An Overview of the Theory of Large Deviations and Applications to Statistical Mechanics

RICHARD S. ELLIS*


We survey a number of results in the theory of large deviations, including Cramér’s Theorem, the Donsker-Varadhan theory, and other modern developments. We then apply the large deviation theorems to three models in statistical mechanics, the Curie-Weiss model, the Curie-Weiss-Potts model, and the Ising model. These models are analyzed by the three respective levels of the Donsker-Varadhan theory: the sample means (level 1), the empirical measures (level 2), and the empirical processes and fields (level 3). In the last section a general approach to the large deviation analysis of models in statistical mechanics is formulated. Key words: Large deviation principle, Curie-Weiss model, Curie-Weiss-Potts model, Ising model.

1. INTRODUCTION

The theory of large deviations studies the exponential decay of certain probabilities. It is one of the most active topics in probability today, and it has applications in numerous areas including statistics, queueing theory, and statistical mechanics. The theory of large deviations has been the subject of countless papers, a number of books, and a host of special conferences. The seed from which this forest grew is the 1938 paper of Harald Cramér (Cramér, 1988) which treats the large deviation theory of the sample means of independent, identically distributed (i.i.d.) random variables. The present paper was delivered in Stockholm at the symposium honoring the centennial of the birth of this gifted and influential mathematician, who also made seminal contributions to a number of other fields besides large deviations. At the Cramér Symposium I learned about the close ties in Sweden between mathematics studied at the university and mathematics applied in the business world, especially insurance and finance. Indeed, Cramér’s interest in large deviations apparently grew out of such applications. In this paper I will discuss Cramér’s contribution, mention a number of modern developments, and treat several applications to statistical mechanics, an area that interested Cramér through the work of Khinchine (1949). It will be seen that the theory of large deviations is a useful tool for exploring the fascinating physical phenomenon of phase transitions.

* This research was supported in part by a grant from the National Science Foundation (NSF-DMS-9123575).

© 1995 Scandinavian University Press. ISSN 0346-1238
Although after the work of Cramér a number of important papers appeared (e.g., Bahadur, 1960; Chernoff, 1952), it was not until the mid-1970's that the subject exploded. First through the efforts of M. D. Donsker and S. R. S. Varadhan in the United States and A. D. Wentzell in the Soviet Union, then through the efforts of a host of many other people, a vast new terrain of large deviation phenomena was opened up. What makes the theory particularly exciting and rich are the connections with other fields including statistical mechanics (Ellis, 1985; Lanford, 1973), stochastic optimal control theory (Dupuis & Ellis, 1995), partial differential equations (Dupuis et al., 1990), and geometry (Varadhan, 1967).

The applications of the theory of large deviations to statistical mechanics that we will treat in this paper are not an afterthought. In my opinion the profundity of the theory of large deviations is in part due to the fact that it is a number of important aspects the theory is the mathematical expression of profound statistical mechanical ideas. Conversely, the exquisite formalism of statistical mechanics readily lends itself to large deviation calculations.

The important paper of Lanford (1973), written in 1973 three years before the beginning of the Donsker-Varadhan and Wentzell "revolutions," is a striking instance of the way in which statistical mechanics has enriched the theory of large deviations. Lanford uses large deviation ideas to explain the fact that while matter is extremely complicated on the microscopic level, it can be described on the macroscopic level by a small number of parameters. Among his techniques are subadditivity methods, which are completely natural in the statistical mechanics setting and which he also applies, as a digression, to the much easier problem of the sample means of i.i.d. random variables. Through the paper (Bahadur & Zabell, 1979), which proves infinite-dimensional versions of Cramér's Theorem, these methods entered mainstream large deviation theory and have become more or less standard. They have been used, for example, in Deuschel & Stroock (1989) and Stroock (1984), where they are used to prove large deviation results for Markov chains, and in somewhat modified form in Dupuis & Ellis (1956), where they are used to prove the large deviation principle for a general class of queueing systems. Lanford's paper is highly recommended as is the work of Wightman (Wightman, 1979). This is a beautiful overview of the thermodynamics of phase transitions, based on the Gibbs geometric approach through convexity, and the statistical mechanics of lattice systems.

Section 2 of this paper gives two elementary examples for which large deviation estimates can be derived using combinatoric arguments. In Section 3 the concept of large deviation principle is defined and several consequences of the large deviation principle are mentioned. Section 4 states and sketches the proof of Cramér's Theorem, which is the large deviation principle for the sample means of i.i.d. random vectors taking values in $\mathbb{R}^d$. In Section 5 we present an extension of Cramér's Theorem due to Gärtner (1977) and myself (1984) and now known in the literature as the Gärtner-Ellis Theorem. This theorem is applied later in the paper when we consider the Ising model in statistical mechanics. In Cramér's Theorem and in the Gärtner-Ellis Theorem the rate function is defined as a Legendre-Fenchel transform and so is always convex. This convexity is not a general feature. Indeed, in Section 6 two examples are presented of large deviation principles having
nonconvex rate functions. In that section a generalization of the Legendre-Fenchel transform is introduced which may be useful in proving the large deviation principle in certain cases when the Gärtner-Ellis Theorem cannot be applied. Section 7 presents the very special case of the Donsker-Varadhan theory which concerns a sequence of i.i.d. random variables taking values in a complete separable metric space. This theory discusses large deviation phenomena on three levels. Level-1 is the level of the sample means; level-2, of the empirical measures; and level-3, of the empirical processes. The remainder of the paper is devoted to applications to statistical mechanics and, in particular, the problem of phase transitions, of which Section 8 provides an overview. In Sections 9, 10, and 11 we apply each of the three levels of the Donsker-Varadhan theory to three specific models in statistical mechanics. The Curie-Weiss model is treated via a level-1 analysis; the Curie-Weiss-Potts model via a level-2 analysis; and the Ising model via a level-3 analysis. As we will see in Section 11, what distinguishes the Ising model is the existence of large deviation phenomena on two different scales. Away from a certain phase transition interval, the scaling is by volume, which is the analogue of the scaling in Cramér’s Theorem. However, inside the phase transition interval the scaling is by surface area. In the last section of the paper, Section 12, a general approach to the large deviation analysis of models in statistical mechanics is formulated.

I thank the organizers of the Cramér Symposium for inviting me to participate.

2. BASIC CONCEPTS AND ELEMENTARY EXAMPLES

Before the contribution of Professor Cramér can be explained, suitable background must be provided. In order to do this, we introduce some notation that will be used throughout this paper. We have a complete separable metric space $\mathcal{X}$ and for each $n \in \mathbb{N}$ a probability space $(\Omega_n, \mathcal{F}_n, P_n)$ and a random variable $W_n$ mapping $\Omega_n$ into $\mathcal{X}$. The reader who is uncomfortable with this generality may initially consider the case $\mathcal{X} = \mathbb{R}$ or $\mathcal{X}$ a finite set. As we will see in this paper, however, one confronts a wide variety of complete separable metric spaces in the course of analysing large deviation problems that arise naturally. The set $\Omega_n$ is a space of points or, to use a more descriptive language suggested by statistical mechanics, it is the configuration space of a random system. $\mathcal{F}_n$ is a $\sigma$-field of subsets of $\Omega_n$ and $P_n$ is a probability measure on $(\Omega_n, \mathcal{F}_n)$. We denote by $b(\cdot, \cdot)$ the metric on $\mathcal{X}$.

In many problems there exists a distinguished point $x_0$ of $\mathcal{X}$ with the property that for each $\varepsilon > 0$

$$\lim_{n \to \infty} P_n \{b(W_n, x_0) \geq \varepsilon\} = 0.$$  \hfill (2.1)

Now let $A$ be a Borel subset of $\mathcal{X}$ whose closure does not contain $x_0$. For example, $A$ may equal the complement of the open ball with center $x_0$ and radius $\varepsilon$. The limit (2.1) implies that $P_n \{W_n \in A\} \to 0$ as $n \to \infty$. A natural question is to investigate the rate at which this probability converges to 0. In many problems there exists a sequence of real numbers $\{a_n, n \in \mathbb{N}\}$ tending to $\infty$ and a function $I$ mapping $\mathcal{X}$
into $[0, \infty)$ such that

$$P_n\{W_n \in A\} \approx \exp[-a_n I(A)], \quad \text{where} \quad I(A) = \inf_{x \in A} I(x). \quad (2.2)$$

This is an example of a large deviation estimate. The meaning of formula (2.2) will be made precise in Section 3 when we define the notion of large deviation principle. In this context the function $I$ is called a rate function.

We next present two simple examples for which large deviation estimates can be derived using combinatoric arguments. A minor modification of the first example will be used in Section 9 to analyze the Curie-Weiss model in statistical mechanics.

**EXAMPLE 2.1: Fair Coin Tossing.** For each $n \in \mathbb{N}$ the configuration space $\Omega_n$ equals $\{0, 1\}^n$. Thus $\Omega_n$ consists of all sequences $\omega = (\omega_j, j = 1, 2, \ldots, n)$, where each $\omega_j$ equals 0 or 1, a 0 representing a tail on the $j$-th toss and a 1 a head. Since the number of configurations in $\Omega_n$ is $2^n$, in order to model a fair coin we assign to each $\omega \in \Omega_n$ the probability $P_n(\omega) = 2^{-n}$. Let $\mathcal{F}_n$ denote the set of all subsets of $\Omega_n$. $P_n$ is extended to a probability measure on $\mathcal{F}_n$ by defining for any subset $A$ of $\Omega_n$

$$P_n\{A\} = \sum_{\omega \in A} \frac{1}{2^n}.$$ 

For each $\omega \in \Omega_n$ we define the random variable

$$S_n(\omega) = \sum_{j=1}^{n} \omega_j,$$

which counts the number of heads in the sequence of coin tosses represented by $\omega$. The law of large numbers for sums of i.i.d. random variables implies that for each $\varepsilon > 0$

$$\lim_{n \to \infty} P_n\{|S_n - n/2| \geq \varepsilon\} = 0. \quad (2.3)$$

We are interested in an estimate of the form of equation (2.2) for the probability that the sample mean $S_n/n$ takes values in certain subsets $A$ of $[0, 1]$. For $x \in [0, 1]$ we define the function

$$I(x) = x \log(2x) + (1-x) \log[2(1-x)]. \quad (2.4)$$

The function $I$ is strictly convex and nonnegative on $[-1, 1]$, and it has a unique minimum point at $x = 1/2$. Thus

$$I(x) > I(1/2) = 0 \quad \text{for} \quad x \in [-1, 1], x \neq 1/2.$$ 

For any $n \in \mathbb{N}$ and integer $k \in \{0, 1, \ldots, n\}$

$$P_n\{\frac{S_n}{n} = \frac{k}{n}\} = \frac{1}{2^n} \frac{n!}{k!(n-k)!} \quad ,$$

and a weak form of Stirling's approximation shows that uniformly over all such $k$

$$P_n\{\frac{S_n}{n} = \frac{k}{n}\} = \exp\left[-nI\left(\frac{k}{n}\right) + O(\log n)\right]. \quad (2.5)$$

It is now straightforward to prove the following result.
THEOREM 2.2. Let $A$ be a closed subset of $[0, 1]$ which is the closure of its interior. For example, as in equation (2.3), let $A = \{0, 1/2 - \varepsilon \} \cup [1/2 + \varepsilon, 1]$ for some $\varepsilon \in (0, 1/2)$. For $x \in [0, 1]$ we define the rate function $I(x)$ by formula (2.4). Then

$$\lim_{n \to \infty} \frac{1}{n} \log P_n \left\{ \frac{S_n}{n} \in A \right\} = - I(A) = - \min_{x \in A} I(x).$$

Proof. For $n \in \mathbb{N}$ we define the set

$$A_n = A \cap \{ z \in [0, 1] : z = k/n \text{ for some } k = 0, 1, \ldots, n \}.$$ 

As shown in [24, Sect. 1.3], formula (2.5) implies that

$$\lim_{n \to \infty} \frac{1}{n} \log P_n \left\{ \frac{S_n}{n} \in A \right\} = - \lim_{n \to \infty} \min_{z \in A_n} I(z).$$

Since $A$ equals the closure of its interior, for each $x \in A$ there exists a sequence $\{ z_n \in \mathbb{N} \}$ such that each $z_n \in A_n$, and $z_n \to x$. It follows that

$$\lim_{n \to \infty} \min_{z \in A_n} I(z) = \min_{x \in A} I(z) = I(A).$$

This completes the proof.

If the closed set $A$ in the theorem does not contain the point $\frac{1}{2}$, then the quantity $I(A)$ is positive and the probability $P_n \left\{ \frac{S_n}{n} \in A \right\}$ converges to 0 exponentially fast as $n \to \infty$.

Here is a useful device for passing from the combinatorial result expressed in equation (2.5) to the limit in Theorem 2.2. Let us summarize equation (2.5) using the notation

$$P_n \left\{ \frac{S_n}{n} \in dx \right\} \sim e^{-nI(x)} dx.$$ 

Then formally we have

$$P_n \left\{ \frac{S_n}{n} \in A \right\} = \int_A P_n \left\{ \frac{S_n}{n} \in dx \right\} \approx \int_A e^{-nI(x)} dx \approx e^{-nI(x)} \text{ as } n \to \infty.$$ 

The last step is a consequence of Laplace’s method, which states that as $n \to \infty$ the main contribution to the integral comes from the largest value of the integrand.

For $x \in [0, 1]$ the rate function $I(x)$ can be written in the form

$$I(x) = \log 2 - H(x), \quad \text{where } H(x) = - x \log x - (1 - x) \log(1 - x).$$

The quantity $H(x)$ is the Shannon entropy of the probability distribution that assigns probability $x$ to the point 0 and probability $1 - x$ to the point 1. The maximum value of $H(x)$ as $x$ ranges over the interval $[0, 1]$ is $\log 2$. Thus, in the felicitous phrase of S. R. S. Varadhan, the rate function $I$ has the form of an "entropy shortfall." This is typical of many large deviation problems. We will return to coin tossing later in the paper.
Our second example is a modest generalization of the coin tossing example. This example will be used in Section 10 to analyze the \( q \)-state Curie-Weiss-Potts model in statistical mechanics.

**EXAMPLE 2.3:** Fair \( q \)-Faced Die. Let \( q \) be an integer exceeding 2. If \( q = 6 \), then the following setup models a standard fair die. Otherwise we are modeling a fair die with \( q \) faces. For each \( n \in \mathbb{N} \) the configuration space \( \Omega_n \) equals \( \{1, 2, \ldots, q\}^n \). Thus \( \Omega_n \) consists of all sequences \( \omega = (\omega_j)_{j=1}^n \), where each \( \omega_j \) takes values in the set \( \{1, 2, \ldots, q\} \). Since the number of configurations in \( \Omega_n \) is \( q^n \), in order to model a fair die we assign to each \( \omega \in \Omega_n \) the probability \( P_n(\omega) = q^{-n} \). Let \( \mathcal{F}_n \) denote the set of all subsets of \( \Omega_n \). \( P_n \) is extended to a probability measure on \( \mathcal{F}_n \) by defining for any subset \( A \) of \( \Omega_n \)

\[
P_n(A) = \sum_{\omega \in A} \frac{1}{q^n}.
\]

In Example 2.1 the random variable \( S_n(\omega) \) counts the number of heads in the sequence of coin tosses represented by \( \omega \). In the present example it makes sense to count the relative frequency with which the values 1, 2, \ldots, \( q \) arise in each sequence of die tosses. Thus for each \( n \in \mathbb{N} \), \( i \in \{1, 2, \ldots, q\} \), and \( \omega \in \Omega_n \) we define the random variable

\[
L_n^i(\omega) = \frac{1}{n} \sum_{j=1}^n \delta_{\omega_j}(i),
\]

where \( \delta_{\omega_j}(i) \) equals 1 if \( \omega_j = i \) and equals 0 if \( \omega_j \neq i \). We also define the random vector

\[
L_n(\omega) = (L_n^1(\omega), L_n^2(\omega), \ldots, L_n^q(\omega)).
\]

This random vector takes values in the set

\[
\mathcal{M} = \left\{ v \in \mathbb{R}^q : v = (v_1, v_2, \ldots, v_q), v_i \geq 0, \sum_{i=1}^q v_i = 1 \right\},
\]

which is the subset of \( \mathbb{R}^q \) consisting of probability vectors. \( L_n(\omega) \) is called the **empirical vector** of the configuration \( \omega \).

