Chapter I

# Introduction to Large Deviations

## I.1. Overview

One of the common themes of probability theory and statistical mechanics is the discovery of regularity in the midst of chaos. The laws of probability theory, which include laws of large numbers and central limit theorems, summarize the behavior of a stochastic system in terms of a few parameters (e.g., mean and variance). In statistical mechanics, one derives macroscopic properties of a substance from a probability distribution that describes the complicated interactions among the individual constituent particles. A central concept linking the two fields is entropy.[1] The term was introduced into thermodynamics by Clausius in 1865 after many years of intensive work by him and others on the second law of thermodynamics. An early important step in its development and enrichment was the discovery by Boltzmann of a statistical interpretaton of entropy. Boltzmann's discovery, which was published in 1877, has three parts. We have augmented part (c) to include the possibility of phase transitions.

(a) Entropy is a measure of randomness or disorder in a statistical mechanical system.

(b) If $S$ is the entropy for a system in a given state and $W$ is the "thermodynamical probability" of that state,* then $S = k \log W$, where $k$ is a positive physical constant.

(c) The equilibrium states, which are the states of the system observed in nature, are those states with the largest thermodynamical probability and thus the largest entropy. By (a), they are the "most random" states of the system consistent with any constraints which the system must satisfy (e.g., conservation of energy). The existence of more than one equilibrium state corresponds to a phase transition.

All the notions of entropy discussed in this book are variations on the Boltzmann theme. In analyzing stochastic or statistical mechanical systems,

*The thermodynamical probability is defined to be the number of microstates compatible with the given state (see Wehrl (1978, page 223)).

one must extrapolate from a microscopic level, on which the system is defined, to a macroscopic level, on which the laws describing the behavior of the system are formulated. Boltzmann shows that entropy is the bridge between these two levels. We will illustrate these ideas and outline the main themes of this book in terms of a basic stochastic model. This discussion is intended for the reader who has a knowledge of probability theory consistent with Appendix A. Other readers may still perceive the global picture without turning to Appendix A at this time. That task may be postponed until the end of the first chapter.

Each of our systems is modeled microscopically by a collection of random variables $\{X_\alpha; \alpha \in \mathscr{A}\}$ which are defined on a probability space $(\Omega, \mathscr{F}, P)$ and which take values in a space $\Gamma$. $\Omega$ is a nonempty set, $\mathscr{F}$ is a $\sigma$-field of subsets of $\Omega$, and $P$ is a probability measure on $(\Omega, \mathscr{F})$. $\Gamma$ is called the *state space* or the *outcome space*; in all of our applications, $\Gamma$ is $\mathbb{R}^d$, $d \in \{1, 2, \ldots\}$, or a subset of $\mathbb{R}^d$. $\mathscr{A}$ is a suitable index set. Our results depend only on the distribution of the random variables $\{X_\alpha; \alpha \in \mathscr{A}\}$. Hence we may take $\Omega$ to be the product space $\Gamma^{\mathscr{A}}$ and the collection $\{X_\alpha; \alpha \in \mathscr{A}\}$ to be the coordinate representation process. That is, given a point $\omega = \{\omega_\alpha; \alpha \in \mathscr{A}\}$ in $\Gamma^{\mathscr{A}}$, we define $X_\alpha(\omega) = \omega_\alpha$, the $\alpha$th coordinate of $\omega$. Each $\omega \in \Gamma^{\mathscr{A}}$ represents a possible configuration or microstate of the system, and the entire space $\Gamma^{\mathscr{A}}$ is the set of all the configurations. The definition of the model is completed by specifying a probability measure $P$ on configuration space. Here are some examples.

**Example I.1.1.** (a) Let $\mathscr{A}$ be the set of integers $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$ and $\Gamma$ a finite set of distinct real numbers $\{x_1, x_2, \ldots, x_r\}$. Define $P$ to be an infinite product measure on $\Gamma^{\mathbb{Z}}$ with identical one-dimensional marginals, which we denote by $\rho$. Thus $\rho$ is a probability measure on $\Gamma$, and it has the form $\sum_{i=1}^r \rho_i \delta_{x_i}$, where $\rho_i > 0$, $\sum_{i=1}^r \rho_i = 1$, and $\delta_{x_i}$ is the unit point measure at $x_i$. The variables $\{X_j; j \in \mathbb{Z}\}$ are independent and identically distributed (i.i.d.) and each has distribution $\rho$. We write $P_\rho$ for $P$.

(b) A simple case of (a) is $\Gamma = \{0, 1\}$. The $\{X_j; j \in \mathbb{Z}\}$ are then Bernoulli trials and $\rho = \rho_0 \delta_0 + \rho_1 \delta_1$. The $\{X_j; j \in \mathbb{Z}\}$ may, for example, represent the successive outcomes of the toss of a coin in an infinite sequence of tosses separated by a constant time interval. Each outcome is recorded as 0 for a tail and 1 for a head. A fair coin corresponds to $\rho = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$.

