The next definitions are reasonable because f^{**} is concave on $\mathbb{R}$.

- f is concave at $u \in \text{dom}\, f$ if $f(u) = f^{**}(u)$.
- f is concave on a convex subset K of dom f if f is concave at all $u \in K$.
- f is nonconcave at $u \in \text{dom}\, f$ if $f(u) < f^{**}(u)$.
- f is strictly concave at $u \in \text{dom}\, f$ if f is concave at u and f^{**} is strictly concave on a convex neighborhood K of dom f containing u; i.e., for all $u \neq v$ in K and all $\lambda \in (0,1)$

$$f^{**}(\lambda u + (1-\lambda)v) > \lambda f^{**}(u) + (1-\lambda) f^{**}(v);$$

These definitions are illustrated for the microcanonical entropy s in Figure 4.

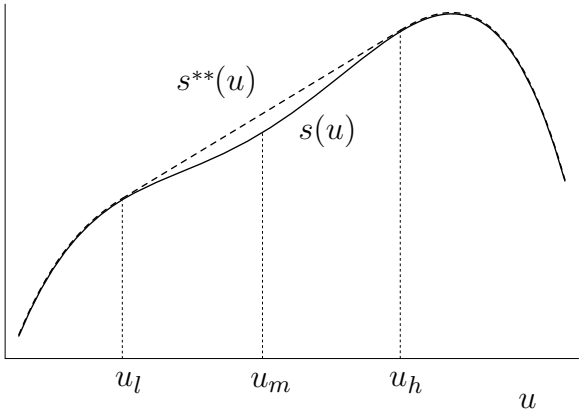

Figure 4: Microcanonical entropy s and its concave, u.s.c. hull s^{**}

- Define s concave at u if $s(u) = s^{**}(u)$.
- Define s strictly concave at u if $s(u) = s^{**}(u)$ and s^{**} strictly concave at u.
- Define s nonconcave at u if $s(u) \neq s^{**}(u)$.

We next state a simplified version of 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 properties of the microcanonical entropy. For simplicity we restrict to points u in the interior of dom s and assume that s is differentiable at all such u.

Theorem 10.6. *In parts* (b)*,* (c)*, and* (d)*, u denotes any point in the interior of* dom s*. We assume that s is differentiable at all such u.*

(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 $\tilde{H}(\mathcal{E}_\beta) \subset \text{dom}\, s$ and*

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

(b) **Full equivalence.** *If s is strictly concave at u, then $\mathcal{E}^u = \mathcal{E}_\beta$ for $\beta = s'(u)$.*

(c) **Partial equivalence.** *If s is concave at u but not strictly concave at u, then $\mathcal{E}^u \subsetneq \mathcal{E}_\beta$ for $\beta = s'(u)$.*

(d) **Nonequivalence.** *If s is not concave at u, then $\mathcal{E}^u \cap \mathcal{E}_\beta = \emptyset$ for all $\beta \in \mathbb{R}$.*

The various possibilities in parts (b), (c), and (d) of Theorem 10.6 are illustrated in [50] for the mean-field Blume-Capel spin model and are shown on the last two pages of this section.

The generalization of Theorem 10.6 to encompass all $u \in \text{dom}\, s$ including boundary points involves support properties of s [20, 41]. This generalization is stated next. It is proved in [41, §4], where it is formulated somewhat differently. The formulation given here specializes Theorem 3.1 in [20] to dimension 1.

Theorem 10.7. *In parts* (b)*,* (c)*, and* (d)*, u denotes any point in* $\text{dom}\, s$*.*

(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 $\tilde{H}(\mathcal{E}_\beta) \subset \text{dom}\, s$ 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 $\mathcal{E}^u \subsetneq \mathcal{E}_\beta$ 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 useful criteria for full or partial equivalence of ensembles and for nonequivalence of ensembles.

- **Full or partial equivalence.** Except possibly for boundary points of $\mathrm{dom}\, s$, s has a supporting line at $u \in \mathrm{dom}\, s$ if and only if s is concave at u [20, Thm. A.5(c)], and thus according to parts (a) and (b) of Theorem 10.6, full or partial equivalence of ensembles holds.

- **Full equivalence.** Assume that $\mathrm{dom}\, s$ is a nonempty interval and that s is strictly concave on the interior of $\mathrm{dom}\, s$; i.e., for all $u \neq v$ in the interior of $\mathrm{dom}\, s$ 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 $\mathrm{dom}\, s$, s has a strictly supporting line at all $u \in \mathrm{dom}\, s$, and thus according to part (a) of the theorem, full equivalence of ensembles holds [20, Thm. A.4(c)].

- **Nonequivalence.** Except possibly for boundary points of $\mathrm{dom}\, s$, s has no supporting line at $u \in \mathrm{dom}\, s$ if and only if s is nonconcave at u [20, Thm. A.5(c)].

A partial proof of the equality in part (a) of Theorems 10.6 and 10.7 is easily provided. Indeed, if $x \in \mathcal{E}_\beta$, then x minimizes $I + \beta\tilde{H}$ over $\mathcal{X}$. Therefore x minimizes $I + \beta\tilde{H}$ over the subset of $\mathcal{X}$ consisting of all x satisfying $\tilde{H}(x) = u$. It follows that x minimizes I over $\mathcal{X}$ subject to the constraint $\tilde{H}(x) = u$ and thus that $x \in \mathcal{E}^{\tilde{H}(x)}$. We conclude that $\mathcal{E}_\beta \subset \cup_{u \in \tilde{H}(\mathcal{E}_\beta)} \mathcal{E}^u$, which is half of the assertion in part (d).

In [42] Theorem 10.7 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 mid-latitude, zone-belt domains on Jupiter.

In [20] we extend the theory developed in [41] and summarized in Theorem 10.7. In [20] it is shown that when the microcanonical ensemble is nonequivalent with 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 with 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 [20] is given in [21], and in [19] it is applied to the Curie-Weiss-Potts model.

The general large deviation procedure presented in the first part of the present section is applied in the next section to the analysis of two models of coherent structures in two-dimensional turbulence, the Miller-Robert model [69, 70, 77, 78] and a related model due to Turkington [86].