We denote by \( \mu^* \) the point in \( \mathcal{M} \) all of whose coordinates equal \( 1/q \) and we write \( \| \cdot \| \) for the Euclidean norm on \( \mathbb{R}^q \). Since

\[
L_n(\omega) = \frac{1}{n} \sum_{j=1}^n (\delta_{\omega_j}(1), \delta_{\omega_j}(2), \ldots, \delta_{\omega_j}(q)),
\]

the law of large numbers for sums of i.i.d. random vectors implies that for each \( \varepsilon > 0 \)

\[
\lim_{n \to \infty} P_n \left( \| L_n - \mu^* \| \geq \varepsilon \right) = 0.
\]

As in Example 2.1, we are interested in an estimate of the form of equation (2.2) for the probability that \( L_n \) takes values in certain subsets \( A \) of \( \mathcal{M} \). For \( v \in \mathcal{M} \) we define the function

\[
I(v \mid \mu^*) = \frac{1}{\varepsilon} \sum_{i=1}^q v_i \log(v_i)q).
\]
This quantity is called the relative entropy of the probability vector \( v \) with respect to the probability vector \( \mu^* \). The function \( I(v \mid \mu^*) \) is strictly convex and nonnegative on \( \mathcal{M} \), and it has a unique minimum point at \( v = \mu^* \). Thus
\[
I(v \mid \mu^*) > I(\mu^* \mid \mu^*) = 0 \quad \text{for } v \in \mathcal{M}, v \neq \mu^*.
\]
For \( n \in \mathbb{N} \) let \( \mathcal{M}_n \) denote the set of all \( q \)-tuples of nonnegative integers \( k = (k_1, k_2, \ldots, k_q) \) satisfying \( \sum_{i=1}^q k_i = n \). Thus for \( k \in \mathcal{M}_n \) the vector \( k/n \) is a probability vector lying in \( \mathcal{M} \). For each \( k \in \mathcal{M}_n \)
\[
P_n \left\{ L^n = \frac{k}{n} \right\} = \frac{1}{q^n} \frac{n!}{\prod_{i=1}^q k_i!},
\]
and as in Example 2.1 a weak form of Stirling's approximation shows that uniformly over all such vectors \( k \)
\[
P_n \left\{ L^n = \frac{k}{n} \right\} = \exp \left[ -n \left( \frac{k}{n} \mid \mu^* \right) + O(\log n) \right]. \quad (2.8)
\]

It is now straightforward to prove the following result.

**THEOREM 2.4.** Let \( A \) be a closed subset of \( \mathcal{M} \) which is the closure of its interior. For example, as in equation (2.6), let \( A = \{ v \in \mathcal{M} : \| v - \mu^* \| \geq \varepsilon \} \) for some sufficiently small \( \varepsilon > 0 \). For \( v \in \mathcal{M} \) we define the relative entropy \( I(v \mid \mu^*) \) by formula (2.7). Then
\[
\lim_{n \to \infty} \frac{1}{n} \log P_n \{ L^n \in A \} = -I(A \mid \mu^*) = -\min_{v \in A} I(v \mid \mu^*).
\]

Proof. For \( n \in \mathbb{N} \) we define the set
\[A_n = A \cap \{ v \in \mathcal{M} : v = k/n \text{ for some } k \in \mathcal{M}_n \} \].

As shown in [24, Sect. 1.4], formula (2.8) implies that
\[
\lim_{n \to \infty} \frac{1}{n} \log P_n \{ L^n \in A \} = -\lim_{n \to \infty} \min_{v \in A_n} I(v \mid \mu^*).
\]

Since \( A \) equals the closure of its interior, for each \( v \in A \) there exists a sequence \( \{ v_n, n \in \mathbb{N} \} \) such that each \( v_n \in A_n \) and \( v_n \to v \). It follows that
\[
\lim_{n \to \infty} \min_{v \in A_n} I(v \mid \mu^*) = \min_{v \in A} I(v \mid \mu^*) = I(A \mid \mu^*).
\]

This completes the proof. \( \square \)

If the set \( A \) in the theorem does not contain the point \( \mu^* \), then the quantity \( I(A \mid \mu^*) \) is positive and the probability \( P_n \{ L^n \in A \} \) converges to 0 exponentially fast as \( n \to \infty \).

For \( v \in \mathcal{M} \) the rate function \( I(v \mid \mu^*) \) in Theorem 2.4 can be written in the form
\[
I(v \mid \mu^*) = \log q - H(v), \quad \text{where } H(v) = -\sum_{i=1}^q v_i \log v_i.
\]
The quantity $H(v)$ is the Shannon entropy of the probability distribution that assigns the probability $v_i$ to the point $i \in \{1, 2, \ldots, q\}$. The maximum value of $H(v)$ as $v$ ranges over the set $\mathcal{M}$ is $\log q$. Hence as in Example 2.1, the rate function $I(v \mid \mu^*)$ has the form of an "entropy shortfall." We will return to the fair $q$-faced die later in the paper.

The two elementary examples just considered motivate the definition of large deviation principle to be given in the next section.

3. DEFINITION OF LARGE DEVIATION PRINCIPLE

In Examples 2.1 and 2.3 we found two instances of random quantities $\{W_n, n \in \mathbb{N}\}$ having the property that for a large class of subsets $A$ the large deviation limit

$$
\lim_{n \to \infty} \frac{1}{n} \log P_n \{W_n \in A\}
$$

exists and can be expressed as the infimum (actually, as the minimum) of a certain function over the set $A$. The proofs were carried out by combinatoric arguments, which obviously are not available in more general cases. Our first task in the present section is to define the concept of rate function followed by the concept of large deviation principle. The latter concept involves a large deviation upper bound and a large deviation lower bound. From these bounds, large deviation limits of the type just discussed can easily be derived.

**DEFINITION 3.1.** Let $\mathcal{X}$ be a complete separable metric space and $I$ a function mapping $\mathcal{X}$ into the extended nonnegative real numbers $[0, \infty]$. We call $I$ a rate function on $\mathcal{X}$ if for all $L \in [0, \infty)$ the level set $\{x \in \mathcal{X} : I(x) \leq L\}$ is compact.

Since a compact subset of a complete separable metric space is closed, it follows that the level sets $\{x \in \mathcal{X} : I(x) \leq L\}$ are all closed. This implies that $I$ is lower semicontinuous; i.e., whenever $\{x_n, n \in \mathbb{N}\}$ is a sequence in $\mathcal{X}$ converging to a point $x \in \mathcal{X}$, then $\liminf_{n \to \infty} I(x_n) \geq I(x)$. The compactness of the level sets of $I$ as expressed in Definition 3.1 is a useful technical property. For $A$ a subset of $\mathcal{X}$ we write $I(A)$ for the infimum of $I$ over $A$.

**DEFINITION 3.2.** Let $\mathcal{X}$ be a complete separable metric space and for each $n \in \mathbb{N}$ let $(\Omega_n, \mathcal{F}_n, P_n)$ be a probability space and $W_n$, a random variable mapping $\Omega_n$ into $\mathcal{X}$. Also let $I$ be a rate function on $\mathcal{X}$ and $\{a_n, n \in \mathbb{N}\}$ a sequence of positive numbers tending to $\infty$ as $n \to \infty$. We say that the sequence $\{W_n, n \in \mathbb{N}\}$ satisfies the large deviation principle on $\mathcal{X}$ with rate function $I$ and norming constants $\{a_n, n \in \mathbb{N}\}$ if the following hold.

(a) For all closed subsets $F$ of $\mathcal{X}$

$$
\limsup_{n \to \infty} \frac{1}{a_n} \log P_n \{W_n \in F\} \leq -I(F).
$$
(b) For all open subsets $G$ of $\mathcal{X}$

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

If for each $n \in \mathbb{N}$, $a_n = n$, then we say that the sequence $\{ W_n, n \in \mathbb{N} \}$ satisfies the large deviation principle on $\mathcal{X}$ with rate function $I$.

We call the limit in part (a) of this definition the large deviation upper bound and the limit in part (b) the large deviation lower bound. If the large deviation principle holds on $\mathcal{X}$ with rate function $I$ and norming constants $\{a_n, n \in \mathbb{N}\}$, then with respect to these norming constants the rate function is unique [10, Lemma 2.1.1]. We summarize the large deviation principle by the notation

$$P_n \{ W_n \in dx \} \asymp e^{-a_n I(x)} \, dx.$$  \hfill (3.1)

Of course, even in the case $\mathcal{X} = \mathbb{R}^d$, where $dx$ can be thought of as Lebesgue measure, this asymptotic result should not be taken literally. As we will see, however, use of this notation will allow us to motivate several important results.

It is convenient to extend the concept of large deviation principle to sequences of probability measures. Let $I$ be a rate function on a complete separable metric space $\mathcal{X}$, $\{Q_n, n \in \mathbb{N}\}$ a sequence of probability measures on the Borel $\sigma$-field of $\mathcal{X}$, and $\{a_n, n \in \mathbb{N}\}$ a sequence of positive numbers tending to $\infty$ as $n \to \infty$. We say that the sequence $\{Q_n, n \in \mathbb{N}\}$ satisfies the large deviation principle on $\mathcal{X}$ with rate function $I$ and normal constants $\{a_n, n \in \mathbb{N}\}$ if for all closed subsets $F$ of $\mathcal{X}$

$$\limsup_{n \to \infty} \frac{1}{a_n} \log Q_n \{ F \} \leq -I(F)$$

and for all open subsets $G$ of $\mathcal{X}$

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

The main thread of this narrative continues in the next section, where we present Cramer's Theorem. We find it useful to list in the next two propositions a number of consequences of the large deviation principle which will be applied later in the paper. Upon first reading, these propositions may be omitted. Part (a) of the first proposition gives a useful criterion for the existence of large deviation limits as in Theorems 2.2 and 2.4. Part (b) shows how to pass from the large deviation principle to law of large numbers-type limits. Part (c) introduces a device known as the contraction principle which allows one to derive a new large deviation principle from a known one.

**Proposition 3.3.** Let $\{W_n, n \in \mathbb{N}\}$ be a sequence of random variables taking values in a complete separable metric space $\mathcal{X}$. We assume that $\{W_n, n \in \mathbb{N}\}$ satisfies the large deviation principle on $\mathcal{X}$ with rate function $I$ and norming constants $\{a_n, n \in \mathbb{N}\}$. The following conclusions hold.
(a) Given $A$ a Borel subset of $\mathcal{X}$ we denote by $\overline{A}$ the closure of $A$ and by $A^o$ the interior of $A$. If $I(\overline{A}) = I(A^o)$, then we have the large deviation limit
\[ \lim_{n \to \infty} \frac{1}{n} \log P_n\{W_n \in A\} = -I(A). \]

(b) The set $\mathcal{E} = \{x \in \mathcal{X} : I(x) = 0\}$ is a nonempty compact subset of $\mathcal{X}$. In addition, if $A$ is a Borel subset of $\mathcal{X}$ whose closure has empty intersection with $\mathcal{E}$, then
\[ \lim_{n \to \infty} P_n\{W_n \in A\} = 0. \]

(c) Let $\mathcal{Y}$ be a complete separable metric and $g$ a continuous function mapping $\mathcal{X}$ into $\mathcal{Y}$. For each $y \in \mathcal{Y}$ define the function
\[ J(y) = \inf\{I(x) : x \in \mathcal{X}, g(x) = y\}. \]
Then $J$ is a rate function on $\mathcal{Y}$, and the sequence $\{g(W_n), n \in \mathbb{N}\}$ satisfies the large deviation principle on $\mathcal{Y}$ with rate function $J$ and norming constants $\{a_n, n \in \mathbb{N}\}$.

Proof. (a) Since $A^o \subset A \subset \overline{A}$, we obtain from the large deviation principle the following string of inequalities:
\[ -I(A^o) \leq \lim inf_{n \to \infty} \frac{1}{a_n} \log P_n\{W_n \in A^o\} \leq \lim inf_{n \to \infty} \frac{1}{a_n} \log P_n\{W_n \in A\} \]
\[ \leq \lim sup_{n \to \infty} \frac{1}{a_n} \log P_n\{W_n \in A\} \leq \lim sup_{n \to \infty} \frac{1}{a_n} \log P_n\{W_n \in \overline{A}\} \leq -I(\overline{A}). \]
The hypothesis that $I(\overline{A}) = I(A^o)$ guarantees that $I(A) = I(\overline{A}) = I(A^o)$ and that
\[ \lim_{n \to \infty} \frac{1}{a_n} \log P_n\{W_n \in A\} = \lim sup_{n \to \infty} \frac{1}{a_n} \log P_n\{W_n \in A\} = -I(A). \]
This completes the proof of part (a).

(b) Since $I$ takes values in the set $[0, \infty]$, the set $\mathcal{E}$ equals the compact level set $\{x \in \mathcal{X} : I(x) = 0\}$ of the rate function $I$. Hence $\mathcal{E}$ is compact. Evaluating the large deviation upper bound and lower bound for the set $F = G = \mathcal{X}$, we see that the infimum of $I$ over the whole space $\mathcal{X}$ equals 0. Since $I$ is lower semicontinuous, it attains its infimum over the nonempty compact level set $K = \{x \in \mathcal{X} : I(x) \leq 1\}$. Thus there exists a point $x_0 \in K$ such that $I(x_0) = I(K) = I(\mathcal{X}) = 0$. This point $x_0$ lies in the set $\mathcal{E}$, which is therefore nonempty. This proves the first assertion in part (b).

If $A$ is a Borel subset of $\mathcal{X}$ whose closure $\overline{A}$ has empty intersection with $\mathcal{E}$, then $I(\overline{A}) > 0$ and by the large deviation upper bound
\[ \lim sup_{n \to \infty} \frac{1}{a_n} \log P_n\{W_n \in A\} \leq \lim sup_{n \to \infty} \frac{1}{a_n} \log P_n\{W_n \in \overline{A}\} \leq -I(\overline{A}). \]
This proves the second assertion in part (b).

(c) We omit the routine calculation that shows that $J$ is a rate function. Let us prove the large deviation upper bound for the sequence $\{g(W_n), n \in \mathbb{N}\}$. The large deviation lower bound is proved similarly. Given $F$ a closed subset of $\mathcal{Y}$, the set
$g^{-1}(F)$ is a closed subset of $\mathcal{F}$, and we have

$$
\lim \sup \frac{1}{a_n} P_n \{ g(W_n) \in F \} = \lim \sup \frac{1}{a_n} \log P_n \{ W_n \in g^{-1}(F) \} \\
\leq -J(g^{-1}(F)) \\
= -\inf_{x \in \mathcal{F}} \{ I(x) : x \in \mathcal{F}, g(x) \in F \} \\
= -\inf_{y \in \mathcal{F}} \{ \inf_{x \in \mathcal{F}} \{ I(x) : x \in \mathcal{F}, g(x) = y \} \} \\
= -\inf_{y \in \mathcal{F}} J(y),
$$

as claimed. The proof of the proposition is complete. \(\square\)

Two additional consequences of the large deviation principle are presented in the next proposition. Part (a) gives Varadhan's basic result on evaluating the asymptotics of certain integrals (Varadhan, 1966). Part (b) uses Varadhan's result to bootstrap the large deviation principle for a sequence of random variables $\{W_n, n \in \mathbb{N}\}$ into a large deviation principle for an associated sequence of measures. Part (b) will be applied several times in our treatment of statistical mechanical models. The assumption in both parts (a) and (b) that $F$ be bounded above may be weakened. See, for example, Theorem II.7.1 and II.7.2 in Ellis (1985) and condition (2.38) there.

**PROPOSITION 3.4.** Let $\{W_n, n \in \mathbb{N}\}$ be a sequence of random variables taking values in a complete separable metric space $\mathcal{F}$. We assume that $\{W_n, n \in \mathbb{N}\}$ satisfies the large deviation principle on $\mathcal{F}$ with rate function $I$ and norming constants $\{a_n, n \in \mathbb{N}\}$. The following conclusions hold.

(a) Let $\Phi$ be a continuous function mapping $\mathcal{F}$ into $\mathbb{R}$ which is bounded above. Then

$$
\lim \frac{1}{a_n} \log E_n \{ \exp [a_n \Phi(W_n)] \} = \lim \frac{1}{a_n} \log \int_{\mathcal{F}} \exp [a_n \Phi(x)] P_n \{ W_n \in dx \} \\
= \sup_{x \in \mathcal{F}} \{ \Phi(x) - I(x) \}.
$$

(b) Let $\Phi$ be a continuous function mapping $\mathcal{F}$ into $\mathbb{R}$ which is bounded above. For $n \in \mathbb{N}$ and Borel subsets $A$ of $\mathcal{F}$ we define the probability measures

$$
Q_n, \Phi(A) = \frac{1}{Z_n} \int_A \exp [a_n \Phi(x)] P_n \{ W_n \in dx \},
$$

where $Z_n$ denotes the normalizing constant

$$
Z_n = \int_{\mathcal{F}} \exp [a_n \Phi(x)] P_n \{ W_n \in dx \}.
$$

Then the sequence of probability measures $\{Q_n, \Phi, n \in \mathbb{N}\}$ satisfies the large deviation principle on $\mathcal{F}$ with norming constants $\{a_n, n \in \mathbb{N}\}$ and rate function

$$
I_\Phi(x) = I(x) - \Phi(x) - \inf_{y \in \mathcal{F}} \{ I(y) - \Phi(y) \}.
$$
Comments on the Proof. (a) This is proved in Varadhan (1966) and also in Section B.1 of Appendix B in Ellis (1985). The result is easy to motivate if we substitute the notation

\[ P_n \{ W_n \in dx \} \leq e^{-a_n(x)} \, dx \]

into the integral

\[ \int_x \exp[a_n \Phi(x)] P_n \{ W_n \in dx \}. \]

Since as \( n \to \infty \) the main contribution to the integral comes from the largest value of the integrand, the limit in part (a) follows.