(c) In (a), the set $\Gamma = \{x_1, x_2, \ldots, x_r\}$ may represent a set of possible velocities of the molecules of an ideal gas which are constrained to move in an interval and which undergo elastic collisions at the endpoints. Then $X_j$ denotes the velocity of the molecule labeled $j$. A configuration $\omega \in \Gamma^{\mathbb{Z}}$ is a specification of the velocities for each molecule. Independence means that the molecules do not interact. This model is treated in Chapter III.

(d) A ferromagnet is modeled by random configurations of spins (microscopic magnets) at sites in the $D$-dimensional integer lattice $\mathbb{Z}^D$, $D \in$

$\{1, 2, \ldots\}$. We set $\mathscr{A} = \mathbb{Z}^D$ and $\Gamma = \{1, -1\}$. The values 1 and $-1$ represent "spin-up" and "spin-down," respectively. Configuration space $\Omega$ is $\{1, -1\}^{\mathbb{Z}^D}$ and $X_j(\omega)$ is the spin at site $j \in \mathbb{Z}^D$ for the configuration $\omega$. A product measure on $\Omega$ is not appropriate for this model since the spins at different sites interact. The ferromagnet is modeled by a probability measure $P$ on $\Omega$ which is translation invariant; i.e., invariant with respect to spatial translations in $\mathbb{Z}^D$. Ferromagnetic models are the subject of Chapters IV and V.

As these examples show, a probability measure on a configuration space provides a microscopic definition of a stochastic or physical system. However, the laws describing the behavior of such a system are macroscopic descriptions which in contrast to the number of all configurations, involve many fewer variables. For each description, the range of possible values of these variables defines the set of macrostates. Each macrostate is compatible with, and hence is a summary of, many microstates. The entropy of a macrostate is a measure of this multiplicity. Those macrostates compatible with the most microstates—i.e., those with the largest entropy—are the ones observed in nature. Generally, a system will have several possible macroscopic descriptions, each differing in complexity and in choice of macrostate. For each description, there is a different entropy concept.

We return to the coin tossing model in order to explain these ideas. This model is represented by the infinite product measure $P_\rho$ on the configuration space $\Omega = \{0, 1\}^{\mathbb{Z}}$ ($\rho = \rho_0 \delta_0 + \rho_1 \delta_1$). Macroscopically, the behavior of the coin can be expressed by a single number, its mean value. The possible mean values are all numbers $z \in (0, 1)$, and there is no harm including the endpoints. We call the set of $z \in [0, 1]$ macrostates. The weak law of large numbers (WLLN) enables one to estimate the macrostate in terms of microstates. Define $S_n(\omega) = \sum_{j=1}^n X_j(\omega)$ for $n = 1, 2, \ldots$ and $\omega \in \Omega$. $S_n(\omega)/n$ is the average value of the tosses $\omega_1, \omega_2, \ldots, \omega_n$. The sum $S_n(\omega)/n$ is called a *microscopic sum* or *n-sum*. Let $m_\rho$ be the mean value of the measure $\rho$ ($m_\rho = 0 \cdot \rho_1 + 1 \cdot \rho_1 = \rho_1$) and $Q_n$ the distribution of $S_n/n$. The WLLN says that for any $\varepsilon > 0$,

$$Q_n\{(m_\rho - \varepsilon, m_\rho + \varepsilon)\} = P_\rho\{\omega \in \Omega : S_n(\omega)/n \in (m_\rho - \varepsilon, m_\rho + \varepsilon)\} \to 1$$

as $n \to \infty$.

In other words, if $n$ is large, then with respect to $P_\rho$ essentially all microscopic $n$-sums are close to the macrostate $m_\rho$. The latter is called the *equilibrium state*.

Here is how entropy arises. Assume for simplicity that the coin is fair ($m_\rho = \frac{1}{2}$). For any $z \in \mathbb{R}$ and $\varepsilon > 0$, let $A_{z,\varepsilon}$ be the interval $(z - \varepsilon, z + \varepsilon)$. By the WLLN, $Q_n\{A_{m_\rho, \varepsilon}\} \to 1$ as $n \to \infty$ while if $z \neq m_\rho$ and $0 < \varepsilon < |z - m_\rho|$, then $Q_n\{A_{z,\varepsilon}\} \to 0$. In the latter case, it is not hard to refine the WLLN and to

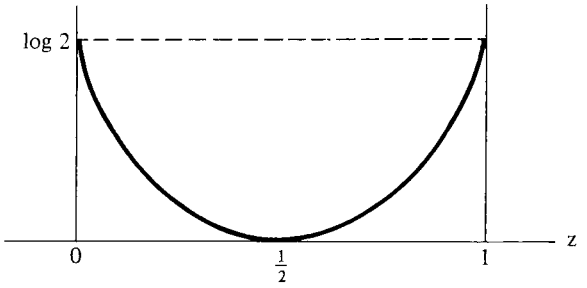


Figure I.1. The entropy function $I(z)$ for the coin tossing model.