10.3 $\mathcal{E}_\beta$, $s'(u)$ and $\mathcal{E}^u$ for the Mean-Field Blume-Capel Model

The mean-field Blume-Capel model was discussed in subsection 9.3. We now illustrate Theorem 10.6 in reference to the sets of equilibrium macrostates $\mathcal{E}_\beta$ and $\mathcal{E}^u$ for the empirical vector with respect to the canonical ensemble and the microcanonical ensemble, respectively.

Case 1. $K = 1.1111$ in the mean-field Blume-Capel model

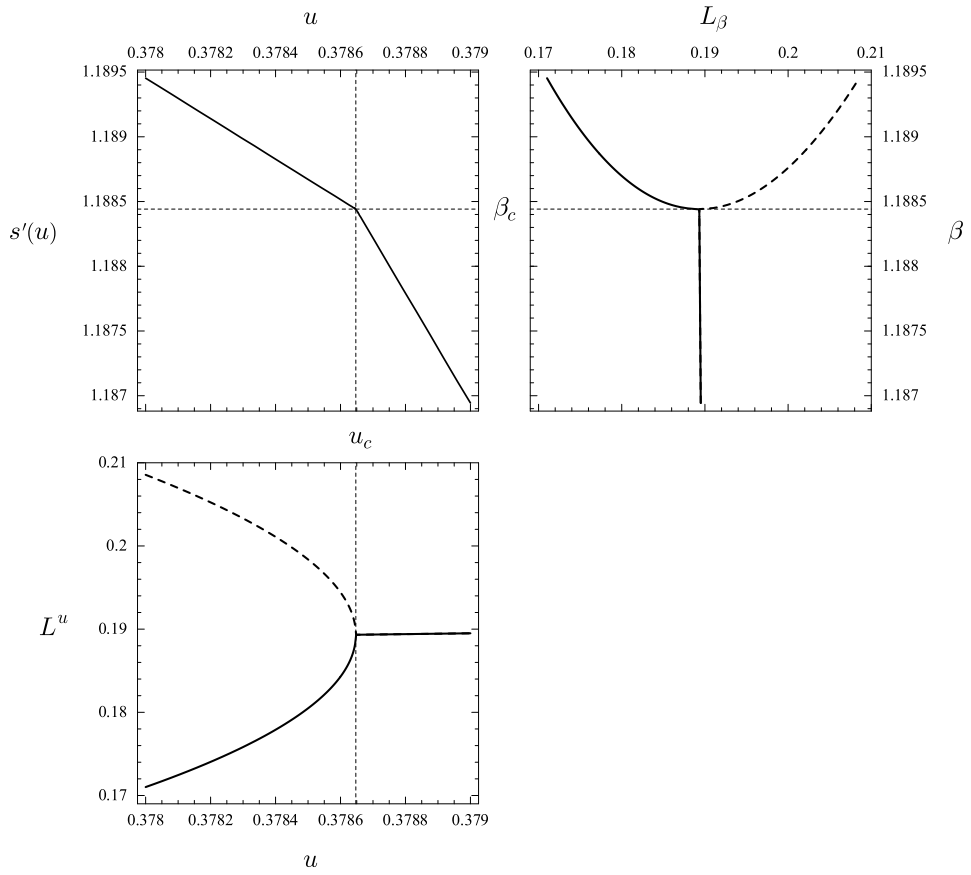

Figure 5: Full equivalence of ensembles for the mean-field Blume-Capel model with $K = 1.1111$. **Top left:** Derivative of the microcanonical entropy $s(u)$. **Top right:** The components L_+ and L_- of the equilibrium empirical vector $L_\beta = (L_-, L_0, L_+)$ in the canonical ensemble as functions of β. For $\beta > \beta_c$ the solid and dashed curves can be taken to represent L_+ and L_-, respectively, or vice versa. **Bottom left:** The components L_+ and L_- of the equilibrium empirical measure $L^u = (L_-, L_0, L_+)$ in the microcanonical ensemble as functions of u. For $u < u_c$ the solid and dashed curves can be taken to represent L_+ and L_-, respectively, or vice versa.

- s' monotonically decreasing $\implies$ s strictly concave
- Full equivalence of ensembles
- Continuous phase transitions in β and u