(b) This is proved in Section B.2 of Appendix B in Ellis (1985). Again the result is easy to motivate if we use the notation

\[ P_n \{ W_n \in dx \} \leq e^{-a_n(x)} \, dx, \]

which leads us to investigate the “measures”

\[ \tilde{Q}_{n, \Phi}(dx) = \frac{1}{Z_n} \exp[a_n(\Phi(x) - I(x))] \, dx. \] (3.2)

By part (a) of the present proposition

\[ \lim_{n \to \infty} \frac{1}{a_n} \log Z_n = \lim_{n \to \infty} \frac{1}{a_n} \log \int_x \exp[a_n \Phi(x)] P_n \{ W_n \in dx \} \]

\[ = \sup_{x \in x} \{ \Phi(x) - I(x) \} \]

\[ = - \inf_{x \in x} \{ I(x) - \Phi(x) \}. \]

Substituting into equation (3.2) the asymptotic formula

\[ Z_n \leq \exp \left[ -a_n \inf_{x \in x} \{ I(x) - \Phi(x) \} \right] \]

leads to the conclusion of part (b); namely,

\[ Q_{n, \Phi}(dx) \leq \exp[-a_n I_\Phi(x)] \, dx. \]

Our comments on the proof of Proposition 3.4 are complete. \( \square \)

We now turn to the statement of Cramér’s Theorem for the sample means of i.i.d. random vectors.

4. CRAMÉR’S THEOREM

Let \( \{ X_j, j \in \mathbb{N} \} \) be a sequence of i.i.d. random vectors taking values in \( \mathbb{R}^d \), where \( d \) is a positive integer. We are interested in the large deviation principle for the sample means \( \{ S_n/n, n \in \mathbb{N} \} \), 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 \( t \in \mathbb{R}^d \). We define for \( t \in \mathbb{R}^d \) the
finite convex function

\[ c(t) = \log E\{\exp\langle t, X_1 \rangle\} \]  

(4.1)

and for \( x \in \mathbb{R}^d \) the Legendre-Fenchel transform

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

(4.2)

In these formulas \( \langle \cdot, \cdot \rangle \) denotes the Euclidean inner product on \( \mathbb{R}^d \).

We first consider the case \( d = 1 \). Let \( \alpha \) be a real number exceeding the mean value \( E\{X_1\} \). Assuming that \( \rho \) has an absolutely continuous component and that certain other conditions hold, Cramér obtained in his 1938 paper (Cramér, 1958) an asymptotic expansion for the probability \( P\{S_n/n \in [\alpha, \infty)\} \) which implies the large deviation limit

\[ \lim_{n \to \infty} \frac{1}{n} \log P\left\{ \frac{S_n}{n} \in [\alpha, \infty) \right\} = -I(\alpha) = - \inf_{x \in [\alpha, \infty)} I(x). \]

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

**THEOREM 4.1.** 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 \( t \in \mathbb{R}^d \). The following conclusions hold.

(a) The sequence of sample means \( \{S_n/n, n \in \mathbb{N}\} \) satisfies the large deviation principle on \( \mathbb{R}^d \) with rate function \( I(x) \) defined in equation (4.2).

(b) The function \( I \) 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 (Bahadur & Zabell, 1979) and (Donsker & Varadhan, 1976, Sect. 5). We will state the theorem given in the latter reference when we discuss the Donsker-Varadhan theory in Section 7.

Let us return to the large deviation limits for fair coin tossing stated in Theorem 2.2 and for a fair \( q \)-faced die stated in Theorem 2.4. They are immediate consequences of Cramér's Theorem and of the criterion for obtaining large deviation limits given in part (a) of Proposition 3.3.

In the case of fair coin tossing, where \( P\{X_1 = 0\} = P\{X_1 = 1\} = \frac{1}{2} \), we have for each \( t \in \mathbb{R} \)

\[ c(t) = \log E\{\exp[tX_1]\} = \log \left[ \frac{1}{2} (1 + e^t) \right] \]

and for each \( x \in \mathbb{R} \)

\[ I(x) = \sup_{t \in \mathbb{R}} \{ tx - c(t) \} = \begin{cases} x \log(2x) + (1 - x) \log[2(1 - x)] & \text{if } x \in [0, 1] \\ \infty & \text{if } x \in \mathbb{R} \setminus [0, 1]. \end{cases} \]

For \( x \in [0, 1] \) this agrees with the function \( I(x) \) defined in equation (2.4) and appearing in Theorem 2.2.
The fair \( q \)-faced die is defined by the probability distribution \( P\{X_i = i\} = 1/q \) for each \( i \in \{1, 2, \ldots, q\} \). Since the empirical vector \( L^n \) is a sum of the i.i.d. random vectors 
\[
Y_j = (\delta_{o_1}(1), \delta_{o_2}(2), \ldots, \delta_{o_d}(q)), \quad j = 1, 2, \ldots, n,
\]
we have for each \( t = (t_1, t_2, \ldots, t_q) \in \mathbb{R}^q \)
\[
c(t) = E[\exp\langle t, Y_1 \rangle] = \log \left( 1 - \frac{1}{q} \sum_{i=1}^q e^{t_i} \right).
\]
We denote by \( \mathcal{M} \) the subset of \( \mathbb{R}^q \) consisting of probability vectors. As shown in (Ellis, 1985, p. 252), we have for each \( v = (v_1, v_2, \ldots, v_q) \in \mathbb{R}^q \)
\[
I(v) = \sup_{t \in \mathbb{R}^q} \{ \langle t, v \rangle - c(t) \} = \begin{cases} 
\frac{1}{q} \sum_{i=1}^q v_i \log(v_i) & \text{if } v \in \mathcal{M} \\
\infty & \text{if } v \in \mathbb{R}^q \setminus \mathcal{M}.
\end{cases}
\]
For \( v \in \mathcal{M} \) this agrees with the relative entropy \( I(v \mid \mu^*) \) appearing in Theorem 2.4.

In a moment we will sketch the proof of Theorem 4.1. A complete proof is given, for example, in Chapters 3 and 4 of Varadhan (1984). The properties of \( I \) stated in part (b) of the theorem as well as other properties are proved in Ellis (1985, Thm. VII.5.5). Before sketching the proof, it is worthwhile to motivate the form of the rate function \( I \), which, we recall, is defined to be the Legendre-Fenchel transform of the finite convex function \( c(t) \) given in equation (4.1). Assuming that the sequence \( \{S_n/n, n \in \mathbb{N}\} \) satisfies the large deviation principle on \( \mathbb{R}^d \) with some convex, lower semicontinuous rate function \( J \), we will prove that \( J = I \). A similar calculation will yield the form of the rate function in part (a) of Theorem 11.1, which is a large deviation result for the Ising model in statistical mechanics.

Although it is not necessary, it will simplify the discussion if we assume that the common support of the random vectors \( \{X_n, n \in \mathbb{N}\} \) is bounded. Then there exists a number \( M < \infty \) such that for all \( n \in \mathbb{N} \), \( \|S_n/n\| \leq M \) with probability 1. By the large derivation lower bound
\[
\liminf_{n \to \infty} \frac{1}{n} \log P\{\|S_n/n\| > M\} \geq - \inf_{\{x \in \mathbb{R}^d, \|x\| > M\}} J(x).
\]
Since the probability \( P\{\|S_n/n\| > M\} \) equals 0, it follows that
\[
\inf_{\{x \in \mathbb{R}^d, \|x\| > M\}} J(x) = \infty
\]
and thus that \( J(x) = \infty \) for every \( x \in \mathbb{R}^d \) satisfying \( \|x\| > M \). Given \( t \in \mathbb{R}^d \), we choose \( \Phi_t \) to be any bounded continuous function mapping \( \mathbb{R}^d \) in \( \mathbb{R} \) such that \( \Phi_t(x) = \langle t, x \rangle \) for every \( x \in \mathbb{R}^d \) satisfying \( \|x\| \leq M \). Then for each \( n \in \mathbb{N} \)
\[
c(t) = \log E[\exp\langle t, X_1 \rangle] = \frac{1}{n} \log E[\exp n \langle t, S_n/n \rangle] \\
= \frac{1}{n} \log \int_{\|x\| \leq M} \exp[n \langle t, x \rangle] P\{S_n/n \in dx\} \\
= \frac{1}{n} \log \int_{\mathbb{R}^d} \exp[n \Phi_t(x)] P\{S_n/n \in dx\},
\]
and so
\[
c(t) = \lim_{n \to \infty} \frac{1}{n} \log \int_{\mathbb{R}^d} \exp[n \Phi_t(x)] P\{S_n \in dx\}.
\]

Using part (a) of Proposition 3.4 and the assumed large deviation principle for \(\{S_n/n, n \in \mathbb{N}\}\) with rate function \(J\), we obtain the variational formula
\[
c(t) = \sup_{x \in \mathbb{R}^d} \left\{ \langle t, x \rangle - J(x) \right\}.
\]

However, \(J(x) = \infty\) for every \(x \in \mathbb{R}^d\) satisfying \(\|x\| > M\), and for every \(x \in \mathbb{R}^d\) satisfying \(\|x\| \leq M \Phi_t(x) = \langle t, x \rangle\). Hence
\[
c(t) = \sup_{x \in \mathbb{R}^d} \{ \langle t, x \rangle - J(x) \}.
\]

The assumed convexity and lower semicontinuity of \(J\) combined with Legendre-Fenchel duality now yields the desired formula; namely, for each \(x \in \mathbb{R}^d\)
\[
J(x) = \sup_{t \in \mathbb{R}^d} \{ \langle t, x \rangle - c(t) \} = I(x).
\]

Legendre-Fenchel duality is explained, for example, in Ellis (1985, Sect. VI.5). This completes the motivation of the form of the rate function in Cramér’s Theorem.

We end this section by sketching 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 (1952) 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 (Cramér, 1938). These same tools for proving the large deviation bounds continue to be used in modern developments of the theory.

**Sketch of the Proof of Theorem 4.1.** We first show that \(I\) is a rate function, then sketch the proof of part (b) followed by sketches of 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. Hence Legendre-Fenchel duality yields the formula
\[
c(t) = \sup_{x \in \mathbb{R}^d} \{ \langle t, x \rangle - I(x) \},
\]
valid for all \(t \in \mathbb{R}^d\). Substituting \(t = 0\) gives
\[
0 = c(0) = \sup_{x \in \mathbb{R}^d} \{ -I(x) \} = - \inf_{x \in \mathbb{R}^d} I(x).
\]

Thus 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 \(K_L = \{x \in \mathbb{R}^d: I(x) \leq L\}\), 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 \(t \in \mathbb{R}^d\)
\[
\langle t, x \rangle \leq c(t) + I(x) \leq c(t) + L.
\]
Fix any positive number $R$. The finite convex function $c$ is bounded on the ball of radius $R$ with center 0, and so there exists a number $\Gamma < \infty$ such that

$$\sup_{|x| \leq R} |c(x)| \leq \sup_{|x| \leq R} c(x) + L \leq \Gamma < \infty.$$ 

This implies that $K_\alpha$ 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 $\mathbb{R}^d$ (part (b) of Proposition 3.3). To show that $x_0$ is unique and equals the mean value $E\{X_1\}$ requires some additional ideas from convex analysis (Ellis, 1985, Thm. VII.5.5).

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

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

If $F$ contains the unique minimum point $E\{X_1\}$ of $I$, then $I(F) = 0$ and the large deviation upper bound is automatic. So we assume that $E\{X_1\} \notin F$. Because of the properties of $I$ given in part (b) of the theorem, it suffices to prove the large deviation upper bound for the closed interval $[-\infty, \beta]$, where $\beta > E\{X_1\}$, and for the closed interval $(\beta, \infty)$, where $\beta < E\{X_1\}$. We carry out the first one; the proof of the second is similar.

By Chebyshev's Inequality, we have for any $t > 0$

$$P\{S_n \in [\beta, \infty)\} = P\{tS_n \geq t\beta\}$$

$$\leq \exp[-nt\beta] \cdot E\{\exp[tS_n]\} = \exp[-nt\beta] \cdot \prod_{i=1}^{n} E\{\exp[tX_i]\} = \exp[-nt(\beta - c(t))].$$

It follows that for any $t > 0$

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

and thus that

$$\limsup_{n \to \infty} \frac{1}{n} \log P\{S_n \in [\beta, \infty)\} \leq -\sup_{t > 0} \{t\beta - c(t)\}.$$

It is easily checked that since $\beta > E\{X_1\} = c'(0)$,

$$\sup_{t > 0} \{t\beta - c(t)\} = \sup_{t \in \mathbb{R}} \{t\beta - c(t)\} = I(\beta)$$

and that since $I$ is convex and has a unique minimum point at $E\{X_1\}$

$$I(\beta) = \inf_{t \in [0, \beta]} I(x).$$

We have proved the large deviation upper bound.
\[
\limsup_{n \to \infty} \frac{1}{n} \log P\{S_n/n \in [\beta, \infty)\} \leq - \inf_{x \in [\beta, \infty)} I(x).
\]

The sketch of the proof of the large deviation upper bound in Theorem 4.1 is complete.

**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 harder than the proof of the large deviation upper bound. For details the reader is referred to Ellis (1985, Ch. VIII).

We denote the common distribution of the random vectors \(\{X_j, j \in \mathbb{N}\}\) by \(\rho(dx) = P\{X_j \in dx\}\).

In general the function
\[
c(t) = \log E[\exp\langle t, X_1 \rangle] = \log \int_{\mathbb{R}^d} \exp\langle t, x \rangle \rho(dx)
\]
is a finite convex differentiable function on \(\mathbb{R}^d\). Although it is not necessary, for simplicity we will assume that the support of \(\rho\) 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 \(\nabla c(t) = z\). For \(z \in \mathbb{R}^d\) and \(\epsilon > 0\), we denote by \(B(z, \epsilon)\) the open ball with center \(z\) and radius \(\epsilon\).

Let \(G\) be an open subset of \(\mathbb{R}^d\). Then for any point \(z_0 \in G\) there exists \(\epsilon > 0\) such that \(B(z_0, \epsilon) \subset G\), and so
\[
P\left\{\frac{S_n}{n} \in G \right\} \geq P\left\{\frac{S_n}{n} \in B(z_0, \epsilon) \right\} = \int_{B(z_0, \epsilon)} \prod_{j=1}^{n} \rho(dx_j).
\]

We choose \(t_0 \in \mathbb{R}^d\) such that \(\nabla c(t_0) = z_0\) and introduce the change of measure given by the exponential family
\[
\rho_{t_0}(dx) = \frac{e^{\langle t_0, x \rangle} \rho(dx)}{e^{c(t_0)}}.
\]

By the definition of \(c(t_0)\) \(\rho_{t_0}\) is a probability measure and the mean of \(\rho_{t_0}\) is \(z_0\). Indeed
\[
\int_{\mathbb{R}^d} x \rho_{t_0}(dx) = \int_{\mathbb{R}^d} x \frac{e^{\langle t_0, x \rangle}}{e^{c(t_0)}} \rho(dx) = \nabla c(t_0) = z_0.
\]

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

Thus
\[
P\left\{\frac{S_n}{n} \in G \right\} \geq \int_{B(z_0, \epsilon)} \prod_{j=1}^{n} \frac{dp(x_j)}{dp_{t_0}} \cdot \prod_{j=1}^{n} \rho_{t_0}(dx_j)
\]
\[
= \int_{B(z_0, \epsilon)} \exp\left[-n\left(\langle t_0, \sum_{j=1}^{n} x_j/n \rangle - c(t_0)\right)\right] \prod_{j=1}^{n} \rho_{t_0}(dx_j)
\]
\[
\geq \exp\left[-n\langle t_0, z_0 \rangle - c(t_0)\right] \cdot \int_{B(z_0, \epsilon)} \prod_{j=1}^{n} \rho_{t_0}(dx_j).
\]
Since the mean of the probability measure $\rho_{z_0}$ equals $z_0$, the weak law of large numbers for i.i.d. random vectors with common distribution $\rho_{z_0}$ implies that

$$\lim_{n \to \infty} \int_{\bigotimes_{i=1}^n A_i} \prod_{j=1}^n \rho_{z_0}(dx_j) = 1.$$ 

Hence

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

$$= -I(z_0) - \| t_0 \| \varepsilon.$$ 

We now send $\varepsilon \to 0$, and since $z_0$ is an arbitrary point in $G$, we can replace $- I(z_0)$ by $- I(G)$. The sketch of the proof of the large deviation lower bound is complete. 

In the next section, Section 5, we present a generalization of Cramér’s Theorem which 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, in Section 6 we present two examples of large deviation principles in which the rate function is not convex, and we suggest a general formalism for proving the large deviation principle which may be effective in a class of such cases. In Section 7 we briefly discuss the wide-ranging generalizations of Cramér’s Theorem which were discovered by Donsker and Varadhan. The remainder of the paper treats applications to statistical mechanics. In discussing these applications, we will use the material in both Sections 5 and 7.

5. THE GÄRTNER-ELLIS THEOREM

Let $\{a_n, n \in \mathbb{N}\}$ be a sequence of positive numbers tending to $\infty$, $(\Omega_n, \mathcal{F}_n, P_n)$, $n \in \mathbb{N}$ a sequence of probability spaces, and for each $n \in \mathbb{N}$, $W_n$ a random vector mapping $\Omega_n$ into $\mathbb{R}^d$. In 1977 Gärtner proved an important generalization of Cramér’s Theorem, assuming only that the limit

$$c(t) \doteq \lim_{n \to \infty} \frac{1}{a_n} \log E_n \{\exp[a_n \langle t, W_n \rangle]\} \quad (5.1)$$

exists and is finite for every $t \in \mathbb{R}^d$ and that $c(t)$ is a differentiable function of $t \in \mathbb{R}^d$ (Gärtner, 1977). Gärtner’s result is that the sequence $\{W_n, n \in \mathbb{N}\}$ satisfies the large deviation principle on $\mathbb{R}^d$ with rate function equal to the Legendre-Fenchel transform

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

and with norming constants $\{a_n, n \in \mathbb{N}\}$. Using ideas from convex analysis, I generalized in 1984 this result of Gärtner by relaxing the condition that $c(t)$ exist.
and be finite for every $t \in \mathbb{R}^d$ (Ellis, 1985). The theorem is now known in the literature as the Gärtner-Ellis Theorem (Bucklew, 1990; Dembo & Zeitouni, 1993). In order to see that this theorem contains Cramér’s Theorem as a special case, let $W_n$ be the $n$-th sample mean of a sequence $\{X_j, j \in \mathbb{N}\}$ 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 equation (5.1) 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 Cramér’s Theorem. In Section 11 we will apply the Gärtner-Ellis Theorem to a problem in statistical mechanics involving dependent random variables.