prove that $Q_n\{A_{z,\varepsilon}\}$ decays to 0 exponentially fast. The exponential rate of decay is defined by $F(z,\varepsilon) = -\lim_{n\to\infty} n^{-1}\log Q_n\{A_{z,\varepsilon}\}$.* By a simple combinatoric argument (as in the proof of Theorem I.3.1), one shows that $F(z,\varepsilon)$ equals the infimum over $A_{z,\varepsilon}$ of the function

$$I(z) = \begin{cases} z\log(2z) + (1-z)\log(2(1-z)) & \text{for } z\in[0,1], \\ \infty & \text{for } z\notin[0,1], \end{cases} \tag{1.1}$$

where $0\log 0 = 0$. The graph of the non-negative convex function $I(z)$ is shown in Figure I.1; clearly, $F(z,\varepsilon) > 0$. $I(z)$ is called the *entropy function* for the coin tossing model. We now interpret it in view of our earlier remarks on entropy.

For any $z\in\mathbb{R}$ and large $n$, $Q_n\{A_{z,\varepsilon}\}$ is approximately $\exp(-nF(z,\varepsilon))$. Since $F(z,\varepsilon)\to I(z)$ as $\varepsilon\to 0$, we may heuristically write

$$Q_n\{A_{z,\varepsilon}\} \approx \exp(-nI(z)) \tag{1.2}$$

for $n$ large and $\varepsilon$ small. If $z\neq m_\rho$, then $I(z)$ is positive and $\exp(-nI(z))\to 0$ as $n\to\infty$. This is consistent with the exponential decay of $Q_n\{A_{z,\varepsilon}\}$ for $0<\varepsilon<|z-m_\rho|$. The heuristic formula (1.2) shows that to a small value of $I(z)$ there corresponds a large probability $Q_n\{A_{z,\varepsilon}\}$ or, in other words, a high multiplicity of microstates. In this sense, $I(z)$ is a measure of the multiplicity of microstates compatible with the macrostate $z$. For another interpretation, given points $z_1$ and $z_2$ in $[0,1]$, it is reasonable to call $z_1$ more random than $z_2$ if $I(z_1) < I(z_2)$; that is, if there are more microstates compatible with $z_1$ than with $z_2$. Thus $I(z)$ also measures the randomness of the macrostate $z$. The equilibrium state $m_\rho = \frac{1}{2}$ is that macrostate which is compatible with the most microstates. In fact,

$$I(m_\rho) = 0 = \min\{I(z): z\in\mathbb{R}\} \quad \text{and} \quad I(z) > 0 \qquad \text{for } z\neq m_\rho.$$

Thus the equilibrium state, being the unique minimum point of $I$, is the most random macrostate. Points $z$ outside of $[0,1]$ are forbidden values for $S_n/n$:

*If $Q_n\{A_{z,\varepsilon}\} = 0$, then set $\log Q_n\{A_{z,\varepsilon}\} = -\infty$.

if $A \cap [0,1]$ is empty, then $Q_n\{A\} = 0$ and $I(z) = \infty$ for each $z \in A$. For $z \neq m_\rho$ and $0 < \varepsilon < |z - m_\rho|$, $Q_n\{A_{z,\varepsilon}\}$ is called a *large deviation probability* since the event $\{\omega \in \Omega : S_n(\omega)/n \in A_{z,\varepsilon}\}$ corresponds to a fluctuation or deviation of $S_n/n$ of order $|z - m_\rho|$ away from the limiting mean. It is a very rare event since $Q_n\{A_{z,\varepsilon}\} \to 0$ exponentially.

An equivalent statement of the WLLN is that the distributions $\{Q_n;$ $n = 1, 2, \ldots\}$ converge weakly to the unit point measure at $m_\rho$ (written $Q_n \Rightarrow \delta_{m_\rho}$). In this book, we will study much more general but analogous situations. A sequence of probability measures $\{Q_n; n = 1, 2, \ldots\}$ on a complete separable metric space $\mathscr{X}$ will converge weakly to the unit point measure at some $x_0 \in \mathscr{X}$. $\{Q_n\}$ will have a large deviation property in the sense that $Q_n\{K\}$ will decay exponentially for all closed sets $K$ in $\mathscr{X}$ which do not contain $x_0$. The decay rate will be given by $-\inf_{x \in K} I(x)$, where $I(x)$ is some non-negative function on $\mathscr{X}$ with a unique minimum point at $x_0$ ($I(x_0) = 0$). $I(x)$ is called the *entropy function of the measures* $\{Q_n\}$.

In a series of important papers beginning in 1975, Donsker and Varadhan have identified three levels of large deviations which fit into the general framework just described. These levels will be treated in detail in this book. Let $\{X_j; j \in \mathbb{Z}\}$ be a sequence of i.i.d. random vectors taking values in $\mathbb{R}^d$. Let $\rho$ be the distribution of $X_1$ and $P_\rho$ the corresponding infinite product measure on $\Omega = (\mathbb{R}^d)^{\mathbb{Z}}$.