Case 2. $K = 1.0817$ in the mean-field Blume-Capel model

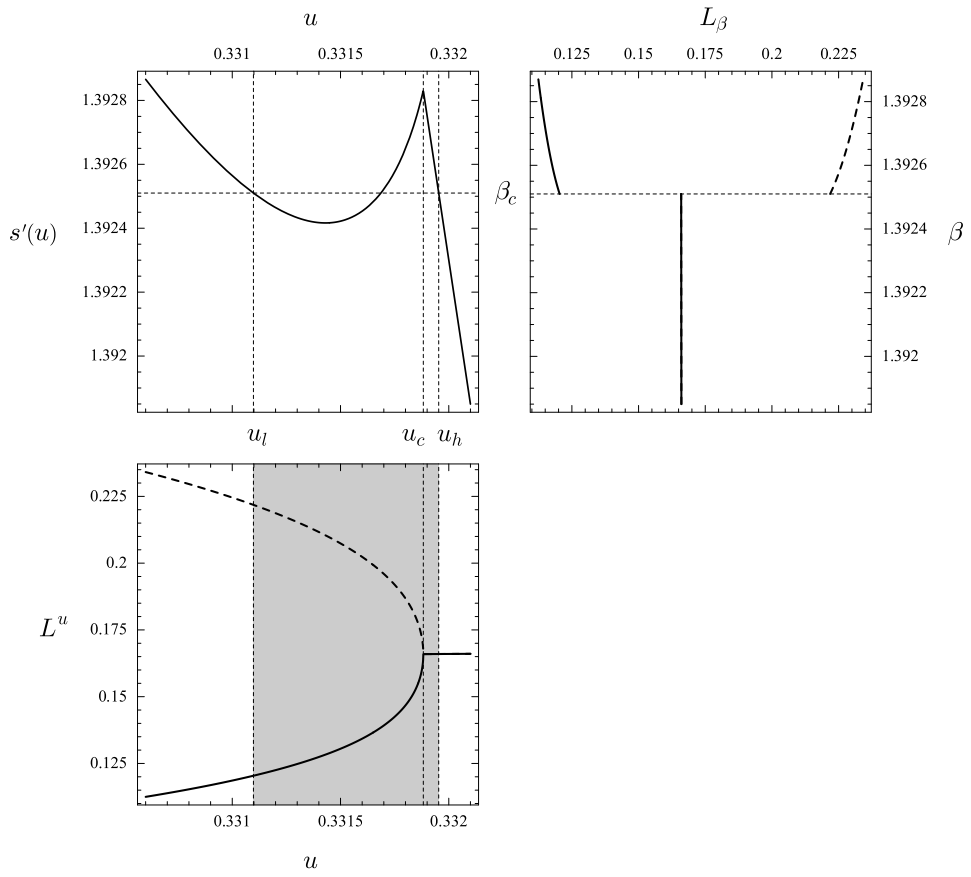

Figure 6: The solid and dashed curves are interpreted as in Figure 5. The shaded area in the bottom left plot corresponds to the region of nonequivalence of ensembles delimited by the open interval (u_l, u_h). The ranges of the inverse temperature and the mean energy used to draw the plots were chosen so as to obtain a good view of the phase transitions.

- s' not decreasing $\implies s$ not concave
- $s(u)$ not concave for $u_l = 0.3311 < u < u_h = 0.33195$
- Canonical phase transition at β_c defined by Maxwell-equal-area line
- Nonequivalence of ensembles: for $u_l < u < u_h$ L^u is not realized by L_β for any β: $\mathcal{E}^u \cap \mathcal{E}_\beta = \emptyset$ for all β.
- First-order phase transition in β versus second-order in u

11 Maximum Entropy Principles in Two-Dimensional Turbulence

This section presents an overview of work in which Gibbs states are used to predict the large-scale, long-lived order of coherent vortices that persist amid the turbulent fluctuations of the vorticity field in two dimensions [8]. This is done by applying a statistical equilibrium theory of the two-dimensional Euler equations, which govern the motion of an inviscid, incompressible fluid. As shown in [16, 67], these equations are reducible to the vorticity transport equations

$$\frac{\partial \omega}{\partial t} + \frac{\partial \omega}{\partial x_1}\frac{\partial \psi}{\partial x_2} - \frac{\partial \omega}{\partial x_2}\frac{\partial \psi}{\partial x_1} = 0 \text{ and } -\triangle \psi = \omega, \tag{11.1}$$

in which ω is the vorticity, ψ is the stream function, and $\triangle = \partial^2/\partial x_1^2 + \partial/\partial x_2^2$ denotes the Laplacian operator on $\mathbb{R}^2$. The two-dimensionality of the flow means that these quantities are related to the velocity field $v = (v_1, v_2, 0)$ according to $(0, 0, \omega) = \text{curl}\, v$ and $v = \text{curl}(0, 0, \psi)$. All of these fields depend upon the time variable $t \in [0, \infty)$ and the space variable $x = (x_1, x_2)$, which runs through a bounded domain in $\mathbb{R}^2$. Throughout this section we assume that this domain equals the unit torus $T^2 = [0, 1) \times [0, 1)$, and we impose doubly periodic boundary conditions on all the flow quantities.

The governing equations (11.1) can also be expressed as a single equation for the scalar vorticity field $\omega = \omega(x, t)$. The periodicity of the velocity field implies that $\int_{T^2} \omega\, dx = 0$. With this restriction on its domain, the Green's operator $G = (-\triangle)^{-1}$ mapping ω into ψ with $\int_{\mathcal{X}} \psi\, dx = 0$ is well-defined. More explicitly, G is the integral operator

$$\psi(x) = G\omega(x) = \int_{\mathcal{X}} g(x - x')\, \omega(x')\, dx',$$

where g is the Green's function defined by the Fourier series

$$g(x - x') = \sum_{0 \neq z \in \mathbb{Z}^2} |2\pi z|^{-2}\, e^{2\pi i \langle z, (x - x') \rangle}\,.$$

Consequently, (11.1) can be considered as an equation in ω alone.

The initial value problem for the equation (11.1) is well-posed for weak solutions whenever the initial data $\omega^0 = \omega(\cdot, 0)$ belongs to $L^\infty(\mathcal{X})$ [67]. However, it is well known that this deterministic evolution does not provide a useful description of the system over long time intervals. When one seeks to quantify the long-time behavior of solutions, therefore, one is compelled to shift from the microscopic, or fine-grained, description inherent in ω to some kind of macroscopic, or coarse-grained, description. We will make this shift by adopting the perspective of equilibrium statistical mechanics. That is, one views the underlying deterministic dynamics as a means of randomizing the microstate ω subject to the conditioning inherent in the conserved quantities for the governing equations (11.1), and one takes the appropriate macrostates to be the canonical Gibbs measures built from these conserved quantities. In doing so, of course, one

accepts an ergodic hypothesis that equates the time averages with canonical ensemble averages. Given this hypothesis, one hopes that these macrostates capture the long-lived, large-scale, coherent vortex structures that persist amid the small-scale vorticity fluctuations. The characterization of these self-organized macrostates, which are observed in simulations and physical experiments, is the ultimate goal of the theory.

The models that we will consider build on earlier and simpler theories, the first of which was due to Onsager [74]. Studying point vortices, he predicted that the equilibrium states with high enough energy have a negative temperature and represent large-scale, coherent vortices. This model was further developed in the 1970's, notably by Montgomery and Joyce [71]. However, the point vortex model fails to incorporate all the conserved quantities for two-dimensional ideal flow.