We next state the Gärtner-Ellis Theorem under the hypotheses of Gärtner (1977).

**Theorem 5.1.** Let $\{a_n, n \in \mathbb{N}\}$ be a sequence of positive numbers tending to $\infty$, $\{\Omega_n, \mathcal{F}_n, P_n, n \in \mathbb{N}\}$ a sequence of probability spaces, and for each $n \in \mathbb{N}$ $W_n$ a random vector mapping $\Omega_n$ into $\mathbb{R}^d$. We assume that the limit

$$c(t) = \lim_{n \to \infty} \frac{1}{a_n} \log E_n[\exp(a_n \langle t, W_n \rangle)]$$

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

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

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 $\mathbb{R}^d$

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

(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 $\mathbb{R}^d$

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

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

The theorem is proved by suitably generalizing the proof of Cramér’s Theorem (see Ellis, 1985, Ch. 7). Since in the Gärtner-Ellis 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 in the next section.

**6. LARGE DEVIATION PRINCIPLES HAVING NONCONVEX RATE FUNCTIONS**

In Cramér’s Theorem and, more generally, in the Gärtner-Ellis Theorem the rate function is defined as a Legendre-Fenchel transform. Hence the rate function is
always convex and lower semicontinuous. In this section we present two examples of large deviation principles having nonconvex rate functions. Additional examples appear in Sections 9 and 10 and in the paper (Dinwoodie & Zabell, 1992). The present section is not used later in the paper and may be omitted.

One of the hypotheses of the Gärtner-Ellis Theorem is the differentiability of the limit function \( c(t) \). A good starting point is to investigate the existence of large deviation principles when this condition is violated. Unfortunately, the situation is complicated and a general theory does not exist. In the first example to be presented in this section, the differentiability of the limit function \( c(t) \) does not hold and the rate function is not given by a Legendre-Fenchel transform. In Theorem 11.1 in Section 11, which concerns the Ising model in statistical mechanics, the same hypothesis of the Gärtner-Ellis Theorem is violated for a certain range of a parameter \( \beta \) but the rate function is defined by the identical Legendre-Fenchel transform appearing in the statement of the Gärtner-Ellis Theorem. We end this section by introducing a generalization of the Legendre-Fenchel transform which may be useful in proving the large deviation principle in certain cases when the Gärtner-Ellis Theorem cannot be applied.

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

**EXAMPLE 6.1**: Let \( X_1 \) be a random variable with 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

\[
W_n = \frac{1}{n} \sum_{j=1}^{n} X_j.
\]

Let us first try to apply the Gärtner-Ellis Theorem to the sequence \( \{W_n, n \in \mathbb{N}\} \). For each \( t \in \mathbb{R} \) and \( x \in \mathbb{R} \) we calculate

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

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, the sequence \( \{W_n, n \in \mathbb{N}\} \) satisfies the large deviation principle on \( \mathbb{R} \) with the rate function

\[
J(x) = \begin{cases} 0 & \text{if } x \in \{1, -1\} \\ \infty & \text{if } x \in \mathbb{R} \setminus \{1, -1\}. \end{cases}
\]  \hspace{1cm} (6.1)

This is easily checked since for each \( n \in \mathbb{N} \) \( W_n \) 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 \( J \). We will return to this example at the end of this section when we point out a generalization of the Legendre-Fenchel transform that gives the rate function \( J \). \qed
Our second example is a generalization of Cramér's Theorem involving a random walk with an interface.

**EXAMPLE 6.2.** We define the sets
\[ \Lambda^{(1)} = \{ x \in \mathbb{R}^d : x_1 \leq 0 \}, \quad \Lambda^{(2)} = \{ x \in \mathbb{R}^d : x_1 > 0 \} \quad \text{and} \quad \partial = \{ x \in \mathbb{R}^d : x_1 = 0 \}, \]
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 \( \mu^{(1)} \) and \( \mu^{(2)} \) 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 \( \{ X^{(1)}_n, n \in \mathbb{N} \} \) and \( \{ X^{(2)}_j, j \in \mathbb{N} \} \) be two sequences of i.i.d. random vectors with probability distributions
\[ P\{ X^{(1)}_n \in dx \} = \mu^{(1)}(dx) \quad \text{and} \quad P\{ X^{(2)}_j \in dx \} = \mu^{(2)}(dx). \]

We consider the stochastic process \( \{ S_n, n \in \mathbb{N} \cup \{0\} \} \), where \( S_0 = 0 \) and \( S_n \) is defined recursively by the formula
\[ S_{n+1} = S_n + 1_{\{ S_n \in \Lambda^{(1)} \}} \cdot X^{(1)}_n + 1_{\{ S_n \in \Lambda^{(2)} \}} \cdot X^{(2)}_n. \]

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 interface \( \partial \), we call this random walk model a model with "discontinuous statistics." In the paper (Dupuis & Ellis, 1992) we show that the sequence \( \{ S_n/n, n \in \mathbb{N} \} \) 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 \( \partial \). 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 \( \mu^{(1)} \) and \( \mu^{(2)} \) coincide, then the main theorem of Dupuis & Ellis (1992) reduces to Cramèr's Theorem.

The large deviation phenomena investigated in the paper (Dupuis & Ellis, 1992) are an example of the fascinating problems that arise in the study of other Markov processes with "discontinuous statistics." A generalization of the main theorem of Dupuis & Ellis (1992), which is a large deviation principle for the entire path of the random walk, is proved in Chapter 6 of the book (Dupuis & Ellis, 1995a). This book presents a new technique for proving large deviation principles which is based on the theory of weak convergence of probability measures and on ideas from stochastic optimal control theory. The paper (Dupuis et al., 1991) proves a large deviation upper bound 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 paper (Dupuis & Ellis, 1995b) proves the large deviation principle for a general class of such systems.

We return to Example 6.1, in which the rate function \( J \) of the sequence \( \{ W_n, n \in \mathbb{N} \} \) is defined in equation (6.1). Let us note a formula for \( J \) which is based on a generalization of the Legendre-Fenchel transform. This formula was pointed out by Paul Dupuis. The generalization replaces the linear "pivot" function \( tx \) in the Legendre-Fenchel transform on \( \mathbb{R} \) by a piecewise linear function that is linear.
on each of the halflines \( \{ x \in \mathbb{R} : x \leq 0 \} \) and \( \{ x \in \mathbb{R} : x \geq 0 \} \). Let \( \alpha \) and \( \beta \) be arbitrary real numbers and for \( x \in \mathbb{R} \) define the function
\[
f_{\alpha \beta}(x) = \begin{cases} 
\alpha x & \text{if } x \leq 0 \\
\beta x & \text{if } x \geq 0.
\end{cases}
\]
We then define
\[
h(\alpha, \beta) = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E} \{ \exp[nf_{\alpha \beta}(W_n)] \} = \lim_{n \to \infty} \frac{1}{n} \log \left( \frac{1}{2} \left[ e^{-n} + e^{n\beta} \right] \right) = (-\alpha) \vee \beta.
\]
For \( x \in \mathbb{R} \) we have the calculation
\[
\sup_{\alpha, \beta \in \mathbb{R}} \{ f_{\alpha \beta}(x) - h(\alpha, \beta) \} = \begin{cases} 
0 & \text{if } x \in \{1, -1\} \\
\infty & \text{if } x \in \mathbb{R} \setminus \{1, -1\}
\end{cases} = J(x),
\]
which gives the rate function \( J \).

This calculation suggests a methodology for treating large deviation problems that cannot be treated by the Görnter-Ellis Theorem.

CONJECTURED METHODOLOGY. We consider the case of random variables taking values in \( \mathbb{R} \). An analogous methodology can be conjectured for random vectors taking values in \( \mathbb{R}^d \). Let \( \{a_n, n \in \mathbb{N}\} \) be a sequence of positive numbers tending to \( \infty \), \( \{\Omega_n, \mathcal{F}_n, P_n\}, n \in \mathbb{N}\) a sequence of probability spaces, and for each \( n \in \mathbb{N} \) \( W_n \) a random variable mapping \( \Omega_n \) into \( \mathbb{R} \). We assume that the limit
\[
h(\alpha, \beta) = \lim_{n \to \infty} \frac{1}{a_n} \log \mathbb{E}_n \{ \exp[a_n f_{\alpha \beta}(W_n)] \}
\]
exists and is finite for every \( x \in \mathbb{R} \) and \( \beta \in \mathbb{R} \). In this case, the function \( h \) is a convex, lower semicontinuous function on \( \mathbb{R}^2 \). For \( x \in \mathbb{R} \) we define the generalized Legendre-Fenchel transform
\[
J(x) = \sup_{\alpha, \beta \in \mathbb{R}} \{ f_{\alpha \beta}(x) - h(\alpha, \beta) \}. \tag{6.3}
\]
This is a lower semicontinuous function on \( \mathbb{R} \) which is convex on each of the halflines \( \{ x \in \mathbb{R} : x \leq 0 \} \) and \( \{ x \in \mathbb{R} : x \geq 0 \} \). We call such a function "halfline convex." The goal is to find conditions on \( h(x, \beta) \) ensuring that the random variables \( \{W_n, n \in \mathbb{N}\} \) satisfy the large deviation principle on \( \mathbb{R} \) with rate function \( J \) and normalizing constants \( \{a_n, n \in \mathbb{N}\} \). This is an open problem.

I am led to conjecture this methodology both because of the calculation, just given, of the rate function \( J \) in Example 6.1 and because the generalized Legendre-Fenchel transform satisfies a duality property that is analogous to one that holds for the standard Legendre-Fenchel transform. The latter property was used in Section 4 just before the sketch of the proof of Theorem 4.1. The analogous duality property satisfied by the generalized Legendre-Fenchel transform is given in the paper (Dupuis et al., 1995), which also treats functions on \( \mathbb{R}^d \).
7. THE DONSKER-VARADHAN THEORY

In a series of papers (Donsker & Varadhan, 1975a; 1975b; 1976; 1983) beginning in 1976, Donsker and Varadhan discovered a vast new array of large deviation phenomena, of which Cramér’s Theorem is one component. Their theory applies to discrete-time Markov chains and continuous-time Markov processes taking values in a complete separable metric space. In summarizing their theory, I will point out a central feature that is needed in our applications to models in statistical mechanics given in Sections 9, 10, and 11. This feature is the existence of three levels of large deviations; namely, sample means, empirical measures, and empirical processes.

In order to ease the exposition, I will consider the special case of the Donsker-Varadhan theory which concerns a sequence of i.i.d. random variables \( \{X_j, j \in \mathbb{N}\} \) taking values in a complete separable metric space \( \mathcal{X} \). We denote by \( \rho \) the common distribution of the random variables \( \{X_j, j \in \mathbb{N}\} \).

LEVEL-1. This is the level of the sample means

\[
S_n = \frac{1}{n} \sum_{j=1}^{n} X_j,
\]

for which Donsker and Varadhan prove the following infinite-dimensional analogue of Cramér’s Theorem (Donsker & Varadhan, 1976, Thm. 5.2).

THEOREM 7.1. 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 variables taking values in \( \mathcal{X} \) and having common distribution \( \rho \). Assume that \( E[\exp(t\|X_1\|)] < \infty \) for every \( t > 0 \). Then the sequence \( \{S_n, n \in \mathbb{N}\} \) 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 \( x = E\{X_1\} = \int_{\mathcal{X}} y \rho(dy) \).

LEVEL-2. Level-2 is the level of the empirical measures. For \( A \) a Borel subset of the complete separable metric space \( \mathcal{X} \) and \( x \) a point in \( \mathcal{X} \), we define the unit point measure \( \delta_x, \{A\} \) to be 1 if \( x \in A \) and 0 if \( x \in \mathcal{X} \setminus A \). Let \( (\Omega, \mathcal{F}, P) \) be the probability space on which the i.i.d. sequence \( \{X_j, j \in \mathbb{N}\} \) is defined. For \( n \in \mathbb{N} \) and \( \omega \in \Omega \) we define the empirical measure

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

We denote by \( \mathcal{P}(\mathcal{X}) \) the space of probability measures on \( \mathcal{X} \). Let \( \{\mu_n, n \in \mathbb{N}\} \) be a sequence in \( \mathcal{P}(\mathcal{X}) \) and \( \mu \) a measure in \( \mathcal{P}(\mathcal{X}) \). We say that the sequence \( \{\mu_n, n \in \mathbb{N}\} \) converges weakly to \( \mu \), and we write \( \mu_n \rightharpoonup \mu \), if for each bounded continuous function \( f \) mapping \( \mathcal{X} \) into \( \mathbb{R} \)

\[
\lim_{n \to \infty} \int_{\mathcal{X}} f(x) \mu_n(dx) = \int_{\mathcal{X}} f(x) \mu(dx).
\]
\( \mathcal{P}(\mathcal{X}) \) can be metrized to be a complete separable metric space, and the metric topology coincides with the topology corresponding to weak convergence of probability measures on \( \mathcal{X} \). For each \( \omega \in \Omega \) \( L_n(\omega, \cdot) \) takes values in \( \mathcal{P}(\mathcal{X}) \).

Let \( f \) be any bounded continuous function mapping \( \mathcal{X} \) into \( \mathbb{R} \). Then for each \( \omega \in \Omega \)
\[
\int_{\mathcal{X}} f(x)L_n(\omega, dx) = \frac{1}{n} \sum_{j=1}^{n} f(X_j),
\]
and so by the law of large numbers for the i.i.d. sequence \( \{f(X_j), j \in \mathbb{N}\} \) we have
\[
P\left\{ \omega \in \Omega : \lim_{n \to \infty} \int_{\mathcal{X}} f(x)L_n(\omega, dx) = \int_{\mathcal{X}} f(x)\rho(dx) \right\} = 1.
\]
In general the \( P \)-null set on which this convergence fails depends on \( f \). By means of a separability argument one shows that there exists a \( P \)-null set \( \mathcal{A} \) such that for all bounded continuous functions \( f \) mapping \( \mathcal{X} \) into \( \mathbb{R} \) and for all points \( \omega \in \Omega \setminus \mathcal{A} \)
\[
\lim_{n \to \infty} \int_{\mathcal{X}} f(x)L_n(\omega, dx) = \int_{\mathcal{X}} f(x)\rho(dx).
\]
It follows that
\[
P\left\{ \omega \in \Omega : L_n(\omega, \cdot) \Rightarrow \rho \right\} = 1.
\]
This leads to the problem of investigating the large deviations of the sequence of empirical measures \( \{L_n, n \in \mathbb{N}\} \) away from the measure \( \rho \).

The next theorem formulates the large deviation principle for this sequence. In the case \( \mathcal{X} = \mathbb{R} \) it is due to Sanov [42]. For \( \nu \) a probability measure on \( \mathcal{X} \) we define the relative entropy of \( \nu \) with respect to \( \rho \) by the formula
\[
I(\nu \mid \rho) = \int_{\mathcal{X}} \left( \log \frac{d\nu}{d\rho}(x) \right) \nu(dx)
\]
whenever \( \nu \) is absolutely continuous with respect to \( \rho \) and \( \log(d\nu/d\rho(x)) \) is \( \nu \)-integrable. In all other cases, we set \( I(\nu \mid \rho) = \infty \).

**Theorem 7.2.** Let \( \{X_j, j \in \mathbb{N}\} \) be a sequence of i.i.d. random variables taking values in a complete separable metric space \( \mathcal{X} \) and having common distribution \( \rho \). Then the sequence \( \{L_n, n \in \mathbb{N}\} \) of the empirical measures of \( \{X_j, j \in \mathbb{N}\} \) satisfies the large deviation principle on \( \mathcal{X} \) with rate function equal to the relative entropy \( I(\cdot \mid \rho) \). The relative entropy \( I(\nu \mid \rho) \) is a convex, lower semicontinuous function on \( \mathcal{P}(\mathcal{X}) \), and it attains its infimum of 0 at the unique measure \( \nu = \rho \).