**Level-1.** Define $Q_n^{(1)}$ to be the distribution of $S_n(\omega)/n = \sum_{j=1}^n X_j(\omega)/n$ on $\mathbb{R}^d$ and assume that $\int_{\mathbb{R}^d} \|x\| \rho(dx)$ is finite. Then by the WLLN the sequence $\{Q_n^{(1)}; n = 1, 2, \ldots\}$ converges weakly to $\delta_{m_\rho}$, where $m_\rho$ is the mean $\int_{\mathbb{R}^d} x\rho(dx)$ of $\rho$. If, furthermore, the moment generating function $\int_{\mathbb{R}^d} \exp\langle t, x\rangle \rho(dx)$ is finite for all $t \in \mathbb{R}^d$, then $Q_n^{(1)}\{K\}$ decays exponentially for all closed subsets $K$ of $\mathbb{R}^d$ which do not contain $m_\rho$ [Theorem II.4.1]. The decay rate is given in terms of a level-1 entropy function $I_\rho^{(1)}$ which is a non-negative convex function on $\mathbb{R}^d$ and which has a unique minimum point at $m_\rho$ ($I_\rho^{(1)}(m_\rho) = 0$). The sums $\{S_n(\omega)/n\}$ are called *level-1 microscopic n-sums* and points $z \in \mathbb{R}^d$ are called *level-1 macrostates*. The mean $m_\rho$ is the unique level-1 equilibrium state.

**Level-2.** For $\omega \in \Omega$, define the empirical measure $L_n(\omega, \cdot) = n^{-1} \sum_{j=1}^n \delta_{X_j(\omega)}(\cdot)$ (for $A \subseteq \mathbb{R}^d$ a Borel set, $L_n(\omega, A) = n^{-1} \sum_{j=1}^n \delta_{X_j(\omega)}\{A\}$). $L_n(\omega)$ takes values in the space $\mathscr{M}(\mathbb{R}^d)$ of probability measures on $\mathbb{R}^d$. $\mathscr{M}(\mathbb{R}^d)$ is a topological space with respect to the topology of weak convergence, and it is metrizable as a complete separable metric space. Let $Q_n^{(2)}$ denote the distribution of $L_n(\omega, \cdot)$ on $\mathscr{M}(\mathbb{R}^d)$. By the ergodic theorem, the sequence $\{L_n(\omega, \cdot); n = 1, 2, \ldots\}$ converges weakly to $\rho$ (almost surely), and this implies that $\{Q_n^{(2)}; n = 1, 2, \ldots\}$ converges weakly to $\delta_\rho$. In addition, $Q_n^{(2)}\{K\}$ decays exponentially for all closed subsets $K$ of $\mathscr{M}(\mathbb{R}^d)$ which do not contain $\rho$ [Theorem II.4.3]. The decay rate is given in terms of a level-2 entropy func-

tion $I_\rho^{(2)}$ which is a non-negative convex function on $\mathcal{M}(\mathbb{R}^d)$ and which has a unique minimum point at $\rho$ ($I_\rho^{(2)}(\rho) = 0$). For $\nu \in \mathcal{M}(\mathbb{R}^d)$, $I_\rho^{(2)}(\nu)$ equals the relative entropy of $\nu$ with respect to $\rho$. The empirical measures $\{L_n(\omega, \cdot)\}$ are called *level-2 microscopic n-sums* and measures $\nu \in \mathcal{M}(\mathbb{R}^d)$ are called *level-2 macrostates*. The distribution $\rho$ is the unique level-2 equilibrium state.

**Level-3.** Let $\mathcal{M}_s(\Omega)$ denote the set of strictly stationary probability measures on $\Omega$. $\mathcal{M}_s(\Omega)$ is a topological space with respect to the topology of weak convergence, and it is metrizable as a complete separable metric space. For $\omega \in \Omega$, one defines the so-called empirical process $R_n(\omega, \cdot)$ which takes values in $\mathcal{M}_s(\Omega)$ [see page 22]. Let $Q_n^{(3)}$ denote the distribution of $R_n(\omega, \cdot)$ on $\mathcal{M}_s(\Omega)$. By the ergodic theorem, the sequence $\{R_n(\omega, \cdot); n = 1, 2, \dots\}$ converges weakly to $P_\rho$ (almost surely). Hence $\{Q_n^{(3)}; n = 1, 2, \dots\}$ converges weakly to $\delta_{P_\rho}$. In addition, $Q_n^{(3)}\{K\}$ decays exponentially for all closed subsets $K$ of $\mathcal{M}_s(\Omega)$ which do not contain $P_\rho$ [Theorem II.4.4]. The decay rate is given in terms of a level-3 entropy function $I_\rho^{(3)}$ which is a non-negative affine function on $\mathcal{M}_s(\Omega)$ and which has a unique minimum point at $P_\rho$ ($I_\rho^{(3)}(P_\rho) = 0$). The empirical processes $\{R_n(\omega, \cdot)\}$ are called *level-3 microscopic n-sums* and measures $P \in \mathcal{M}_s(\Omega)$ are called *level-3 macrostates*. The measure $P_\rho$ is the unique level-3 equilibrium state.