These conserved quantities are the energy, or Hamiltonian functional, and the family of generalized enstrophies, or Casimir functionals [67]. Expressed as a functional of ω, the kinetic energy is

$$H(\omega) = \frac{1}{2} \int_{\mathcal{X} \times \mathcal{X}} g(x - x')\, \omega(x)\, \omega(x')\, dx\, dx'. \tag{11.2}$$

The so-called generalized enstrophies are the global vorticity integrals

$$A(\omega) = \int_{\mathcal{X}} a(\omega(x))\, dx,$$

where a is an arbitrary continuous real function on the range of the vorticity. In terms of these conserved quantities, the canonical ensemble is defined by the formal Gibbs measure

$$P_{\beta,a}(d\omega) = Z(\beta, a)^{-1} \exp[-\beta H(\omega) - A(\omega)]\, \Pi(d\omega),$$

where $Z(\beta, a)$ is the associated partition function and $\Pi(d\omega)$ denotes some invariant product measure on some phase space of all admissible vorticity fields ω. Of course, this formal construction is not meaningful as it stands due to the infinite dimensionality of such a phase space. We therefore proceed to define a sequence of lattice models on T^2 in order to give a meaning to this formal construction.

One lattice model that respects conservation of energy and also the generalized enstrophy constraints was developed by Miller et. al. [69, 70] and Robert et. al. [77, 78]; we will refer to it as the Miller-Robert model. A related model, which discretizes the continuum dynamics in a different way, was developed by Turkington [86]. These authors use formal arguments to derive maximum entropy principles that are argued to be equivalent to variational formulas for the equilibrium macrostates. In terms of these macrostates, coherent vortices of two-dimensional turbulence can be studied. The purpose of this section is to outline how the large deviation analysis presented in section 10 can be applied to derive these variational formulas rigorously. References [8] and [86] discuss in detail the physical background.

The variational formulas will be derived for the following lattice model that includes both the Miller-Robert model and the Turkington model as special cases. Let T^2 denote the unit

torus $[0,1) \times [0,1)$ with periodic boundary conditions and let $\mathcal{L}$ be a uniform lattice of $n = 2^{2m}$ sites s in T^2, where m is a positive integer. The intersite spacing in each coordinate direction is 2^{-m}. We make this particular choice of n to ensure that the lattices are refined dyadically as m increases, a property that is needed later when we study the continuum limit obtained by sending $n \to \infty$ along the sequence $n = 2^{2m}$. In correspondence with this lattice we have a dyadic partition of T^2 into n squares called microcells, each having area $1/n$. For each $s \in \mathcal{L}$ we denote by $M(s)$ the unique microcell having the site s in its lower left corner. Although $\mathcal{L}$ and $M(s)$ depend on n, this is not indicated in the notation.

The configuration spaces for the lattice model are the product spaces $\Omega_n = \mathcal{Y}^n$, where $\mathcal{Y}$ is a compact set in $\mathbb{R}$. Configurations in Ω_n are denoted by $\zeta = \{\zeta(s), s \in \mathcal{L}\}$, which represents the discretized vorticity field. Let ρ be a probability measure on $\mathcal{Y}$ and let P_n denote the product measure on Ω_n with one-dimensional marginals ρ. As discussed in [8], the Miller-Robert model and the Turkington model differ in their choices of the compact set $\mathcal{Y}$ and the probability measure ρ.

For $\zeta \in \Omega_n$ the Hamiltonian for the lattice model is defined by

$$H_n(\zeta) = \frac{1}{2n^2} \sum_{s,s' \in \mathcal{L}} g_n(s - s')\, \zeta(s)\, \zeta(s'),$$

where g_n is the lattice Green's function defined by the finite Fourier sum

$$g_n(s - s') = \sum_{0 \neq z \in \mathcal{L}^*} |2\pi z|^{-2}\, e^{2\pi i \langle z, s - s' \rangle}$$

over the finite set $\mathcal{L}^* = \{z = (z_1, z_2) \in \mathbb{Z}^2 : -2^{m-1} < z_1, z_2 \leq 2^{m-1}\}$. Let a be any continuous function mapping $\mathcal{Y}$ into $\mathbb{R}$. For $\zeta \in \Omega_n$ we also define functions known as the generalized enstrophies by

$$A_{n,a}(\zeta) = \frac{1}{n} \sum_{s \in \mathcal{L}} a(\zeta(s)),$$

In terms of these quantities we define the partition function

$$Z_n(\beta, a) = \int_{\Omega_n} \exp[-\beta H_n(\zeta) - A_{n,a}(\zeta)]\, P_n(d\zeta)$$

and the canonical ensemble $P_{n,\beta,a}$, which is the probability measure that assigns to a Borel subset B of Ω_n the probability

$$P_{n,\beta,a}\{B\} = \frac{1}{Z_n(\beta, a)} \int_B \exp[-\beta H_n(\zeta) - A_{n,a}(\zeta)]\, P_n(d\zeta). \tag{11.3}$$

These probability measures are parametrized by the constant $\beta \in \mathbb{R}$ and the function $a \in \mathcal{C}(\mathcal{Y})$. The dependence of Gibbs measures on the inverse temperature β is standard, while their dependence on the function a that determines the enstrophy functional is a novelty of this

particular statistical equilibrium problem. The Miller-Robert model and the Turkington model also differ in their choices of the parameter β and the function a.

The main theorem in this section applies the theory of large deviations to derive the continuum limit $n \to \infty$ of the lattice model just introduced. Because the interactions $g_n(s - s')$ in the lattice model are long-range, one must replace β and a by $n\beta$ and na in order to obtain a nontrivial continuum limit [8, 69, 70]. Replacing β and a by $n\beta$ and na in the formulas for the partition function and the Gibbs state is equivalent to replacing H_n and A_n by nH_n and nA_n and leaving β and a unscaled. We carry out the large deviation analysis of the lattice model by applying the general procedure specified in the preceding section, making the straightforward modifications necessary to handle both the Hamiltonian and the generalized enstrophy. Thus we seek a space of macrostates, a sequence of macroscopic variables Y_n, representation functions $\tilde{H}$ and $\tilde{A}_a$ for the Hamiltonian and for the generalized enstrophy, and a large deviation principle for Y_n with respect to the product measures P_n. The first marginal of a probability measure μ on $T^2 \times \mathcal{Y}$ is defined to be the probability measure $\mu_1\{A\} = \mu\{A \times \mathcal{Y}\}$ for Borel subsets A of T^2.

- **Space of macrostates.** This is the space $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$ of probability measures on $T^2 \times \mathcal{Y}$ with first marginal θ, where $\theta(dx) = dx$ is Lebesgue measure on T^2.