Since the sequence of empirical measures \( \{L_n, n \in \mathbb{N}\} \) is the sequence of sample means of the i.i.d. random variables \( \{\delta_{X_j}, j \in \mathbb{N}\} \), it is not unreasonable to suppose that Theorem 7.2 can be derived as a consequence of a suitable infinite-dimensional version of Cramér's Theorem. This is carried out in Deuschel & Stroock (1989, Thm. 3.2.17) and in Stroock (1984, Lemma 3.37–3.38). Theorem 2.4, which treats the fair \( q \)-faced die, is the special case of Theorem 7.2 for \( \mathcal{X} \) the finite set \( \{1, 2, \ldots, q\} \). Our discussion of level-2 of the Donsker-Varadhan theory is complete.
LEVEL-3. Natural extensions of the empirical measures are the bivariate empirical measures

\[ \frac{1}{n} \sum_{j=1}^{n} \delta_{(x_j, x_{j+1})}, \]

the trivariate empirical measures

\[ \frac{1}{n} \sum_{j=1}^{n} \delta_{(x_j, x_{j+1}, x_{j+2})}, \]

and the \( z \)-variate analogues for \( z \in \{4, 5, \ldots\} \). It is useful to modify these definitions slightly so that for each \( z \geq 2 \) the first and \( z \)-th marginals of the \( z \)-variance empirical measure both equal the \((z-1)\)-variate empirical measure. Thus we introduce the symmetric bivariate empirical measure

\[ \frac{1}{n} \sum_{j=1}^{n-1} \delta_{(x_j, x_{j+1})} + \frac{1}{n} \delta_{(x_n, x_1)}, \]

the symmetric trivariate measures

\[ \frac{1}{n} \sum_{j=1}^{n-2} \delta_{(x_j, x_{j+1}, x_{j+2})} + \frac{1}{n} \delta_{(x_{n-1}, x_n, x_1)} + \frac{1}{n} \delta_{(x_n, x_1, x_2)}, \]

and the symmetric \( z \)-variate analogues for \( z \in \{4, 5, \ldots\} \). Instead of considering each of these cases separately, Donker and Varadhan proposed an infinite dimensional extension of the symmetric \( z \)-variate empirical measures known as the empirical processes. For each \( z \in \mathbb{N} \) exceeding 1, the \( z \)-dimensional marginal of the empirical process is the symmetric \( z \)-variate empirical measure while its one-dimensional marginal is the empirical measure. Hence from the large deviation principle for the empirical processes, one may derive the large deviation principle for the symmetric \( z \)-variate empirical measures by means of the contraction principle (part (c) of Proposition 3.3).

Level-3 is the level of the empirical processes, which we now define. Let \( \mathcal{F} \) be a complete separable metric space, \( \{X_j, j \in \mathbb{N}\} \) a sequence of i.i.d. random variables taking values in \( \mathcal{F} \) and having the common distribution \( \rho \), and \( T \) the shift mapping on the sequence space \( \mathcal{F}^\mathbb{N} \). We denote by \( \Omega \) the space on which the sequence \( \{X_j, j \in \mathbb{N}\} \) is defined. For each \( n \in \mathbb{N} \) and \( \omega \in \Omega \) we repeat the sequence \( X_1(\omega), X_2(\omega), \ldots, X_n(\omega) \) periodically into a doubly infinite sequence, obtaining a point \( X(n, \omega) \in \mathcal{F}^\mathbb{Z} \). For Borel subsets \( A \) of \( \mathcal{F}^\mathbb{Z} \) we define the empirical process

\[ R_n(A) = R_n(\omega, A) = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{T^k X(n, \omega)}(A), \]

where \( T^0 \) is the identity mapping and \( T^k = T(T^{k-1}) \). We denote by \( \mathcal{P}(\mathcal{F}) \) the space of strictly stationary probability measures on \( \mathcal{F}^\mathbb{Z} \). This space is a closed subset of the space \( \mathcal{P}(\mathcal{F}^\mathbb{Z}) \) of probability measures on \( \mathcal{F}^\mathbb{Z} \) and so can be metrized as a complete separable metric space. Since for each \( \omega \in \Omega \) \( X(n, \omega) \) is periodic of period \( n \), \( R_n(\omega, \cdot) \) takes value in \( \mathcal{P}(\mathcal{F}) \).
As a consequence of the ergodic theorem and a separability argument, one proves that with probability 1 the sequence \( \{R_n, n \in \mathbb{N}\} \) converges weakly to the infinite produce measure with identical one-dimensional marginals \( \rho \). We denote this measure by the symbol \( P_\rho \). This leads to the problem of investigating the large deviations of the sequence of empirical processes \( \{R_n, n \in \mathbb{N}\} \) away from the measure \( P_\rho \). The next theorem states the large deviation principle for this sequence. The definition of the rate function is somewhat complicated, and since its explicit form is not needed in the sequel, we will omit it.

**Theorem 7.3.** Let \( \{X_j, j \in \mathbb{N}\} \) be a sequence of i.i.d. random variables taking values in a complete separable metric space \( \mathcal{X} \) and having common distribution \( \rho \). Then the sequence \( \{R_n, n \in \mathbb{N}\} \) of the empirical processes of \( \{X_j, j \in \mathbb{N}\} \) satisfies the large deviation principle on \( \mathcal{P}(\mathcal{X}^n) \) with a rate function that is convex and lower semicontinuous and attains its infimum of 0 at the unique measure \( P_\rho \).

Let \( \{X(t), t \in [0, \infty)\} \) be a continuous-time Markov process taking values in a complete separable metric space. For such a process one defines the analogue of the empirical process. The corresponding large deviation principle is proved in Donsker & Varadhan (1983); see also Sections 9–12 of Varadhan (1984). Theorem 7.3 is proved in Olla (1988) using similar ideas. In the special case where \( \mathcal{X} \) is a finite set, the theorem is proved in Ellis (1985, Ch. 9). When we study the Ising model in Section 11, we will need the large deviation principle for a multidimensional extension of the empirical processes known as the empirical fields. Our discussion of level-3 of the Donsker-Varadhan theory is now complete.

The remaining sections of this paper treat applications to statistical mechanics.

**8. Spin Systems and Phase Transitions in Statistical Mechanics**

One of the most interesting problems in equilibrium statistical mechanics is to explain phase transitions in terms of the probability distributions on configuration space which describe the microscopic behavior of physical systems. The simplest systems for which this is possible are ferromagnetic models on a lattice. We will devote this section and the next three sections to the study of these models. Phase transitions arise as a result of two competing microscopic effects. The first effect, which tends to order the system, is caused by attractive forces of interaction and is measured by energy. The second effect, which tends to randomize the system, is caused by thermal excitations and is measured by entropy. At sufficiently low temperatures the energy effect predominates and a phase transition becomes possible. All the models that we study exhibit phase transitions in the manner just described.

One of the main tools used to study phenomena in statistical mechanics is probability theory, and the systems that one encounters in statistical mechanics are perfectly suited to large deviation analysis. However, for the newcomer obstacles
arise because of differences in terminology. It is instructive to compile a "dictionary" associating to each concept in statistical mechanics its probabilistic equivalent. Let us start such a dictionary for the spin systems that we will study. Each system is defined on subsets \( \{ \Lambda_n, n \in \mathbb{N} \} \) of the D-dimensional integer lattice \( \mathbb{Z}^D \), the cardinality of which, \( |\Lambda_n| \), tends to \( \infty \) as \( n \to \infty \). The finite-volume Gibbs states of the spin systems are the sequence of probability distribution describing the systems on the subsets \( \{ \Lambda_n, n \in \mathbb{N} \} \). A macroscopic observable is a sequence of random variables \( \{ Y_n, n \in \mathbb{N} \} \) such that with respect to the finite-volume Gibbs states the sequence \( \{ Y_n/|\Lambda_n|, n \in \mathbb{N} \} \) has a limit as \( n \to \infty \). An example of such a \( Y_n \) is the total spin \( S_n \) in the subset \( \Lambda_n \); the quantity \( S_n/|\Lambda_n| \), or the spin per site in \( \Lambda_n \), corresponds to the sample mean in classical probability theory. If the sequence \( \{ Y_n/|\Lambda_n|, n \in \mathbb{N} \} \) satisfies the large deviation principle with some rate function \( I \), then the equilibrium states of the corresponding macroscopic observable are the points at which the rate function \( I \) attains its infimum of 0. The set of these points arises in part (b) of Proposition 3.3 and is denoted by \( \delta \). The finite-volume Gibbs states that we will encounter are all defined in terms of a positive parameter \( \beta \) representing the inverse absolute temperature. In each case there exists a critical value \( \beta_c \) of \( \beta \) satisfying \( \beta_c \in (0, \infty) \). For all \( \beta \in (0, \beta_c) \) the sequence \( \{ S_n/|\Lambda_n|, n \in \mathbb{N} \} \) of spins per site in the subsets \( \{ \Lambda_n, n \in \mathbb{N} \} \) satisfies a law of large numbers. If \( \beta_c < \infty \), then for all \( \beta > \beta_c \) this law of large numbers for the sequence \( \{ S_n/|\Lambda_n|, n \in \mathbb{N} \} \) breaks down. This breakdown corresponds to a phase transition.

In the first two spin systems that we consider, the large deviation principle for the sequence of spins per site is straightforward to derive. By identifying the set of equilibrium states and using part (b) of Proposition 3.3 together with a symmetry argument, we are able to prove the law of large numbers and to calculate the limit that replaces the law of large numbers when it breaks down. This passage from the large deviation principle to the law of large numbers and its breakdown reverses the order in many other probabilistic systems, where the law of large numbers precedes the large deviation principle.

We now turn to our study of three spin systems, the Curie-Weiss model, the Curie-Weiss-Potts model, and the Ising model. Each of these systems will be analyzed using one of the three levels of the Donsker-Varadhan theory: the Curie-Weiss model by a level-1 analysis, in which the basic random quantities are the sequence of sample means (in this case, the spin per site); the Curie-Weiss-Potts model by a level-2 analysis, in which the basic random quantities are the sequence of empirical measures (in this case, the empirical vectors of the spin random variables); and the Ising model by a level-3 analysis, in which the basic random quantities are the sequence of empirical fields of the spin random variables. The empirical fields are multidimensional analogues of the empirical processes introduced in the previous section.

9. THE CURIE-WEISS MODEL
For each \( n \in \mathbb{N} \) we define \( \Lambda_n \) to be the set \( \{1, 2, \ldots, n\} \). The first result in this section is the large deviation principle for the spin per site in the Curie-Weiss
model, which is a spin system on the sequence of subsets \( \{\Lambda_n, n \in \mathbb{N}\} \). The Curie-Weiss model has a critical inverse temperature \( \beta_c \in (0, \infty) \). For all values of the inverse absolute temperature \( \beta \) satisfying \( \beta \in (0, \beta_c] \) the rate function in the large deviation principle is strictly convex on the interval \([0, 1]\) and it has a unique minimum point at \( 0 \). For all values \( \beta > \beta_c \) the rate function is nonconvex and finite on this interval and it has two symmetric minimum points. These properties lead immediately to a law of large numbers for the spin per site when \( \beta \in (0, \beta_c] \) and a replacement limit theorem when \( \beta > \beta_c \) (Theorem 9.2).

The configuration space of the Curie-Weiss model is the set \( \Omega_n = \{1, -1\}^{\Lambda_n} \). \( \Omega_n \) consists of all sequences \( \omega = \{\omega_j, j = 1, 2, \ldots, n\} \), where each \( \omega_j \) represents the value of the spin at the site \( j \in \Lambda_n \). A spin value \( 1 \) corresponds to "spin-up" and a spin value \( -1 \) corresponds to "spin-down." As in the case of fair coin tossing, we define for each \( \omega \in \Omega_n \) the quantity \( P_n\{\omega\} = 2^{-n} \) and extend this in the usual way to a probability measure \( P_n \) on the set \( \mathcal{F} \) of all subsets of \( \Omega_n \). With respect to \( P_n \), the coordinates \( \{\omega_j, j \in \Lambda_n\} \) are i.i.d. with common distribution \( \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1} \).

In order to complete the definition of the Curie-Weiss model we define for each \( \omega \in \Omega_n \) the quantity

\[
H_n(\omega) = -\frac{1}{2n} \sum_{i,j=1}^{n} \omega_i \omega_j = -\frac{n}{2} \left( \frac{1}{n} \sum_{i=1}^{n} \omega_i \right)^2,
\]

which is called the energy or Hamiltonian of the configuration \( \omega \). Let \( \beta \) be a positive parameter representing the inverse absolute temperature. The probability of a configuration \( \omega \in \Omega_n \) corresponding to a given value of \( \beta > 0 \) is defined by the formula

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

(9.1)

The quantity \( Z_n(\beta) \) is a normalization defined by the formula

\[
Z_n(\beta) = \int_{\Omega_n} \exp[-\beta H_n(\omega)] P_n(d\omega)
\]

and called the partition function. \( P_{n, \beta} \) is extended to a probability measure on \( \mathcal{F} \) by defining for any subset \( A \) of \( \Omega_n \)

\[
P_{n, \beta}\{A\} = \sum_{\omega \in A} P_{n, \beta}\{\omega\}.
\]

The resulting probability measure \( P_{n, \beta} \) is called the finite-volume Gibbs state on \( \Lambda_n \).

We recall the definition of weak convergence of probability measures. Let \( \{\mu_n, n \in \mathbb{N}\} \) be a sequence of probability measures on \( \mathbb{R} \) and \( \mu \) a probability measure on \( \mathbb{R} \). We say that the sequence \( \{\mu_n, n \in \mathbb{N}\} \) converges weakly to \( \mu \), and we write \( \mu_n \Rightarrow \mu \), if for each bounded continuous function \( f \) mapping \( \mathbb{R} \) into \( \mathbb{R} \)

\[
\lim_{n \to \infty} \int_{\mathbb{R}} f(x) \mu_n(dx) = \int_{\mathbb{R}} f(x) \mu(dx).
\]
The measure $\mu$ is called the weak limit of the sequence $\{\mu_n, n \in \mathbb{N}\}$. For $n \in \mathbb{N}$ and $\omega \in \Omega_n$ we define the total spin random variable by the formula

$$S_n(\omega) = \sum_{j=1}^{n}(\omega_j).$$

The sample mean $S_n/n$ is called the spin per site. Our goal is to prove for each $\beta > 0$ a law of large numbers, as well as a replacement limit theorem when the law of large numbers breaks, for the sequence $\{S_n/n, n \in \mathbb{N}\}$ of spins per site. We do this by evaluating the weak limit of the sequence of distributions $\{P_{n,\beta}\{S_n/n \in dx\}, n \in \mathbb{N}\}$ as $n \to \infty$. This will be carried out by first deriving the large deviation principle for the sequence $\{P_{n,\beta}\{S_n/n \in dx\}, n \in \mathbb{N}\}$.

In order to motivate the form of the weak limit of the sequence $\{P_{n,\beta}\{S_n/n \in dx\}, n \in \mathbb{N}\}$, let us denote by $\omega^-$ the configuration in $\Omega_n$ in which all the coordinates $(\omega^+)$, equal 1 and by $\omega^-$ the configuration in $\Omega_n$ in which all the coordinates $(\omega^-)$ equal $-1$. The Hamiltonian $H_n(\omega)$ attains its minimum value over $\Omega_n$ at the configuration $\omega = \omega^+$ and, by symmetry, at the configuration $\omega = \omega^-$. $H_n(\omega)$ is a near minimum for configurations in which most of the spin values agree in sign. Correspondingly, the finite-volume Gibbs state $P_{n,\beta}$ assigns maximum probability to the two configurations $\omega^+$ and $\omega^-$ and nearly maximum probability to configurations in which most of the spin values agree in sign. In this sense the finite-volume Gibbs state models a ferromagnet. Since the effect of the Hamiltonian is controlled by the parameter $\beta$, one might expect that a law of large numbers is valid for sufficiently small values of $\beta$, i.e.; that as $n \to \infty$

$$P_{n,\beta}\{S_n/n \in dx\} = \delta_0.$$  

In addition, one might expect that for sufficiently large values of $\beta$ the limit breaks down. We will see that this is indeed the case.

In order to motivate the limit results further, note that for $n \in \mathbb{N}$ and $\beta = 0$ $P_{n,\beta}$ reduces to the fair coin tossing measure $P_n$, with respect to which $S_n/n$ satisfies the law of large numbers. On the other hand, for $n \in \mathbb{N}$ and $\{\beta_k, k \in \mathbb{N}\}$ an arbitrary sequence tending to $\infty$, we have the weak convergence

$$P_{n,\beta} \Rightarrow P_{n,\infty} = \frac{1}{2}(\delta_{\omega^+} + \delta_{\omega^-}).$$

With respect to the limit measure a law of large numbers fails; in fact

$$P_{n,\infty}\{S_n/n = 1\} = P_{n,\infty}\{S_n/n = -1\} = 1/2.$$  

The large deviation principle for the sequence of $P_{n,\beta}$-distribution of the spins per site is straightforward to derive once we recall some facts about fair coin tossing. In Example 2.1, and again after the statement of Cramér's Theorem, fair coin tossing was considered with respect to the configuration space $\{0, 1\}^\mathbb{N}$. We must now translate these results to the configuration space $\{1, -1\}^\mathbb{N}$. Let $\{X_j, j \in \mathbb{N}\}$ be a sequence of i.i.d. random variables that are defined on this space and that have the common distribution $P_n\{X_j = 1\} = P_n\{X_j = -1\} = \frac{1}{2}$. Cramér's Theorem implies that the sequence $\{P_n\{S_n/n \in dx\}, n \in \mathbb{N}\}$ satisfies the large deviation principle.
with rate function
\[ I(x) = \begin{cases} 
\frac{1-x}{2} \log(1-x) + \frac{1+x}{2} \log(1+x) & \text{if } |x| \leq 1, \\
\infty & \text{if } |x| > 1.
\end{cases} \] 
(9.2)

We summarize this large deviation principle by the notation
\[ P_n \{ S_n/n \in dx \} \sim e^{-\beta I(x)} dx. \]

The large deviation principle for the sequence \( \{ P_{n,\beta} \{ S_n/n \in dx \}, n \in \mathbb{N} \} \) is now easily derived. Indeed, for any Borel subset \( A \) of \( \mathbb{R} \)
\[
P_{n,\beta} \{ S_n/n \in A \} = \frac{1}{Z_n(\beta)} \int_{\{ \omega \in \Omega_n : S_n(\omega)/n \in A \}} \exp[ -\beta H_n(\omega) ] P_n(\omega) \] 
\[
= \frac{1}{Z_n(\beta)} \int_{\{ \omega \in \Omega_n : S_n(\omega)/n \in A \}} \exp \left[ \frac{n\beta}{2} \left( \frac{S_n(\omega)}{n} \right)^2 \right] P_n(\omega) 
\]
\[
= \frac{1}{Z_n(\beta)} \int_A \exp \left[ \frac{n\beta}{2} x^2 \right] P_n \{ S_n/n \in dx \},
\]

where
\[
Z_n(\beta) = \int_{\Omega_n} \exp[ -\beta H_n(\omega) ] P_n(\omega) = \int_A \exp \left[ \frac{n\beta}{2} x^2 \right] P_n \{ S_n/n \in dx \}. \]

Part (b) of Proposition 3.4 yields the following theorem.