For each of the three levels, we may heuristically express the asymptotic behavior of the distributions $Q_n^{(i)}(dx)$ by the formula $\exp(-nI_\rho^{(i)}(x))dx$. $Q_n^{(i)}(dx)$ is a measure on the complete separable metric space $\mathcal{X} = \mathbb{R}^d$, $\mathcal{M}(\mathbb{R}^d)$, or $\mathcal{M}_s(\mathbb{R}^d)$ for $i = 1$, 2, or 3, respectively. By analogy with coin tossing, we may interpret each entropy function $I_\rho^{(i)}(x)$ as a measure of the multiplicity of microstates compatible with the macrostate $x \in \mathcal{X}$. In that sense, $I_\rho^{(i)}(x)$ is also a measure of the randomness of $x$.

Varadhan (1966) gave a useful application of the large deviation property to calculate the asymptotics of certain integrals. The heuristic formula $Q_n^{(i)}(dx) \approx \exp(-nI_\rho^{(i)}(x))dx$ suggests that if $F$ is a bounded continuous function on $\mathcal{X}$, then

$$\lim_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{X}} \exp(nF(x)) Q_n^{(i)}(dx) = \lim_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{X}} \exp[n(F(x) - I_\rho^{(i)}(x))]dx$$
$$= \sup_{x \in \mathcal{X}} \{F(x) - I_\rho^{(i)}(x)\}.$$

This limit is valid under suitable hypotheses [Theorem II.7.1] and will be applied a number of times in the book.

So far we have discussed large deviations for i.i.d. random vectors. Statistical mechanical systems have a similar three-level structure with one additional feature: there need not be a unique equilibrium state for a given level. This lack of a unique equilibrium state corresponds physically to a

phase transition and probabilistically to a breakdown in the law of large numbers (or ergodic theorem) for the corresponding microscopic $n$-sums. But in general, whether or not there is a phase transition, we may consider the Legendre–Fenchel transform of the corresponding entropy function. This transform defines a convex function which in statistical mechanics is called the free energy. Free energy functions will play a central role in analyzing statistical mechanical systems in Chapters III–V.

In the remainder of this chapter we will introduce level-1, 2, and 3 large deviations by considering i.i.d. random variables with a finite state space. The corresponding entropy functions will be calculated by means of elementary combinatorics. In Chapter II, the three levels of large deviations will be generalized to i.i.d. random vectors taking values in $\mathbb{R}^d$. Section II.6 presents additional large deviation results, which are particularly suited for applications to statistical mechanics. The proofs of the theorems in Chapter II are detailed and will be postponed until Chapters VI–IX. In Chapters III–V the large deviation results will be applied to an ideal gas model and to ferromagnetic spin models in statistical mechanics.

## I.2. Large Deviations for I.I.D. Random Variables with a Finite State Space

In its simplest form the theory of large deviations refines the classical law of large numbers. Let $S_n$ be the $n$th partial sum of independent, identically distributed random variables $X_1, X_2, \ldots$. The strong law of large numbers states that if the expectation $E\{|X_1|\}$ is finite, then $S_n/n$ converges to $E\{X_1\}$ almost surely. This was proved by Kolmogorov (1930). It implies the weak law, which states that $S_n/n$ converges to $E\{X_1\}$ in probability. The first large deviation results were those of Cramér (1938) and Chernoff (1952). They showed that if $X_1$ has a finite moment generating function in a neighborhood of 0, then the probability that $S_n/n$ deviates from $E\{X_1\}$ by a small amount $\varepsilon > 0$ is exponentially small as $n \to \infty$. After Chernoff, these results were applied and extended in statistics and probability by many people, and they have played a key role in information theory. In a series of papers starting in 1975, Donsker and Varadhan have generalized these results to Markov processes with general state spaces and have found many interesting new applications.

The Donsker–Varadhan theory identifies three levels of large deviations, which were mentioned in the previous section. An elementary way of introducing the three levels is by means of well-known but instructive examples involving i.i.d. random variables with a finite state space. The rest of this chapter focuses upon these examples. Later chapters will generalize the large deviation results beyond this elementary setting.