- **Macroscopic variables.** For each $n \in \mathbb{N}$, Y_n is the measure-valued function mapping $\zeta \in \Omega_n$ to $Y_n(\zeta, dx \times dy) \in \mathcal{P}_\theta(T^2 \times \mathcal{Y})$ defined by

$$Y_n(dx \times dy) = Y_n(\zeta, dx \times dy) = dx \otimes \sum_{s \in \mathcal{L}} 1_{M(s)}(x)\, \delta_{\zeta(s)}(dy).$$

  Thus for Borel subsets A of $T^2 \times \mathcal{Y}$

$$Y_n\{A\} = \sum_{s \in \mathcal{L}} \int_A 1_{M(s)}(x)\, dx\, \delta_{\zeta(s)}(dy).$$

  Since $\sum_{s \in \mathcal{L}} 1_{M(s)}(x) = 1$ for all $x \in T^2$, the first marginal of Y_n equals dx.

- **Hamiltonian representation function.** $\tilde{H} : \mathcal{P}_\theta(T^2 \times \mathcal{Y}) \mapsto \mathbb{R}$ is defined by

$$\tilde{H}(\mu) = \frac{1}{2} \int_{(T^2 \times \mathcal{Y})^2} g(x - x')\, y\, y'\, \mu(dx \times dy)\, \mu(dx' \times dy'),$$

  where

$$g(x - x') = \sum_{0 \neq z \in \mathcal{Z}^2} |2\pi z|^{-2} \exp[2\pi i \langle z, x - x' \rangle].$$

  As proved in [8, Lem. 4.4], $\tilde{H}$ is bounded and continuous and there exists $C < \infty$ such that

$$\sup_{\zeta \in \Omega_n} |H_n(\zeta) - \tilde{H}(Y_n(\zeta, \cdot))| \leq C \left(\frac{\log n}{n} \right)^{1/2} \quad \text{for all } n \in \mathbb{N}. \tag{11.4}$$

- **Generalized enstrophy representation function.** $\tilde{A}_a : \mathcal{P}_\theta(T^2 \times \mathcal{Y}) \mapsto \mathbb{R}$ is defined by

$$\tilde{A}_a(\mu) = \int_{T^2 \times \mathcal{Y}} a(y)\, \mu(dx \times dy).$$

$\tilde{A}_a$ is bounded and continuous and

$$A_{n,a}(\zeta) = \tilde{A}_a(Y_n(\zeta, \cdot)) \text{ for all } \zeta \in \Omega_n. \tag{11.5}$$

- **Large deviation principle for Y_n.** With respect to the product measures P_n, Y_n satisfies the large deviation principle on $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$ with rate function the relative entropy

$$I_{\theta \times \rho}(\mu) = \begin{cases} \displaystyle\int_{T^2 \times \mathcal{Y}} \left(\log \frac{d\mu}{d(\theta \times \rho)} \right) d\mu & \text{if } \mu \ll \theta \times \rho \\ \infty & \text{otherwise.} \end{cases}$$

We first comment on the last item. The large deviation principle for Y_n with respect to P_n is far from obvious and in fact is one of the main contributions of [8]. We will address this issue after specifying the large deviation behavior of the model in Theorem 11.1. Concerning (11.4), since $\theta\{M(s)\} = 1/n$, it is plausible that

$$\tilde{H}(Y_n(\zeta, \cdot)) = \frac{1}{2} \sum_{s,s' \in \mathcal{L}} \int_{M(s) \times M(s')} g(x - x')\, dx\, dx'\, \zeta(s)\zeta(s')$$

is a good approximation to $H_n(\zeta) = [1/(2n^2)] \sum_{s,s' \in \mathcal{L}} g_n(s - s')\zeta(s)\zeta(s')$. Concerning (11.5), for $\zeta \in \Omega_n$ we have

$$\tilde{A}_a(Y_n(\zeta, \cdot)) = \int_{T^2 \times \mathcal{Y}} a(y)\, Y_n(\zeta, dx \times dy) = \frac{1}{n} \sum_{s \in \mathcal{L}} a(\zeta(s)) = A_{n,a}(\zeta).$$

The proofs of the boundedness and continuity of $\tilde{A}_a$ are straightforward.

Part (a) of Theorem 11.1 gives the asymptotic behavior of the scaled partition functions $Z_n(n\beta, na)$, and part (b) states the large deviation principle for Y_n with respect to the scaled canonical ensemble $P_{n,n\beta,na}$. The rate function has the familiar form

$$I_{\beta,a} = I_{\rho \times \theta} + \beta \tilde{H} + \tilde{A} - \varphi(\beta, a),$$

where $\varphi(\beta, a)$ denotes the canonical free energy. In the formula for $I_{\beta,a}$ the relative entropy $I_{\rho \times \theta}$ arises from the large deviation principle for Y_n with respect to P_n, and the other terms arise from (11.4), (11.5), and the form of $P_{n,n\beta,na}$. Part (c) of the theorem gives properties of the set $\mathcal{E}_{\beta,a}$ of equilibrium macrostates. $\mathcal{E}_{\beta,a}$ consists of measures μ at which the rate function $I_{\beta,a}$ in part (b) attains its infimum of 0 over $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$. The proof of the theorem is omitted since it is similar to the proof of Theorem 10.4, which adapts Theorems 10.2 and 10.3 to the setting of models of coherent structures in turbulence.