**THEOREM 9.1.** For each \( n \in \mathbb{N} \) and \( \beta > 0 \) we define the finite-volume Gibbs state \( P_{n,\beta} \) of the Curie-Weiss model by equation (9.1). Then for each \( \beta > 0 \) the sequence of distributions \( \{ P_{n,\beta} \{ S_n/n \in dx \}, n \in \mathbb{N} \} \) of the spins per site satisfies the large deviation principle on \( \mathbb{R} \) with rate function
\[
I_{\beta}(x) = I(x) - \frac{\beta}{2} x^2 - \inf_{y \in \mathbb{R}} \left\{ I(y) - \frac{\beta}{2} y^2 \right\}.
\]

The function \( I(x) \) is defined in equation (9.2). The rate function \( I_{\beta} \) is finite on the interval \([-1, 1]\) and equals \( \infty \) on the complement of \([-1, 1]\).

Because similar constructions will recur in subsequent sections, we want to emphasize the three steps leading to the derivation of this large deviation principle.

- The Hamiltonian \( H_n \) is a quadratic functional of the random quantity \( S_n/n \).
- We want to prove that the sequence of \( P_{n,\beta} \)-distributions of the same random quantities \( S_n/n \) satisfies the large deviation principle.
- With respect to the sequence of fair coin tossing measures \( \{ P_n, n \in \mathbb{N} \} \), we have the large deviation principle summarized by the notation \( P_n \{ S_n/n \in dx \} \sim e^{-n I_{\beta}(x)} dx \).

The form of the rate function \( I_{\beta} \) in Theorem 9.1 can be explained physically in terms of two competing effects. The large deviations of the sequence of spins per site \( S_n/n \) with respect to the fair coin tossing measures \( P_n \) represent randomizing or
"thermal" effects in the Curie-Weiss model. These effects give rise to the term \( I(x) \) in the rate function \( I_\beta \). They are balanced by the contribution of the Hamiltonian \( H_n = -(n/2)(S_n/n)^2 \), which represent the ordering effects due to the interactions in the model. The latter give rise to the quadratic term in the rate function \( I_\beta \).

In order to complete our analysis of the Curie-Weiss model we refer to Ellis (1985, Sect. IV.4), where the following facts are shown. For each \( \beta \in (0, 1) \) the rate function \( I_\beta \) attains its infimum at the unique point \( x = 0 \), which is also the unique minimum point of \( I \). Furthermore, for these values of \( \beta \), \( I_\beta \) is a strictly convex function on \([-1, 1]\). By contrast, for each \( \beta > 1 \), \( I_\beta \) is not convex. In fact there exists a point \( m^+(\beta) \in (0, 1) \) such that \( I_\beta \) attains its infimum at 0 at the two points \( m^+(\beta) \) and \(-m^+(\beta)\) and is positive for all other values of \( x \). Thus, if for each \( \beta > 0 \) we define the set

\[
\mathcal{E}_\beta = \{ x \in \mathbb{R} : I_\beta(x) = 0 \},
\]

then

\[
\mathcal{E}_\beta = \begin{cases} 
\{0\} & \text{if } \beta \in (0, 1) \\
\{m^+(\beta), -m^+(\beta)\} & \text{if } \beta > 1.
\end{cases}
\]

For the macroscopic observable which is the spin per site, the value \( \{0\} \) and the values \( \{m^+(\beta), -m^+(\beta)\} \) represent the equilibrium values for \( \beta \in (0, 1) \) and for \( \beta > 1 \), respectively. Part (b) of Proposition 3.3 and a symmetry argument now yield the following limits.

**Theorem 9.2.** For \( n \in \mathbb{N} \) we define the finite-volume Gibbs state \( P_{n,\beta} \) of the Curie-Weiss model by equation (9.1). Then as \( n \to \infty \) we have the weak limits

\[
P_{n,\beta} \left( \begin{array}{c}
S_n \\
n \in \mathbb{R}
\end{array} \right) \Rightarrow \begin{cases}
\delta_0 & \text{if } \beta \in (0, 1) \\
\frac{1}{2} \delta_{m^+(\beta)} + \frac{1}{2} \delta_{-m^+(\beta)} & \text{if } \beta > 1.
\end{cases}
\]

In other words, for \( \beta \in (0, 1) \) a law of large numbers is valid while for \( \beta > 1 \) the law of large numbers is replaced by a limit having atoms at the two symmetric points \( m^+(\beta) \) and \(-m^+(\beta)\).

This theorem justifies calling the value \( \beta = \beta_c = 1 \) the critical inverse temperature of the Curie-Weiss model. The limit replacing the law of large numbers for \( \beta > \beta_c \) corresponds to a phase transition. Because \( m^+(\beta) \to 0 \) as \( \beta \to 1^- \) while \((m^+(\beta))' \to \infty \) as \( \beta \to 1^+ \), we speak of a second order phase transition. This phase transition is closely related to the existence of spontaneous magnetization for the Curie-Weiss model (Ellis, 1985, Ch. IV). Additional limit theorems for the Curie-Weiss model are proved in Ellis & Newman (1978a; 1978b) and Ellis et al. (1980). Beside being a source of interesting probabilistic limit theorems, the Curie-Weiss model is an approximation to the behavior of much more complicated models including the Ising model (Ellis, 1985, pp. 183–186).

From the viewpoint of large deviations, we used a level-1 analysis to treat the Curie-Weiss model because the Hamiltonian is a quadratic functional of the level-1
quantity \( S_n/n \). The model to be studied in the next section will be treated by a level-2 analysis.

**10. THE CURIE-WIESS-POTTS MODEL**

In Section 2 of this paper we presented the two elementary examples of fair coin tossing and the fair \( q \)-faced die. The first of these played a key role in the analysis of the Curie-Weiss model. The second will play a similar role in the analysis of the model to be considered in the present section.

Let \( q \) be an integer exceeding 2 and as in the previous section, for each \( n \in \mathbb{N} \) define \( \Lambda_n \) to be the set \( \{1, 2, \ldots, n\} \). The first result in this section is the large deviation principle for the empirical vector in the Curie-Weiss-Potts model, which is a spin system on the sequence of subsets \( \{\Lambda_n, n \in \mathbb{N}\} \). The Curie-Weiss-Potts model has a critical inverse temperature \( \beta_c \in (0, \infty) \). For all values of the inverse absolute temperature \( \beta \) satisfying \( \beta \in (0, \beta_c) \) the rate function in the large deviation principle is strictly convex on the set of probability vectors in \( \mathbb{R}^\mathbb{N} \) and it has a unique minimum point. For all values \( \beta > \beta_c \) the rate function is nonconvex and finite on this set and it has \( q \) symmetric minimum points. For \( \beta = \beta_c \) the rate function is nonconvex and finite on this set and it has \( q \) symmetric minimum points together with an additional \( (q+1) \)-th minimum point. These properties lead immediately to a law of large numbers for the empirical vector when \( \beta \in (0, \beta_c) \) and replacement limit theorems when \( \beta = \beta_c \), and when \( \beta > \beta_c \) (Theorem 10.2). Analogous limit theorems for the spin per site can also be deduced.

The configuration space of the Curie-Weiss-Potts model is the set \( \Omega_n = \{1, 2, \ldots, q\}^\mathbb{N} \). \( \Omega_n \) consists of all sequences \( \omega = (\omega_j, j = 1, 2, \ldots, n) \), where each \( \omega_j \) represents the value of the spin at the site \( j \in \Lambda_n \). The possible spin values are any integer in \( \{1, 2, \ldots, q\} \). As in the case of the fair \( q \)-faced die, we define for each \( \omega \in \Omega_n \) the quantity \( P_n[\omega] = q^{-n} \) and extend this in the usual way to a probability measure \( P_n \) on the set \( \mathcal{P}_n \) of all subsets of \( \Omega_n \). With respect to \( P_n \), the coordinates \( \{\omega_j, j \in \Lambda_n\} \) are i.i.d. with common distribution \( \frac{1}{q} \delta_1 + \frac{1}{q} \delta_{-1} \).

For each \( n \in \mathbb{N} \) and \( \omega \in \Omega_n \) we introduce the empirical vector

\[
L_n^\omega(\omega) = (L_n^1(\omega), L_n^2(\omega), \ldots, L_n^q(\omega)),
\]

where for each \( i \in \{1, 2, \ldots, q\} \)

\[
L_n^i(\omega) = \frac{1}{n} \sum_{j=1}^n \delta_{\omega_j}(i).
\]

The empirical vector takes values in the set

\[
\mathcal{H} = \left\{ \nu \in \mathbb{R}^q : \nu = (\nu_1, \nu_2, \ldots, \nu_q), \nu_i \geq 0, \sum_{i=1}^q \nu_i = 1 \right\}
\]

consisting of the probability vectors in \( \mathbb{R}^q \).

Given real numbers \( a \) and \( b \) we define the quantity \( \delta(a, b) \) to be 1 if \( a = b \) and to be 0 if \( a \neq b \). For the Curie-Weiss-Potts model the Hamiltonian of a configuration
\( \omega \in \Omega_n \) is defined by the formula

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

where \( \langle \cdot, \cdot \rangle \) denotes the Euclidean inner product in \( \mathbb{R}^q \). Just as the Hamiltonian in the Curie-Weiss model is a quadratic function of the level-1 quantity \( S_n/n \), so the Hamiltonian in the Curie-Weiss-Potts model is a quadratic function of the level-2 quantity \( L^n \). For the Curie-Weiss-Potts model the probability of a configuration \( \omega \in \Omega_n \) corresponding to a given value of \( \beta > 0 \) is defined by the formula

\[
P_{n,\beta}(\omega) = \frac{1}{Z_n(\beta)} \cdot \exp[-\beta H_n(\omega)] \cdot P_n(\omega), \tag{10.1}
\]

in which \( Z_n(\beta) \) is the partition function

\[
Z_n(\beta) = \int_{\Omega_n} \exp[-\beta H_n(\omega)] P_n(d\omega).
\]

\( P_{n,\beta} \) is extended to a probability measure on \( \mathcal{F}_n \) by defining for any subset \( A \) of \( \Omega_n \)

\[
P_{n,\beta}(A) = \sum_{\omega \in A} P_{n,\beta}(\omega).
\]

The resulting probability measure \( P_{n,\beta} \) is the finite-volume Gibbs state on \( \Lambda_n \) for the model.

Our goal is first to derive for each \( \beta > 0 \) the large deviation principle for the sequence \( \{P_{n,\beta} \{L^n \in dv\}, n \in \mathbb{N}\} \) and then to deduce the weak limit of this sequence of distributions. From these results for the empirical vectors one can easily derive the corresponding results for the spins per site \( S_n/n \). In order to proceed, we recall from Cramér's Theorem that with respect to the sequence of fair \( q \)-faced die measures \( \{P_n, n \in \mathbb{N}\} \) the empirical vectors \( L_n \) satisfy the large deviation principle on \( \mathbb{R}^q \) with rate function

\[
I(v) = \begin{cases} 
\sum_{i=1}^{q} v_i \log(qv_i) & \text{if } v \in \mathcal{M} \\
\infty & \text{if } v \notin \mathbb{R}^q \setminus \mathcal{M}.
\end{cases} \tag{10.2}
\]

We summarize this large deviation principle by the notation \( P_n \{L^n \in dv\} \approx e^{-nI(v)} dv \).

As in the case of the Curie-Weiss model, we have for any Borel subset \( A \) of \( \mathbb{R}^q \)

\[
P_{n,\beta} \{L^n \in A\} = \frac{1}{Z_n(\beta)} \cdot \int_{\omega \in \Omega_n, L^n(\omega) \in A} \exp[-\beta H_n(\omega)] P(d\omega)
\]

\[
= \frac{1}{Z_n(\beta)} \cdot \int_{\omega \in \Omega_n, L^n(\omega) \in A} \exp \left[ \frac{n\beta}{2} \langle L^n(\omega), L^n(\omega) \rangle \right] P_n(d\omega)
\]

\[
= \frac{1}{Z_n(\beta)} \cdot \int_{A} \exp \left[ \frac{n\beta}{2} \langle v, v \rangle \right] P_n \{L^n \in dv\},
\]
where
\[ Z_n(\beta) = \int_{\Omega_n} \exp[-\beta H_n(\omega)] P_n(d\omega) = \int_{\mathbb{R}^n} \exp\left[ -\frac{n\beta}{2} \langle v, v \rangle \right] P_n[L^n \in dv]. \]

Part (b) of Proposition 3.3 yields the following theorem.

**THEOREM 10.1.** For each \( n \in \mathbb{N} \) and \( \beta > 0 \) we define the finite-volume Gibbs state \( P_{n,\beta} \) of the Curie-Weiss-Potts model by equation (10.1). Then for each \( \beta > 0 \) the sequence of distributions \( \{P_{n,\beta}[L^n \in dv], n \in \mathbb{N}\} \) of the empirical vectors satisfies the large deviation principle on \( \mathbb{R} \) with rate function
\[ I(\nu) = I(\nu) - \frac{\beta}{2} \langle \nu, \nu \rangle - \inf_{\mu \in \mathcal{M}} \left\{ I(\mu) - \frac{\beta}{2} \langle \mu, \mu \rangle \right\}. \]

The function \( I(\nu) \) is defined in equation (10.2). The rate function \( I_{\beta} \) is finite on the set \( \mathcal{M} \) of probability vectors in \( \mathbb{R}^n \) and equals \( \infty \) on the complement of \( \mathcal{M} \).

From this large deviation principle one can derive the large deviation principle for the sequence of \( P_{n,\beta} \)-distributions of the spins per site by applying the contraction principle (part (c) of Proposition 3.3).

Again, we want to emphasize the three steps leading to the derivation of the large deviation principle in Theorem 10.1.

- The Hamiltonian \( H_n(\omega) \) is a quadratic functional of the random quantity \( L^n \).
- We want to prove that the sequence of \( P_{n,\beta} \)-distributions of the same random quantities \( L^n \) satisfies the large deviation principle.
- With respect to the sequence of fair \( q \)-faced die measures \( \{P_n, n \in \mathbb{N}\} \), we have the large deviation principle summarized by the notation \( P_n[L^n \in dv] \approx e^{-nI_{\beta}(v)} dv \).

We define the quantity
\[ \beta_\ast = 2 \left( \frac{q-1}{q-2} \right) \log(q-1), \]
which represents the critical inverse temperature of the Curie-Weiss-Potts model, and denote by \( \mu^* \) the point in \( \mathcal{M} \) all of whose coordinates equal \( 1/q \). In order to complete our analysis of this model we refer to Ellis & Wang (1990), where the following facts are shown. For each \( \beta \in (0, \beta_\ast) \) the rate function \( I_{\beta} \) attains its infimum of 0 at the unique point \( \nu = \mu^* \), which is also the unique minimum point of the relative entropy \( I(\nu | \mu^*) \). Furthermore, for these values of \( \beta \), \( I_{\beta} \) is a strictly convex function on \( \mathcal{M} \). By contrast, for each \( \beta \geq \beta_\ast \), \( I_{\beta} \) is not convex. For \( \beta \geq \beta_\ast \), there exist \( q \) distinct points \( \{m^{(i)}(\beta), i = 1, 2, \ldots, q\} \) in \( \mathcal{M} \) such that for \( \beta > \beta_\ast \), \( I_{\beta} \) attains its infimum of 0 at precisely these \( q \) points, while for \( \beta = \beta_\ast \), \( I_{\beta} \) attains its infimum of 0 at the \((q+1)\) distinct points \( \{\mu^*, m^{(1)}(\beta_\ast), \ldots, m^{(q)}(\beta_\ast)\} \). For all other values of \( \nu \), \( I_{\beta}(\nu) \) is positive. Thus, if for each \( \beta > 0 \) we define the set
\[ \mathcal{D}_\beta = \{ v \in \mathbb{R}^n : I_{\beta}(v) = 0 \} , \]

then

\[ \mathcal{E}_\beta = \begin{cases} \{0\} & \text{if } \beta \in (0, \beta_c) \\ \{m^{(1)}(\beta_c), \ldots, m^{(q)}(\beta_c)\} & \text{if } \beta = \beta_c \\ \{m^{(1)}(\beta_c), m^{(2)}(\beta_c), \ldots, m^{(q)}(\beta_c)\} & \text{if } \beta > \beta_c. \end{cases} \]

For \( \beta \geq \beta_c \), the first \((q-1)\) components of the vector \(m^{(1)}(\beta)\) are equal and differ from the last component, and the vectors \(\{m^{(i)}(\beta), i = 2, 3, \ldots, q\}\) are obtained from \(m^{(1)}(\beta)\) by permutations. In the cases where \(\beta \in (0, \beta_c)\) and \(\beta > \beta_c\), part (b) of Proposition 3.3 and a symmetry argument now yield the following limits. The derivation of the limit when \(\beta = \beta_c\) requires a different argument (Ellis & Wang, 1990).