Let $r \geq 2$ be an integer and consider a finite set $\Gamma = \{x_1, x_2, \ldots, x_r\}$, where $x_1 < x_2 < \cdots < x_r$ are real numbers. Let $\mathscr{B}(\Gamma)$ denote the set of all subsets of $\Gamma$. We fix a probability measure $\rho$ on $\mathscr{B}(\Gamma)$ for which $\rho_i = \rho\{x_i\} > 0$ for each $x_i \in \Gamma$. Thus $\rho$ has the form $\sum_{i=1}^r \rho_i \delta_{x_i}$. For $A$ a subset of $\Gamma$ $\rho\{A\}$ equals $\sum_{i=1}^r \rho_i \delta_{x_i}\{A\}$, where $\delta_{x_i}\{A\}$ equals 1 if $x_i \in A$ and equals 0 if $x_i \notin A$. Denote by $\omega$ the doubly infinite sequence $(\ldots, \omega_{-2}, \omega_{-1}, \omega_0, \omega_1, \omega_2, \ldots)$ with each $\omega_j \in \Gamma$. Configuration space $\Omega$ is the set of all such sequences; thus $\Omega = \Gamma^{\mathbb{Z}}$. Let $P_\rho$ be the infinite product measure on $\Omega$ with identical one-dimensional marginals $\rho$. To a cylinder set of the form

$$\Sigma = \{\omega \in \Omega : \omega_{m+1} \in F_1, \ldots, \omega_{m+k} \in F_k\}, \quad F_1, \ldots, F_k \text{ subsets of } \Gamma, \tag{1.3}$$

$P_\rho$ assigns the probability $P_\rho\{\Sigma\} = \prod_{j=1}^k \rho\{F_j\}$. $P_\rho$ is uniquely determined by these probabilities.* For each integer $j$ define the coordinate function $X_j$ on $\Omega$ by $X_j(\omega) = \omega_j$. The functions $\{X_j; j \in \mathbb{Z}\}$ form a sequence of i.i.d. random variables with finite state space $\Gamma$ and distribution $\rho$.

Until Section I.5, we will work with level-1 and level-2. These are defined by two random quantities, called level-1 and level-2 (microscopic) $n$-sums. The level-1 $n$-sum is the average value $S_n(\omega)/n = \sum_{j=1}^n X_j(\omega)/n$, $n = 1, 2, \ldots$. The level-2 $n$-sum is defined in terms of the empirical frequency $L_{n,i}(\omega)$ with which $x_i$ appears in the sequence $X_1(\omega), \ldots, X_n(\omega)$:

$$L_{n,i}(\omega) = \frac{1}{n} \sum_{j=1}^n \delta_{X_j(\omega)}\{x_i\}, \qquad n = 1, 2, \ldots, \quad i = 1, \ldots, r. \tag{1.4}$$

For each $\omega$, the numbers $\{L_{n,i}(\omega);\ i = 1, \ldots, r\}$ define a probability measure $L_n(\omega, \cdot)$ on the set of all subsets of $\Gamma$. For $A$ a subset of $\Gamma$,

$$L_n(\omega, A) = \sum_{x_i \in A} L_{n,i}(\omega) = \frac{1}{n} \sum_{j=1}^n \delta_{X_j(\omega)}\{A\},^\dagger \tag{1.5}$$

where $\delta_{X_j(\omega)}\{A\}$ equals 1 if $X_j(\omega) \in A$ and equals 0 if $X_j(\omega) \notin A$. The measure $L_n(\omega, \cdot)$ is the level-2 $n$-sum. It is called the *empirical measure* corresponding to $X_1(\omega), \ldots, X_n(\omega)$. The average value $S_n(\omega)/n$ can be calculated by multiplying each $x_i$ by the empirical frequency $L_{n,i}(\omega)$ and summing over $\Gamma$; i.e.,

$$S_n(\omega)/n = \sum_{i=1}^r x_i L_{n,i}(\omega). \tag{1.6}$$

The right-hand side is the mean of the empirical measure.

With respect to the measure $P_\rho$, the asymptotic behavior of $S_n(\omega)/n$ and of $L_n(\omega, \cdot)$ follows from the law of large numbers. Indeed the summands in $S_n(\omega)/n$ are i.i.d. with mean

$$m_\rho = \int_\Omega X_j(\omega) P_\rho(d\omega) = \sum_{i=1}^r x_i \rho_i,$$

*Appendix A summarizes all the properties of probability measures that are needed in the text.

†The sum over an empty set is defined to be 0.

while the summands in $L_{n,i}(\omega)$ are i.i.d. with mean

$$\int_\Omega \delta_{X_j(\omega)}\{x_i\} P_\rho(d\omega) = P_\rho\{\omega \in \Omega: X_j(\omega) = x_i\} = \rho_i.$$

Hence for any $\varepsilon > 0$

$$\begin{aligned} &\lim_{n\to\infty} P_\rho\{\omega \in \Omega: |S_n(\omega)/n - m_\rho| \geq \varepsilon\} = 0, \\ &\lim_{n\to\infty} P_\rho\{\omega \in \Omega: \max_{i=1,\ldots,r} |L_{n,i}(\omega) - \rho_i| \geq \varepsilon\} = 0. \end{aligned} \tag{1.7}$$

The vector $(\rho_1, \rho_2, \ldots, \rho_r)$ is the limiting mean of the random vector $(L_{n,i}(\omega), \ldots, L_{n,r}(\omega))$. The probabilities in (1.7) represent large deviations since they involve fluctuations of order $\varepsilon$ of the respective $n$-sums away from the limiting means, and $\varepsilon$ is fixed. Below we show by elementary combinatorial arguments that each of these probabilities decays exponentially.