Theorem 11.1. *For each $\beta \in \mathbb{R}$ and $a \in \mathcal{C}(\mathcal{Y})$ the following conclusions hold.*

(a) *The canonical free energy $\varphi(\beta, a) = -\lim_{n\to\infty} \frac{1}{n} \log Z_n(n\beta, na)$ exists and is given by the variational formula*

$$\varphi(\beta, a) = \inf_{\mu \in \mathcal{P}_\theta(T^2 \times \mathcal{Y})} \{\beta \tilde{H}(\mu) + \tilde{A}_a(\mu) + I_{\rho\times\theta}(\mu)\}.$$

(b) *With respect to the scaled canonical ensemble $P_{n,n\beta,na}$ defined in* (11.3)*, Y_n satisfies the large deviation principle on $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$ with scaling constants n and rate function*

$$I_{\beta,a}(\mu) = I_{\rho\times\theta}(\mu) + \beta \tilde{H}(\mu) + \tilde{A}_a(\mu) - \varphi(\beta, a).$$

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

$$\mathcal{E}_{\beta,a} = \{\mu \in \mathcal{P}_\theta(T^2 \times \mathcal{Y}) : I_{\beta,a}(\mu) = 0\}.$$

Then $\mathcal{E}_{\beta,a}$ is a nonempty, compact subset of $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$. In addition, if A is a Borel subset of $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$ such that $\overline{A} \cap \mathcal{E}_{\beta,a} = \emptyset$, then $I_{\beta,a}(\overline{A}\,) > 0$ and for some $C < \infty$

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

In section 3 of [8] we discuss the physical implications of the theorem and the relationship between the following concepts in the context of the Miller-Robert model and the Turkington model: $\mu \in \mathcal{P}_\theta(T^2 \times \mathcal{Y})$ is a canonical equilibrium macrostate (i.e., $\mu \in \mathcal{E}_{\beta,a}$) and μ satisfies a corresponding maximum entropy principle. In the Miller-Robert model, the maximum entropy principle takes the form of minimizing the relative entropy $I_{\theta\times\rho}(\mu)$ over $\mu \in \mathcal{P}_\theta(T^2 \times \mathcal{Y})$ subject to the constraints

$$\tilde{H}(\mu) = H(\omega^0) \quad \text{and} \quad \int_{T^2} \mu(dx \times \cdot) = \int_{T^2} \delta_{\omega^0(x)}(\cdot)\, dx,$$

where ω^0 is an initial vorticity field and $H(\omega^0)$ is defined in (11.2). By analogy with our work in the preceding section, this constrained minimization problem defines the set of equilibrium macrostates with respect to the microcanonical ensemble for the Miller-Robert model. The fact that each $\mu \in \mathcal{E}_{\beta,a}$ is also a microcanonical equilibrium macrostate is a consequence of part (d) of Theorem 10.6 adapted to handle both the Hamiltonian and the generalized enstrophy. In the Turkington model, the maximum entropy principle takes a somewhat related form in which the second constraint appearing in the Miller-Robert maximum entropy principle is relaxed to a family of convex inequalities parametrized by points in $\mathcal{Y}$. Understanding for each model the relationship between equilibrium macrostates μ and the corresponding maximum entropy principle allows one to identify a steady vortex flow with a given equilibrium macrostate μ. Through this identification, which is described in [8], one demonstates how the equilibrium macrostates capture the long-lived, large-scale, coherent structures that persist amid the small-scale vorticity fluctuations.

We spend the rest of this section outlining how the large deviation principle is proved for the macroscopic variables

$$Y_n(dx \times dy) = dx \otimes \sum_{s \in \mathcal{L}} 1_{M(s)}(x)\, \delta_{\zeta(s)}(dy)$$

with respect to the product measures P_n. The proof is based on the innovative technique of approximating Y_n by a doubly indexed sequence of random measures $W_{n,r}$ for which the large deviation principle is, at least formally, almost obvious. This doubly indexed sequence, obtained from Y_n by averaging over an intermediate scale, clarifies the physical basis of the large deviation principle and reflects the multiscale nature of turbulence. A similar large deviation principle is derived in [68, 76] by an abstract approach that relies on a convex analysis argument. That approach obscures the role of spatial coarse-graining in the large deviation behavior.

In order to define $W_{n,r}$, we recall that $\mathcal{L}$ contains $n = 2^{2m}$ sites s. For even $r < 2m$ we consider a regular dyadic partition of T^2 into 2^r macrocells $\{D_{r,k}, k = 1, 2, \ldots, 2^r\}$. Each macrocell contains $n/2^r$ lattice sites and is the union of $n/2^r$ microcells $M(s)$, where $M(s)$ contains the site s in its lower left corner. We now define

$$W_{n,r}(dx \times dy) = W_{n,r}(\zeta, dx \times dy) = dx \otimes \sum_{k=1}^{2^r} 1_{D_{r,k}}(x)\, \frac{1}{n/2^r} \sum_{s \in D_{r,k}} \delta_{\zeta(s)}(dy).$$

$W_{n,r}$ is obtained from Y_n by replacing, for each $s \in D_{r,k}$, the point mass $\delta_{\zeta(s)}$ by the average $(n/2^r)^{-1} \sum_{s \in D_{r,k}} \delta_{\zeta(s)}$ over the $n/2^r$ sites contained in $D_{r,k}$.

We need the key fact that with respect to a suitable metric d on $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$, $d(Y_n, W_{n,r}) \leq \sqrt{2}/2^{r/2}$ for all $n = 2^{2m}$ and all even $r \in \mathbb{N}$ satisfying $r < 2m$. The proof of this approximation property uses the fact that the diameter of each macrocell $D_{r,k}$ equals $\sqrt{2}/2^{r/2}$ [8, Lem. 4.2]. The next theorem states the two-parameter large deviation principle for $W_{n,r}$ with respect to the product measures P_n. The approximation property $d(Y_n, W_{n,r}) \leq \sqrt{2}/2^{r/2}$ implies that with respect to P_n, Y_n satisfies the Laplace principle, and thus the equivalent large deviation principle, with the same rate function $I_{\theta \times \rho}$ [8, Lem. 4.3]. Subtleties involved in invoking the Laplace principle are discussed in the proof of that lemma.