**Theorem 10.2.** For \(n \in \mathbb{N}\) we define the finite-volume Gibbs state \(P_{n,\beta}\) of the Curie-Weiss-Potts model by equation (10.1). Then as \(n \to \infty\) we have the weak limits

\[ P_{n,\beta} \{ L^n \in d\nu\} = \begin{cases} \delta_{\beta} & \text{if } \beta \in (0, \beta_c) \\ \gamma_1 \sum_{i=1}^q \delta_{m^{(i)}(\beta)} + \gamma_2 \delta_{\beta} & \text{if } \beta = \beta_c \\ \frac{1}{q} \sum_{i=1}^q \delta_{m^{(i)}(\beta)} & \text{if } \beta > \beta_c. \end{cases} \]

In line two of this limit, \(\gamma_1\) and \(\gamma_2\) are explicitly given positive numbers summing to 1. In other words, for \(\beta \in (0, \beta_c)\) a law of large numbers is valid, while for \(\beta = \beta_c\), and for \(\beta > \beta_c\), the law of large numbers is replaced by a limit having atoms at \(q+1\) points and at \(q\) points, respectively.

This completes our analysis of the Curie-Weiss-Potts model. In the next section we consider the Ising model, the treatment of which requires a much more sophisticated analysis at level-3. From the limits expressed in Theorem 10.2, it is straightforward to derive the weak limits, for \(\beta \in (0, \beta_c)\), for \(\beta = \beta_c\), and for \(\beta > \beta_c\), of the sequence \(\{P_{n,\beta} \{ S_n /n \in d\lambda\}, n \in \mathbb{N}\}\) of distributions of the spins per site. On the other hand, without studying first the limiting behavior of the empirical vectors, it would be difficult to derive directly the weak limits of the sequence \(\{P_{n,\beta} \{ S_n /n \in d\lambda\}, n \in \mathbb{N}\}\). We will see that a similar situation arises in the case of the Ising model. Other limit theorems for the Curie-Weiss-Potts model are proved in Ellis & Wang (1990; 1992), Kesten & Schonmann (1990) and Orey (1988). Just as the Curie-Weiss model is an approximation to the behavior of much more complicated models including the Ising models, so the Curie-Weiss-Potts model is an approximation to the behavior of the much more complicated Potts model, which has been the subject of many studies (Wu, 1982; 1984).

**11. The Ising Model**

Both in the Curie-Weiss model and the Curie-Weiss-Potts model the Hamiltonians \(H_\omega(\omega)\) are symmetric functions of the spin random variables \(\{\omega_j\}\). Thus in each set \(A_\omega\) each spin \(\omega_i\) has the same interaction with all the other spins \(\omega_j, j \neq i\). To put it another way, neither the Curie-Weiss model nor the Curie-Weiss-Potts model
depends on the geometry of the underlying lattice. By contrast, a distinguishing feature of the Ising model is a strong geometric dependence.

Let $D \geq 2$ be a positive integer and for each $n \in \mathbb{N}$ let $\{\Lambda_n, n \in \mathbb{N}\}$ be a sequence of expanding hypercubes of side length $n$ in the $D$-dimensional integer lattice $\mathbb{Z}^D$. Thus the cardinality of $\Lambda_n | \Lambda_n|$, equals $n^D$. The Ising model is a spin system on the sequence of subsets $\{\Lambda_n, n \in \mathbb{N}\}$ having a critical inverse temperature $\beta_c \in (0, \infty)$. The main result in this section is the large deviation principle, with norming constants $\{|\Lambda_n|, n \in \mathbb{N}\}$, for the spins per site $\{S_n | / |\Lambda_n|, n \in \mathbb{N}\}$ (Theorem 11.1). By analogy with the two models just considered, for all values of the inverse absolute temperature $\beta$ satisfying $\beta < \beta_c$, the rate function in the large deviation principle is strictly convex on the interval $[-1, 1]$ and it has a unique minimum point at 0. However, in contrast to these two models, for all values $\beta > \beta_c$, the rate function is convex and finite on this interval, but it has an entire interval of minimum points. This interval, which we call the phase transition interval, takes the form $[ -m^*(\beta), m^*(\beta) ]$ for some number $m^*(\beta) \in (0, 1)$ that increases to 1 as $\beta \to \infty$.

These properties have dramatic consequences for the asymptotic behavior of the spins per site $\{S_n | / |\Lambda_n|, n \in \mathbb{N}\}$. First consider the case where $\beta \in (0, \beta_c)$ and $A$ is any closed interval of $[-1, 1]$ having nonempty interior and not containing the origin. Since the infimum of the rate function over $A$ is positive, the probability that $S_n | / |\Lambda_n|$ lies in $A$ decays to 0 like $\exp \{-const \cdot |\Lambda_n|\}$ as $n \to \infty$. Exactly the same exponential decay to 0 occurs when $\beta > \beta_c$ and $A$ is any closed interval of $[-1, 1]$ having nonempty interior and empty intersection with the phase transition interval. The difference occurs in the case when $\beta > \beta_c$ and $A$ is any closed interval having nonempty interior and nonempty intersection with the phase transition interval. Since on the phase transition interval the rate function equals 0, the large deviation principle yields the limit

$$\lim_{n \to \infty} \frac{1}{|\Lambda|} \log P_{n, \beta} \{S_n | / |\Lambda_n| \in A\} = 0,$$

where $\{P_{n, \beta}, n \in \mathbb{N}\}$ is the corresponding sequence of finite-volume Gibbs states. This limit gives no information at all concerning the decay of the probabilities.

In the special case of the Ising model on $\mathbb{Z}^2$ the works (Chayes et al., 1987; Dobrushin et al., 1989; 1992; Pfister, 1991; Schonmann, 1987) show that inside the phase transition interval large deviation probabilities decay to 0, not like $\exp \{-const \cdot n^2\}$, but like $\exp \{-const \cdot n\}$ as $n \to \infty$. This is a boundary effect since $n$ is proportional to the perimeter of the square $\Lambda_n \in \mathbb{Z}^2$. Related phenomena are explored in Föllmer & Ort (1988). The study of boundary effects in the large deviation behavior of the Ising model is much more subtle and complex than the proof of the large deviation principle stated in Theorem 11.1.

We do not discuss the Ising model on $\mathbb{Z}$ because from the viewpoint of statistical mechanics it is uninteresting. In fact, the Ising model on $\mathbb{Z}$ is essentially a finite state Markov chain, and the critical inverse temperature $\beta_c$ equals $\infty$.

The configuration space of the Ising model is the set $\Omega_n = \{1, -1\}^{|n|}$. For each $\omega \in \Omega_n$ we define the quantity $P_{n, \beta} |\omega| = 2^{-|\omega|}$ and extend this in the usual way to $n$.
probability measure $P_n$ on the set $\mathcal{F}_n$ of all subsets of $\Omega_n$. With respect to $P_n$, the coordinates $\{\omega_j, j \in \Lambda_n\}$ are i.i.d. with common distribution $\frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$. In the Ising model the Hamiltonian of a configuration $\omega \in \Omega_n$ is given by the formula

$$H_n(\omega) = -\frac{1}{2} \sum_{(\omega_i, \omega_j) \in \Lambda_n \setminus \Lambda_n^{\prime \prime}} \omega_i \omega_j.$$ 

Thus the double sum is taken over all nearest neighbor pairs in $\Lambda_n$. The factor $\frac{1}{2}$ appears because each pair is counted twice.

Given the definition of the Hamiltonian, we complete the definition of the Ising model as in the previous two cases. The probability of a configuration $\omega \in \Omega_n$ corresponding to a given value of $\beta > 0$ is defined by the formula

$$P_{n, \beta}(\omega) = \frac{1}{Z_n(\beta)} \cdot \exp[-\beta H_n(\omega)] \cdot P_{\omega}(\omega), \quad \text{(11.1)}$$

in which $Z_n(\beta)$ is the partition function

$$Z_n(\beta) = \int_{\Omega_n} \exp[-\beta H_n(\omega)] P_{\omega}(d\omega).$$

$P_{n, \beta}$ is extended to a probability measure on $\mathcal{F}_n$ by defining for any subset $A$ of $\Omega_n$

$$P_{n, \beta}[A] = \sum_{\omega \in A} P_{n, \beta}(\omega).$$

The resulting probability measure $P_{n, \beta}$ is the finite-volume Gibbs state on $\Lambda_n$ for the model.

There are many variations on this definition. For example, for each $n \in \mathbb{N}$ let us define the boundary of $\Lambda_n$ by the formula

$$\partial \Lambda_n = \{ k \in \mathbb{Z}^D \setminus \Lambda_n : \| k - j \| = 1 \text{ for some } j \in \Lambda_n \}.$$ 

Then one can include boundary conditions in the finite volume Gibbs states by fixing a set of spin values $\omega_j$ for all lattice points $j \in \partial \Lambda_n$ and by replacing the sum in the Hamiltonian by a sum over nearest neighbor pairs in $\Lambda_n \cup \partial \Lambda_n$. For example, the choice $\omega_j = 1$ for all $j \in \partial \Lambda_n$ gives the plus boundary condition while the choice $\omega_j = -1$ for all $j \in \partial \Lambda_n$ gives the minus boundary condition. The Ising Hamiltonian, being a sum over nearest neighbor pairs, is said to have range 1. This can be replaced by a Hamiltonian having finite range $R \geq 2$ or even infinite range. In addition, one can replace the hypercubes $\{\Lambda_n, n \in \mathbb{N}\}$ by other sequences of subsets of $\mathbb{Z}^D$. These and other generalizations are described in the enormous literature on the Ising model. A good place to start is Chapters IV and V of Ellis (1985). The Ising model on $\mathbb{Z}^2$ is special because it lends itself to exact calculations (Baxter, 1982; McCoy & Wu, 1973).

Our goal is to prove for each $\beta > 0$ the large deviation principle for the sequence of $P_{n, \beta}$ distributions of the spins per site. For each $\omega \in \Omega_n$ these are the random variables

$$S_n(\omega) = \frac{1}{|\Lambda_n|} \sum_{j=1}^n \omega_j.$$
The derivation of the large deviation principle is harder than in the cases of the Curie-Weiss model and the Curie-Weiss-Potts model, for which it was derived as an application of part (b) of Proposition 3.4. We recall that in order to apply the latter, we wrote the corresponding Hamiltonian as a function of a process for which the large deviation principle with respect to the sequence of measures \( \{P_n, n \in \mathbb{N}\} \) is available. In the case of the Curie-Weiss model this process was the spin per site \( S_n/n \) and in the case of the Curie-Weiss-Potts model this process was the empirical vector \( L^n \). However for the Ising model, it is not at all clear which process plays the role that these processes played earlier. Although we will answer this question later, it is instructive first to try to prove the large deviation principle for the Ising model using the Gärtner-Ellis Theorem, stated in Theorem 5.1.

According to this theorem, we must evaluate for each \( t \in \mathbb{R} \) the limit function

\[
\psi(t) = \lim_{n \to \infty} \frac{1}{|\Lambda_n|} \log \int_{\Omega_n} \exp[t S_n(\omega)] P_{n, \beta}(d\omega). \tag{11.2}
\]

This limit function can be expressed in terms of the specific Gibbs free energy for the Ising model, as I now explain. For \( h \) a real number and \( \omega \in \Omega_n \) we define the Hamiltonian

\[
H_{n, h}(\omega) = -\frac{1}{2} \sum_{\{i, j \in \Lambda_n : \|i-j\| \leq 1\}} \omega_i \omega_j - h \sum_{i \in \Lambda_n} \omega_i
\]

and the associated partition function

\[
Z_n(\beta, h) = \int_{\Omega_n} \exp[-\beta H_{n, h}(\omega)] P_n(d\omega).
\]

The parameter \( h \) represents an external magnetic field. We have the following facts, which are proved, for example, in Ellis (1985, Chs. IV-V).

- For each \( \beta > 0 \) and \( h \in \mathbb{R} \) the limit
  
  \[
  \psi(\beta, h) = \lim_{n \to \infty} \frac{1}{|\Lambda_n|} \log Z_n(\beta, h) \tag{11.3}
  \]
  
  exists. The function \( -h^{-1} \psi(\beta, h) \) is called the specific Gibbs free energy.

- There exists a critical inverse temperature \( \beta_c \in (0, \infty) \) such that for all \( \beta \in (0, \beta_c) \) \( \psi(\beta, h) \) is a differentiable function of \( h \in \mathbb{R} \).

- For all \( \beta > \beta_c \) \( \psi(\beta, h) \) is a differentiable function of \( h \neq 0 \) but is not differentiable at \( h = 0 \). This nondifferentiability is closely related to the phase transition in the Ising model associated with the phenomenon of "spontaneous magnetization."

These facts are proved by a combination of subadditivity arguments, convexity arguments, and moment inequalities for the finite-volume Gibbs states. When \( \beta = \beta_c \) and \( D = 2 \psi(\beta, h) \) is also a differentiable function at \( h = 0 \). When \( \beta = \beta_c \) and \( D \geq 3 \psi(\beta, h) \) is a differentiable function of \( h \neq 0 \), but the differentiability of \( \psi(\beta, h) \) at \( h = 0 \) is not known.
We now return to the evaluation of the Gärtner-Ellis limit function \( c_\beta(t) \) defined in equation (11.2). Since for each \( \omega \in \Omega \),
\[
t S_n(\omega) - \beta H_n(\omega) = -\beta H_{n, \Omega}(\omega),
\]
we have
\[
c_\beta(t) = \lim_{n \to \infty} \frac{1}{|\Lambda_n|} \log \left( \frac{1}{Z_n(\beta)} \int_{\Omega_n} \exp[tS_n - \beta H_n(\omega)]P_n(d\omega) \right)
\]
\[
= \lim_{n \to \infty} \frac{1}{|\Lambda_n|} \log \left( \frac{1}{Z_n(\beta)} \int_{\Omega_n} \exp[-\beta H_{n, \Omega}(\omega)]P_n(d\omega) \right)
\]
\[
= \lim_{n \to \infty} \frac{1}{|\Lambda_n|} \log \left( \frac{Z_n(\beta, t/\beta)}{Z_n(\beta)} \right)
\]
\[
= \psi(\beta, t/\beta) - \psi(\beta, 0).
\]
(11.4)

For each \( \beta \in (0, \beta_0) \) the function \( \psi(\beta, h) \) is a differentiable function of \( h \in \mathbb{R} \). Hence for each \( \beta \in (0, \beta_0) \) the limit function \( c_\beta(t) \) is a differentiable function of \( t \in \mathbb{R} \). The Gärtner-Ellis Theorem implies that for this set of \( \beta \) the sequence of distribution \( \{P_{n, \beta} [S_n / |\Lambda_n| \in dx], n \in \mathbb{N} \} \) satisfies the large deviation principle on \( \mathbb{R} \) with rate function
\[
I_\beta(x) = \sup_{t \in \mathbb{R}} \{tx - c_\beta(t)\}
\]
and norming constants \( \{|\Lambda_n|, n \in \mathbb{N}\} \). However, for each \( \beta > \beta_0 \) the differentiability of \( \psi(\beta, h) \) at \( h = 0 \), and thus the differentiability of \( c_\beta(t) \) at \( t = 0 \), fail. Hence for this set of \( \beta \) the Gärtner-Ellis Theorem is not applicable. Nevertheless, as the next theorem shows, for each \( \beta > \beta_0 \) the large deviation principle holds with the same rate function and the same norming constants as in the case \( \beta \in (0, \beta_0) \). The identical large deviation principle is valid if the sequence \( \{P_{n, \beta}, n \in \mathbb{N}\} \) is replaced by any sequence of finite-volume Gibbs states with nearest neighbor interactions and any choice of boundary conditions.

**THEOREM 11.1.** For each \( n \in \mathbb{N} \) and \( \beta > 0 \) we define the finite-volume Gibbs state \( P_{n, \beta} \) of the Ising model by equation (11.1). The following conclusions hold.

(a) For each \( \beta > 0 \) the sequence of distributions \( \{P_{n, \beta} [S_n / |\Lambda_n| \in dx], n \in \mathbb{N}\} \) of the spins per site satisfies the large deviation principle on \( \mathbb{R} \) with rate function
\[
I_\beta(x) = \sup_{t \in \mathbb{R}} \{tx - c_\beta(t)\}.
\]

For \( t \in \mathbb{R} \) the function \( c_\beta(t) \) is given by the formula
\[
c_\beta(t) = \psi(\beta, t/\beta) - \psi(\beta, 0),
\]
where for \( \beta > 0 \) and \( h \in \mathbb{R} \) \( \psi(\beta, h) \) is defined by the limit (11.3). The rate function \( I_\beta \) is finite on the interval \([ -1, 1] \) and equals \( \infty \) on the complement of \([-1, 1]\).

(b) For each \( \beta \in (0, \beta_0) \) the rate function \( I_\beta \) is a symmetric strictly convex function on \([ -1, 1] \) and attains its infimum of 0 at the unique point \( x = 0 \).
(c) For each $\beta > \beta_c$ the right-hand derivative $D^+c_\beta(0)$ of $c_\beta(t)$ at $t = 0$ exists and satisfies $D^+c_\beta(0) \in (0, 1)$. For each such $\beta$ the rate function $I_\beta$ is a symmetric convex function on $\mathbb{R}$ but is not strictly convex. In fact, $I_\beta$ attains its infimum of 0 on the entire interval $[-m^+(\beta), m^-(\beta)]$, where $m^+(\beta) = D^+c_\beta(0) > 0$. We call this interval the phase transition interval.