## I.3. Levels-1 and 2 for Coin Tossing

Coin tossing is defined by the state space $\Gamma = \{0, 1\}$ and the measure $\rho = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. The value 0 represents a tail and 1 a head. The proof for this simple case will set a pattern of proof for the more general large deviation results which follow. We have $\rho_1 = \rho_2 = \frac{1}{2}$, $m_\rho = \frac{1}{2}$, $L_{n,1}(\omega) = 1 - S_n(\omega)/n$, and $L_{n,2}(\omega) = S_n(\omega)/n$. Hence for each $\omega$

$$|L_{n,1}(\omega) - \rho_1| = |L_{n,2}(\omega) - \rho_2| = |S_n(\omega)/n - m_\rho|,$$

and so the level-1 and 2 probabilities coincide:

$$P_\rho\{|S_n/n - m_\rho| \geq \varepsilon\} = P_\rho\{\max_{i=1,2} |L_{n,i} - \rho_i| \geq \varepsilon\}. \tag{1.8}$$

Let $Q_n^{(1)}$ be the $P_\rho$-distribution of $S_n/n$ on $\mathbb{R}$ and define the closed set

$$A = \{z \in \mathbb{R}: |z - m_\rho| \geq \varepsilon\}, \qquad \text{where } 0 < \varepsilon < \tfrac{1}{2}.$$

The set $A \cap [0, 1]$ is nonempty and $Q_n^{(1)}\{A\} = P_\rho\{|S_n/n - m_\rho| \geq \varepsilon\}$ is positive for all sufficiently large $n$. Since $A$ does not contain $m_\rho$, $Q_n^{(1)}\{A\} \to 0$. According to the next theorem, $Q_n^{(1)}\{A\}$ decays exponentially, and the decay rate is given in terms of the entropy function

$$I_\rho^{(1)}(z) = \begin{cases} z\log(2z) + (1-z)\log(2(1-z)) & \text{for } z \in [0, 1], \\ \infty & \text{for } z \notin [0, 1], \end{cases} \tag{1.9}$$

where $0 \log 0 = 0$. $I_\rho^{(1)}(z)$ is convex, is symmetric about $z = m_\rho = \frac{1}{2}$, and attains its minimum value of 0 at the unique point $z = m_\rho$. $I_\rho^{(1)}$ is depicted in Figure I.1.

**Theorem I.3.1.**

$$\lim_{n\to\infty}\frac{1}{n}\log Q_n^{(1)}\{A\} = \lim_{n\to\infty}\frac{1}{n}\log P_\rho\{|S_n/n - m_\rho| \geq \varepsilon\} = -\min_{z\in A} I_\rho^{(1)}(z).$$

*Since the set $A$ is closed and does not contain $m_\rho$, $\min_{z\in A} I_\rho^{(1)}(z) > I_\rho^{(1)}(m_\rho) = 0$. Hence $Q_n^{(1)}\{A\}$ converges to zero exponentially fast as $n \to \infty$.*

*Proof.* Let $\Omega_n$ be the finite configuration space consisting of all sequences $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$, with each $\omega_j \in \Gamma = \{0, 1\}$; thus $\Omega_n = \Gamma^n$. If $\pi_n P_\rho$ is the finite product measure on $\Omega_n$ with identical one-dimensional marginals $\rho$, then $Q_n^{(1)}\{A\} = \pi_n P_\rho\{\omega \in \Omega_n : S_n(\omega)/n \in A\}$. For fixed $n$ and $\omega \in \Omega_n$, $S_n(\omega)$ may take any value $k \in \{0, 1, \ldots, n\}$. $S_n(\omega)/n$ is in $A$ if and only if $k$ is in the set $A_n = \{k \in \{0, 1, \ldots, n\} : |k/n - \frac{1}{2}| \geq \varepsilon\}$. For $k \in A_n$, define $C(n, k) = n!/(k!(n-k)!)$. There are $C(n, k)$ points $\omega$ in $\Omega_n$ for which $S_n(\omega) = k$, and $\pi_n P_\rho\{\omega\} = 2^{-n}$ for each $\omega \in \Omega_n$. Hence

$$(1.10)\quad Q_n^{(1)}\{A\} = \sum_{k\in A_n} \pi_n P_\rho\{\omega \in \Omega_n : S_n(\omega)/n = k/n\} = \sum_{k\in A_n} C(n,k)\frac{1}{2^n}.$$

Since there are no more than $n + 1$ terms in the sum,

$$\max_{k\in A_n} C(n,k)\frac{1}{2^n} \leq Q_n^{(1)}\{A\} \leq (n+1)\max_{k\in A_n} C(n,k)\frac{1}{2^n},$$

and since log is an increasing function

$$(1.11)\quad \max_{k\in A_n}\left[\frac{1}{n}\log\left(C(n,k)\frac{1}{2^n}\right)\right] \leq \frac{1}{n}\log Q_n^{(1)}\{A\} \leq \frac{\log(n+1)}{n} + \max_{k\in A_n}\left[\frac{1}{n}\log\left(C(n,k)\frac{1}{2^n}\right)\right].$$

Thus the asymptotic behavior of $Q_n^{(1)}\{A\}$ is governed by the asymptotic behavior of the largest summand in (1.10). Entropy arises by the following lemma.