Theorem 11.2. *With respect to the product measures P_n, the sequence $W_{n,r}$ satisfies the following two-parameter large deviation principle on $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$ with rate function $I_{\theta \times \rho}$: for any closed subset F of $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$*

$$\limsup_{r \to \infty} \limsup_{n \to \infty} \frac{1}{n} \log P_n\{W_{n,r} \in F\} \leq -I_{\theta \times \rho}(F)$$

and for any open subset G of $\mathcal{P}_\theta(T^2 \times \mathcal{Y})$

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

Our purpose in introducing the doubly indexed process $W_{n,r}$ is the following. The local averaging over the sets $D_{r,k}$ introduces a spatial scale that is intermediate between the macroscopic scale of the torus T^2 and the microscopic scale of the microcells $M(s)$. As a result, $W_{n,r}$ can be written in the form

$$W_{n,r}(dx \times dy) = dx \otimes \sum_{k=1}^{2^r} 1_{D_{r,k}}(x)\, L_{n,r,k}(dy), \tag{11.6}$$

where

$$L_{n,r,k}(dy) = L_{n,r,k}(\zeta, dy) = \frac{1}{n/2^r} \sum_{s \in D_{r,k}} \delta_{\zeta(s)}(dy).$$

Since each $D_{r,k}$ contains $n/2^r$ lattice sites s, with respect to P_n the sequence $\{L_{n,r,k}, k = 1, \ldots, 2^r\}$ is a family of i.i.d. empirical measures. For each r and each $k \in \{1, \ldots, 2^r\}$ Sanov's Theorem 6.7 implies that as $n \to \infty$, $L_{n,r,k}$ satisfies the large deviation principle on $\mathcal{P}(\mathcal{Y})$ with scaling constants $n/2^r$ and rate function I_ρ.

We next motivate the large deviation principle for $W_{n,r}$ stated in Theorem 11.2. Suppose that $\mu \in \mathcal{P}_\theta(T^2 \times \mathcal{Y})$ has finite relative entropy with respect to $\theta \times \rho$ and has the special form

$$\mu(dx \times dy) = dx \otimes \tau(x, dy), \quad \text{where } \tau(x, dy) = \sum_{k=1}^{2^r} 1_{D_{r,k}}(x)\, \tau_k(dy) \tag{11.7}$$

and $\tau_1, \ldots, \tau_{2^r}$ are probability measures on $\mathcal{Y}$. The representation (11.6), Sanov's Theorem, and the independence of $L_{n,r,1}, \ldots, L_{n,r,2^r}$ suggest that

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log P_n \{W_{n,r} \sim \mu\} \\
&\quad = \lim_{n\to\infty} \frac{1}{n} \log P_n \{L_{n,r,k} \sim \tau_k, k = 1, \ldots, 2^r\} \\
&\quad = \frac{1}{2^r} \sum_{k=1}^{2^r} \lim_{n\to\infty} \frac{1}{n/2^r} \log P_n \{L_{n,r,k} \sim \tau_k\} \\
&\quad \approx -\frac{1}{2^r} \sum_{k=1}^{2^r} I_\rho(\tau_k) = -\int_{T^2} I_\rho(\tau(x, \cdot))\, dx \\
&\quad = -\int_{T^2} \int_{\mathcal{Y}} \left(\log \frac{d\tau(x, \cdot)}{d\rho(\cdot)}(y) \right) \tau(x, dy)\, dx \\
&\quad = -\int_{T^2 \times \mathcal{Y}} \left(\log \frac{d\mu}{d(\theta \times \rho)}(x, y) \right) \mu(dx \times dy) \\
&\quad = -I_{\theta\times\rho}(\mu).
\end{aligned}$$

Because of this calculation, the two-parameter large deviation principle for $W_{n,r}$ with rate function $I_{\theta\times\rho}$ is certainly plausible, in view of the fact that any measure $\mu \in \mathcal{P}_\theta(T^2 \times \mathcal{Y})$ can be well approximated, as $r \to \infty$, by a sequence of measures of the form (11.7) [9, Lem. 3.2].

The reader is referred to [8] for an outline of the proof of this two-parameter large deviation principle. The large deviation principle for $W_{n,r}$ is a special case of a large deviation principle proved in [9] for an extensive class of random measures that includes $W_{n,r}$ as a special case.

This completes our application of the theory of large deviations to models of two-dimensional turbulence. The asymptotic behavior of these models is stated in Theorem 11.1. One of the main components of the proof is the large deviation principle for the macroscopic variables Y_n, which in turn follows by approximating Y_n by the doubly indexed sequence $W_{n,r}$ and proving the large deviation principle for this sequence. This proof relies on Sanov's Theorem, which generalizes Boltzmann's 1877 calculation of the asymptotic behavior of multinomial probabilities. Earlier in the paper we used the elementary form of Sanov's Theorem stated in Theorem 3.4 to derive the form of the Gibbs state for the discrete ideal gas and to motivate the version of Cramér's Theorem needed to analyze the Curie-Weiss model [Cor. 6.6]. It is hoped that both the importance of Boltzmann's 1877 calculation and the applicability of the theory of large deviations to problems in statistical mechanics have been amply demonstrated in these lectures. It is also hoped that these lectures will inspire the reader to discover new applications.

References

[1] R. R. Bahadur and S. Zabell. Large deviations of the sample mean in general vector spaces. *Ann. Prob.* 7:587–621, 1979.

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

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

[4] M. Blume, Theory of the first-order magnetic phase change in UO_2, *Phys. Rev.* 141:517–524, 1966.

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

[6] L. Boltzmann. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respecktive den Sätzen über das Wärmegleichgewicht (On the relationship between the second law of the mechanical theory of heat and the probability calculus). *Wiener Berichte* 2, no. 76, 373–435, 1877.

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

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

[9] C. Boucher, R. S. Ellis, and B. Turkington. Spatializing random measures: doubly indexed processes and the large deviation principle. *Ann. Prob.* 27:297–324, 1999. Erratum: *Ann. Prob.* 30:2113, 2002.

[10] C. Cercignani. *Ludwig Boltzmann: The Man Who Trusted Atoms*. Oxford: Oxford Univ. Press, 1998.

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

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

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

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

[15] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* 23:493–507, 1952.

[16] A. J. Chorin, *Vorticity and Turbulence*, New York: Springer, 1994.

[17] M. Costeniuc, R. S. Ellis and P. T.-H. Otto. Multiple Critical Behavior of Probabilistic Limit Theorems in the Neighborhood of a Tricritical Point. *J. Stat. Phys.* 127:495–552, 2007.

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

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

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

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

[22] H. Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles* 736:2–23, 1938. Colloque consacré à la théorie des probabilités, Vol. 3, Hermann, Paris.