For the value $\beta = \beta_c$ and $D = 2$ the shape of the rate function is like in part (b). For the value $\beta = \beta_c$ and $D \geq 3$ the shape of the rate function is either like in part (b) or like in part (c), but it is not known which.

Parts (a) and (b) of Theorem 11.1 imply that for each $\beta \in (0, \beta_c)$

$$P_{n,\beta}[S_n/|\Lambda_n| \in d\lambda] \to \delta_0 \quad \text{as } n \to \infty.$$  

The identical limit is valid if the sequence $\{P_{n,\beta}, n \in \mathbb{N}\}$ is replaced by any sequence of finite-volume Gibbs states with nearest neighbor interactions and any choice of boundary conditions. However, for $\beta > \beta_c$ the existence and the form of the weak limit of $\{P_{n,\beta}[S_n/|\Lambda_n|] \in d\lambda, n \in \mathbb{N}\}$ is not a consequence of part (c) of Theorem 11.1 but require special techniques. For $\beta > \beta_c$ one may also investigate this question when the sequence $\{P_{n,\beta}, n \in \mathbb{N}\}$ is replaced by a sequence of finite-volume Gibbs states with nearest neighbor interactions and some choice of boundary conditions. The existence and form of the weak limit depend upon the particular boundary conditions that are chosen.

Parts (b) and (c) of Theorem 11.1 follow from the properties of the specific Gibbs free energy $\psi(\beta, h)$ summarized earlier and from convex analysis. The details are omitted. We spend the remainder of this section sketching the proof of part (a) of the theorem. Rather than obtain this large deviation principle directly, we will first derive the large deviation principle of another process, which we call $\{\Xi_n, n \in \mathbb{N}\}$ and which has the following properties.

- The process $\{\Xi_n, n \in \mathbb{N}\}$ takes values in a complete separable metric space $\mathcal{X}$. There exists a bounded continuous function $\Phi$ mapping $\mathcal{X}$ into $\mathbb{R}$ such that for each $n \in \mathbb{N}$ and $\omega \in \Omega_n$ the Hamiltonian $H_n(\omega)$ can be written in the form

$$-H_n(\omega) = |\Lambda_n| \cdot \Phi(\Xi_n(\omega)) + o(|\Lambda_n|) \quad \text{uniformly for } \omega \in \Omega_n;$$  

i.e.,

$$\lim_{n \to \infty} \max_{\omega \in \Omega_n} \frac{1}{|\Lambda_n|} |H_n(\omega) - |\Lambda_n| \cdot \Phi(\Xi_n(\omega))| = 0.$$

- With respect to the measures $P_n$ that assign probability $2^{-|\Lambda_n|}$ to each $\omega \in \Omega_n$, the processes $\{\Xi_n, n \in \mathbb{N}\}$ satisfy the large deviation principle with some rate function $K$ on $\mathcal{X}$ and with norming constants $|\Lambda_n|, n \in \mathbb{N}$.

- There exists a continuous function $g$ mapping $\mathcal{X}$ into $\mathbb{R}$ such that for each $n \in \mathbb{N}$ and $\omega \in \Omega_n$, $S_n(\omega)/|\Lambda_n| = g(\Xi_n(\omega))$.

The first two properties, together with a minor modification of part (b) of Proposition 3.4 which allows one to handle the error $o(|\Lambda_n|)$, imply that for each
\( \beta > 0 \) the \( P_{x, \beta} \)-distributions of the processes \( \{\Xi_n, n \in \mathbb{N}\} \) satisfy the large deviation principle on \( \mathcal{X} \) with rate function

\[
K_\beta(Q) = K(Q) - \beta \Phi(Q) - \inf_{P \in \mathcal{K}} \{K(P) - \beta \Phi(P)\}
\]

and with norming constants \( \{\|\Lambda_n\|, n \in \mathbb{N}\} \). The third property, together with the contraction principle (part (c) of Proposition 3.3), implies that for each \( \beta > 0 \) the \( P_{x, \beta} \)-distributions of the spins per site \( \{S_n/\Lambda_n, n \in \mathbb{N}\} \) satisfy the large deviation principle on \( \mathcal{R} \) with some rate function \( J_\beta \) defined in terms of \( K_\beta \) and with norming constants \( \{\|\Lambda_n\|, n \in \mathbb{N}\} \). The form of \( J_\beta \) is not important, only the fact that it exists.

The final step is to show that for \( x \in \mathbb{R} \) \( J_\beta(x) \) equals the function

\[
I_\beta(x) \equiv \sup_{t \in \mathbb{R}} \{tx - c_\beta(t)\}
\]

which appears in part (a) of Theorem 11.1. This is done exactly as in Section 4 where we motivated the form of the rate function in Cramèr’s Theorem just before sketching the proof of that theorem. We will not bother to repeat the calculation.

The magic process \( \{\Xi_n, n \in \mathbb{N}\} \) having all these wonderful properties is a straightforward generalization to \( D \) dimensions of the level-3 empirical process defined in Section 7. For \( n \in \mathbb{N} \) and \( \omega \in \Omega_n \) let \( X(n, \omega) \) be the point in \( \{-1, 1\}^2 \) obtained by repeating the sequence \( \omega_1, \omega_2, \ldots, \omega_n \) periodically into a doubly infinite sequence. For Borel subsets \( A \) of \( \{-1, 1\}^2 \), we recall that the empirical process is defined by the formula

\[
R_n(A) = R_n(\omega, A) \equiv \frac{1}{n} \sum_{k=0}^{n-1} \delta_{T^k X(n, \omega)}(A),
\]

where \( T^0 \) is the identity mapping, \( T \) is the shift mapping on the sequence space \( \{1, -1\}^2 \), and \( T^k = T(T^{k-1}) \). For each \( \omega \in \Omega_n \), \( R_n(\omega, \cdot) \) takes values in the space \( \mathcal{P}_c(\{-1, 1\}^2) \) of strictly stationary probability measures on \( \{-1, 1\}^2 \).

We now define a \( D \)-dimensional generalization of the empirical process. For \( n \in \mathbb{N} \) and \( \omega \in \Omega_n \) consider the \( \Lambda_n \)-periodized point \( X(n, \omega) \) in \( \{-1, 1\}^{2^D} \) having coordinates \( (X(n, \omega)_j) = \omega_j \) for each \( j \) in the hypercube \( \Lambda_n \). We denote by \( T_1, T_2, \ldots, T_D \) the shift mappings in the \( D \) coordinate directions and for \( j \in \Lambda_n \), set \( T_j = T_{j1} T_{j2} \cdots T_{jd} \). For Borel subsets \( A \) of \( \{-1, 1\}^{2^D} \) we define the empirical field

\[
\Xi_n(A) = \Xi_n(\omega, A) \equiv \frac{1}{|\Lambda_n|} \sum_{k=0}^{2^D-1} \delta_{T^k X(n, \omega)}(A).
\]

For each \( \omega \in \Omega_n \), \( \Xi_n(\omega, \cdot) \) takes values in the space \( \mathcal{P}_c(\{-1, 1\}^{2^D}) \) of strictly stationary probability measures on \( \{-1, 1\}^{2^D} \). This space is a closed subset of the space \( \mathcal{P}_c(\{-1, 1\}^{2^D}) \) of probability measures on \( \{-1, 1\}^{2^D} \) and so can be metrized as a complete separable metric space.

Let \( 0 \) denote the origin in \( \mathbb{Z}^D \). For \( Q \) a measure in \( \mathcal{P}_c(\{-1, 1\}^{2^D}) \) we define the function

\[
\Phi(Q) \equiv \frac{1}{2} \int_{\mathbb{Z}^D \cap [-1, 1]} \int_{\mathbb{Z}^D \cap [-1, 1]} \delta_0 \delta_0 Q(d\tilde{\omega})
\]

\( 11.7 \)
and the function

\[
g(Q) = \int_{[1,-1]^{2\Omega}} \tilde{\omega} Q(\tilde{\omega}).
\]

(11.8)

Both \(\Phi\) and \(g\) are bounded continuous affine functions mapping \(\mathcal{P}_s([1,-1]^{2\Omega})\) into \(\mathbb{R}\). It is straightforward to check that for each \(n \in \mathbb{N}\) and \(\omega \in \Omega_n\) the Ising model Hamiltonian \(H_n(\omega)\) can be written in the form

\[
-H_n(\omega) = |\Lambda_n| \cdot \Phi(\Xi_n(\omega, \cdot)) + O(|\partial \Lambda_n|) \quad \text{uniformly for } \omega \in \Omega_n.
\]

Since \(|\partial \Lambda_n|/|\Lambda_n| \to 0\) like \(1/n\), we have verified formula (11.5). In addition, for each \(n \in \mathbb{N}\) and \(\omega \in \Omega_n\) the spin per site in the hypercube \(\Lambda_n\) is related to the empirical field by the formula

\[
\frac{S_n(\omega)}{n} = g(\Xi_n(\omega, \cdot)).
\]

In [38] it is proved that with respect to the measures \(P_n\) that assign probability \(2^{-|\Lambda_n|}\) to each \(\omega \in \Omega_n\), the processes \(\{\Xi_n, n \in \mathbb{N}\}\) satisfy the large deviation principle with norming constants \(\{|\Lambda_n|, n \in \mathbb{N}\}\). The paper [38] also identifies the form of the rate function. This takes care of the last of the three properties needed to prove that the \(P_n\)-distributions of the spins per site \(\{S_n/|\Lambda_n|, n \in \mathbb{N}\}\) satisfy the large deviation principle on \(\mathbb{R}\) with norming constants \(\{|\Lambda_n|, n \in \mathbb{N}\}\). The sketch of the proof of part (a) of Theorem 11.1 is complete.

As we have mentioned above, a consequence of the fact that the \(P_n\)-distributions of the empirical fields satisfy the large deviation principle on \(\mathcal{P}_s([1,-1]^{2\Omega})\) with some rate function \(K\) (Olla, 1988) is that for each \(\beta > 0\) the \(P_n\)-distributions of the empirical fields satisfy the large deviation principle on \(\mathcal{P}_s([1,-1]^{2\Omega})\) with rate function \(K_\beta\) defined in equation (11.6). The latter fact is also proved in Cornets (1986) and in Föllmer & Orey (1987), which like Olla (1988) consider a large class of models that include the Ising model as a special case. In the next section we present a general approach to the large deviation analysis of models in statistical mechanics.

12. A GENERAL APPROACH TO THE LARGE DEVIATION ANALYSIS OF MODELS IN STATISTICAL MECHANICS

By abstracting the calculations in the last three sections, we can give a general approach to the large deviation analysis of models in statistical mechanics. We consider a model that is defined in terms of the following data.

- A sequence of configuration spaces \(\{\Omega_n, n \in \mathbb{N}\}\).
- For each \(n \in \mathbb{N}\) the Hamiltonian \(H_n(\omega)\) of \(\omega \in \Omega_n\).
- A sequence of positive norming constants \(\{a_n, n \in \mathbb{N}\}\).
- For each \(n \in \mathbb{N}\) a probability measure \(P_n\) on \(\Omega_n\).
In terms of these quantities we define for each \( n \in \mathbb{N}, \beta > 0, \) and subset \( A \) of \( \Omega_n \) the partition function

\[
Z_n(\beta) = \int_{\Omega_n} \exp[-\beta H_n(\omega)] P_n(d\omega)
\]

and the finite-volume Gibbs state

\[
P_{n,\beta}(A) = \frac{1}{Z_n(\beta)} \int_A \exp[-\beta H_n(\omega)] P_n(d\omega).
\]

For \( \beta > 0 \) we define a function \( \psi(\beta) \) by the limit

\[
\psi(\beta) = \lim_{n \to \infty} \frac{1}{a_n} \log Z_n(\beta)
\]

if the limit exists. The function \( -\beta^{-1} \psi(\beta) \) is the specific Gibbs free energy for the model.

By analogy with our work in the previous section on the Ising model, we see that in order to carry out a large deviation analysis of the model the following four items are needed:

- A complete separable metric space \( \mathcal{F} \), called the **hidden space**.
- For each \( n \in \mathbb{N} \) a random variable \( \zeta_n \) mapping \( \Omega_n \) into \( \mathcal{F} \). We call the sequence \( \{\zeta_n, n \in \mathbb{N}\} \) the **hidden process**.
- For each \( n \in \mathbb{N} \) a bounded continuous function \( \Phi \) mapping \( \mathcal{F} \) into \( \mathbb{R} \) such that

\[
-H_n(\omega) = a_n \Phi(\zeta_n(\omega)) + o(a_n)
\]

uniformly for \( \omega \in \Omega_n \);

i.e.,

\[
\lim_{n \to \infty} \sup_{\omega \in \Omega_n} \frac{1}{a_n} \left| -H_n(\omega) - a_n \Phi(\zeta_n(\omega)) \right| = 0.
\]

We call the function \( \Phi \) the **energy function**.

- A **rate function** \( I \) mapping \( \mathcal{F} \) into \( \mathbb{R} \) such that the \( P_n \)-distributions of \( \{\zeta_n, n \in \mathbb{N}\} \) satisfy the large deviation principle on \( \mathcal{F} \) with rate function \( I \) and normalizing constants \( \{a_n, n \in \mathbb{N}\} \).

For example, in the case of the Curie-Weiss model \( \mathcal{F} \) equals \( \mathbb{R} \), \( \zeta_n \) equals the spin per site \( S_n/n \), and for \( x \in \mathbb{R} \) the energy function is given by the formula \( \Phi(x) = \frac{1}{2} x^2 \).

In the case of the Curie-Weiss-Potts model \( \mathcal{F} \) equals \( \mathbb{R}^p \), \( \zeta_n \) equals the empirical vector \( L_n^* \), and for \( v \in \mathcal{F} \) the energy function is given by the formula \( \Phi(v) = \frac{1}{2} \langle v, v \rangle \).

Finally, in the case of the Ising model \( \mathcal{F} \) equals the set of strictly stationary probability measures on the space \( \{1, -1\}^{2^d} \), \( \zeta_n \) equals the empirical field \( \Xi_n \), and for \( Q \in \mathcal{F} \) the energy function is given by formula (11.7). For numerous other models the hidden space, the hidden process, and the energy function can be identified.

We now return to the general case. The large deviation analysis of the general model is summarized in the next theorem.
THEOREM 12.1. We assume that for the given model there exists a hidden space \( \mathcal{X} \), a hidden process \( \{\zeta_n, n \in \mathbb{N}\} \), and an energy function \( \Phi \) and that the \( P_n \)-distributions of the hidden process satisfy the large deviation principle on \( \mathcal{X} \) with some rate function \( I \) and norming constants \( a_n \). The following conclusions hold.

(a) For each \( \beta > 0 \) the quantity \( \psi(\beta) \) exists and is given by the variational formula

\[
\psi(\beta) = \sup_{x \in \mathcal{X}} \{ \beta \Phi(x) - I(x) \}.
\]

(b) For each \( \beta > 0 \) the \( P_{n, \beta} \)-distributions of the hidden process \( \{\zeta_n, n \in \mathbb{N}\} \) satisfy the large deviation principle on \( \mathcal{X} \) with rate function

\[
I_\beta(x) = I(x) - \beta \Phi(x) - \inf_{y \in \mathcal{X}} \{ I(y) - \beta \Phi(y) \}
\]

and norming constants \( \{a_n, n \in \mathbb{N}\} \).

(c) For each \( \beta > 0 \) we define the set of equilibrium states

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

Then \( \mathcal{E}_\beta \) is a nonempty compact subset of \( \mathcal{X} \). In addition, if \( A \) is a Borel subset of \( \mathcal{X} \) whose closure has nonempty intersection with \( \mathcal{E}_\beta \), then

\[
\lim_{n \to \infty} P_{n, \beta} (\zeta_n \in A) = 0.
\]

Also, any subsequence of the positive integers has a subsequence \( \{n'\} \) such that the measures \( P_{n, \beta} (\zeta_n \in dx) \) have a weak limit as \( n' \to \infty \). Any weak limit of

\[
P_{n, \beta} (\zeta_n \in dx)
\]

has support in the set \( \mathcal{E}_\beta \).

Comments on the Proof. Parts (a) and (b) are consequences of parts (a) and (b) of Proposition 3.4. The first and second assertions in part (c) are consequences of part (b) of the present proposition and part (b) of Proposition 3.3. The third and fourth assertions in part (c) can be proved using the remark on page 49 of Parthasarathy (1967).

This completes our overview of the theory of large deviations and applications to statistical mechanics. We began this paper by presenting large deviation results for i.i.d. random variables taking values in a finite set. We then stated Cramér's Theorem and the Gärtner-Ellis Theorem and discussed the three levels of the Donsker-Varadhan theory. In the last part of the paper we applied each of the three levels of the Donsker-Varadhan theory to three models in statistical mechanics. The phase transition structure of each of these models is reflected in properties of the rate function that appears in the corresponding large deviation principle. The exploration of these phenomena is but one example of the many problems in which the theory of large deviations gives considerable insight.

REFERENCES


Address for correspondence:
Richard S. Ellis
Department of Mathematics and Statistics
University of Massachusetts
Amherst, MA 01003
email: rseliss@math.umass.edu