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

[24] T. Dauxois, V. Latora, A. Rapisarda, S. Ruffo, and A. Torcini. 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*, volume 602 of *Lecture Notes in Physics*, pp. 458–487, New York: Springer, 2002.

[25] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Second edition. New York: Springer, 1998.

[26] J.-D. Deuschel and D. W. Stroock. *Large Deviations*. Boston: Academic Press, 1989.

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

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

[29] I. H. Dinwoodie and S. L. Zabell. Large deviations for exchangeable random vectors. *Ann. Prob.* 20:1147–1166, 1992.

[30] R. L. Dobrushin and S. B. Shlosman. Large and moderate deviations in the Ising model. *Adv. Soviet Math.* 20:91–219, 1994.

[31] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, I. *Comm. Pure Appl. Math.* 28:1–47, 1975.

[32] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, III. *Comm. Pure Appl. Math.* 29:389–461, 1976.

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

[34] P. Dupuis and R. S. Ellis. Large deviations for Markov processes with discontinuous statistics, II: random walks. *Probab. Th. Relat. Fields* 91:153–194, 1992.

[35] P. Dupuis and R. S. Ellis. The large deviation principle for a general class of queueing systems, I. *Trans. Amer. Math. Soc.* 347:2689–2751, 1995. An error in the proof of the large deviation upper bound in Theorem 4.3 (see the third display on page 2730) was corrected by I. Ignatiouk-Robert, Large deviations for processes with discontinuous statistics, *Ann. Prob.* 33:1479–1508, 2005.

[36] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. New York: Wiley, 1997.

[37] P. Dupuis, R. S. Ellis, and A. Weiss. Large deviations for Markov processes with discontinuous statistics, I: general upper bounds. *Ann. Prob.* 19:1280–1297, 1991.

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

[39] R. S. Ellis. Large deviations for a general class of random vectors. *Ann. Prob.* 12:1–12, 1984.

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

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

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

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

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

[45] R. S. Ellis, J. Machta, and P. T. Otto. Refined asymptotics of the spin and finite-size scaling of the magnetization. In preparation, 2008.

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

[47] R. S. Ellis and C. M. Newman. The statistics of Curie-Weiss models. *J. Stat. Phys.* 19:149–161, 1978.

[48] R. S. Ellis, C. M. Newman, and J. S. Rosen. 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, 1980.

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

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

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

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

[53] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. New York: Wiley, 1986.

[54] W. R. Everdell. *The First Moderns.* Chicago: The University of Chicago Press, 1997.

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

[56] W. Feller. *An Introduction to Probability Theory and Its Applications*. Vol. I, second edition. New York: Wiley, 1957.

[57] H. Föllmer and S. Orey. Large deviations for the empirical field of a Gibbs measure. *Ann. Prob.* 16:961–977, 1987.

[58] J. Gärtner. On large deviations from the invariant measure. *Th. Prob. Appl.* 22:24–39, 1977.

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

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

[61] M. K.-H. Kiessling and J. L. Lebowitz. The micro-canonical point vortex ensemble: beyond equivalence. *Lett. Math. Phys.* 42:43–56, 1997.

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

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

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

[65] D. Lindley. *Boltzmann's Atom: The Great Debate That Launched a Revolution in Physics*. New York: Free Press, 2001.

[66] D. Lynden-Bell and R. Wood. The gravo-thermal catastrophe in isothermal spheres and the onset of red-giant structure for stellar systems. *Mon. Notic. Roy. Astron. Soc.* 138:495, 1968.

[67] C. Marchioro and M. Pulvirenti. *Mathematical Theory of Incompressible Nonviscous Fluids*. New York: Springer, 1994.

[68] J. Michel and R. Robert. Large deviations for Young measures and statistical mechanics of infinite dimensional dynamical systems with conservation law. *Comm. Math. Phys.* 159:195–215, 1994.

[69] J. Miller. Statistical mechanics of Euler equations in two dimensions. *Phys. Rev. Lett.* 65:2137–2140 (1990).

[70] J. Miller, P. Weichman and M. C. Cross. Statistical mechanics, Euler's equations, and Jupiter's red spot. *Phys. Rev. A* 45:2328–2359, 1992.

[71] D. Montgomery and G. Joyce. Statistical mechanics of negative temperature states. *Phys. Fluids* 17:1139–1145, 1974.

[72] P. Ney. Private communication, 1997.

[73] S. Olla. Large deviations for Gibbs random fields. *Prob. Th. Rel. Fields* 77:343–359, 1988.

[74] L. Onsager. Statistical hydrodynamics. *Suppl. Nuovo Cim.* 6:279–287, 1949.

[75] G. Parisi. *Statistical Field Theory*. Redwood City, CA: Addison-Wesley, 1988.

[76] R. Robert. Concentration et entropie pour les mesures d'Young. *C. R. Acad. Sci. Paris* 309, Série I:757–760, 1989.

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

[78] R. Robert and J. Sommeria. Statistical equilibrium states for two-dimensional flows. *J. Fluid Mech.* 229:291–310, 1991.

[79] R. T. Rockafellar. *Convex Analysis*. Princeton: Princeton Univ. Press, 1970.

[80] E. Seneta. *Non-Negative Matrices and Markov Chains.* Second edition. New York: Springer, 1981.

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

[82] D. W. Stroock. *An Introduction to the Theory of Large Deviations*. New York: Springer, 1984.

[83] W. Thirring. Systems with negative specific heat. *Z. Physik*, 235:339–352, 1970.

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

[85] H. Touchette. The large deviation approach to statistical mechanics. Submitted for publication, 2008. Posted at http://arxiv.org/abs/0804.0327.

[86] B. Turkington. Statistical equilibrium measures and coherent states in two-dimensional turbulence. *Comm. Pure Appl. Math.* 52:781–809, 1999.

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

[88] S. R. S. Varadhan. Asymptotic properties and differential equations. *Comm. Pure Appl. Math.* 19:261–286, 1966.

[89] A. S. Wightman. 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 Univ. Press, 1979.

[90] F. Y. Wu. The Potts model. *Rev. Mod. Phys.* 54:235–268, 1982.