**Lemma I.3.2.** *Uniformly in $k \in \{0, 1, \ldots, n\}$,*

$$\frac{1}{n}\log C(n,k) = -\frac{k}{n}\log\frac{k}{n} - \left(1 - \frac{k}{n}\right)\log\left(1 - \frac{k}{n}\right) + O\left(\frac{\log n}{n}\right) \qquad \text{as } n \to \infty.$$

*Proof.* Since $C(n, 0) = C(n, n) = 1$ and $C(n, 1) = C(n, n-1) = n$, the lemma holds for all $n \geq 1$ and $k = 0, 1, n-1, n$. A weak form of Stirling's approximation states that for all $n \geq 2$, $\log(n!) = n\log n - n + \beta_n$, where $|\beta_n| = O(\log n)$ [Problem I.8.1]. Hence for $2 \leq k \leq n-2$,

$$\frac{1}{n}\log C(n,k) = \log n - \frac{k}{n}\log k - \frac{n-k}{n}\log(n-k) + \frac{1}{n}(\beta_n - \beta_k - \beta_{n-k}).$$

Write

$$\log n = -\frac{k}{n}\log\frac{1}{n} - \frac{n-k}{n}\log\frac{1}{n}$$

and combine these terms with the other log terms to give

$$\frac{1}{n}\log C(n,k) = -\frac{k}{n}\log\frac{k}{n} - \left(1-\frac{k}{n}\right)\log\left(1-\frac{k}{n}\right) + \frac{1}{n}(\beta_n - \beta_k - \beta_{n-k}).$$

For $2 \le k \le n-2$, the last term can be bounded by $O(n^{-1}\log n)$ uniformly in $k$. This completes the proof. □

The lemma shows that

$$\frac{1}{n}\log\left(C(n,k)\frac{1}{2^n}\right) = \log\frac{1}{2} - \frac{k}{n}\log\frac{k}{n} - \left(1-\frac{k}{n}\right)\log\left(1-\frac{k}{n}\right) + O\left(\frac{\log n}{n}\right).$$

The first three terms are exactly $-I_\rho^{(1)}(k/n)$, where $I_\rho^{(1)}$ is defined in (1.9). Thus

$$\frac{1}{n}\log Q_n^{(1)}\left\{\frac{k}{n}\right\} = \frac{1}{n}\log\left(C(n,k)\frac{1}{2^n}\right) = -I_\rho^{(1)}\left(\frac{k}{n}\right) + O\left(\frac{\log n}{n}\right). \tag{1.12}$$

Since $n^{-1}\log n$ and $n^{-1}\log(n+1)$ both tend to zero, we have by (1.11)

$$\lim_{n\to\infty}\frac{1}{n}\log Q_n^{(1)}\{A\} = \lim_{n\to\infty}\max_{k\in A_n}\left(-I_\rho^{(1)}\left(\frac{k}{n}\right)\right) = -\lim_{n\to\infty}\min_{k\in A_n} I_\rho^{(1)}\left(\frac{k}{n}\right). \tag{1.13}$$

For each $n$ the set $\{z\in[0,1]: z = k/n \text{ for some } k\in A_n\}$ is a subset of $A\cap[0,1]$. Since $I_\rho^{(1)}(z) = \infty$ for $z\notin[0,1]$, we conclude using Problem I.8.2 that

$$\lim_{n\to\infty}\frac{1}{n}\log Q_n^{(1)}\{A\} = -\min_{z\in A\cap[0,1]} I_\rho^{(1)}(z) = -\min_{z\in A} I_\rho^{(1)}(z). \qquad \square$$

## I.4. Levels-1 and 2 for I.I.D. Random Variables with a Finite State Space

In general, the state space $\Gamma$ equals $\{x_1, x_2, \ldots, x_r\}$, where $x_1 < x_2 < \cdots < x_r$ are real numbers. The exponential decay rates of the two probabilities in (1.7) are expressed in terms of a function called the relative entropy. Let $\mathscr{B}(\Gamma)$ denote the set of all subsets of $\Gamma$ and $\mathscr{M}(\Gamma)$ the set of probability measures on $\mathscr{B}(\Gamma)$. Each $\nu\in\mathscr{M}(\Gamma)$ has the form $\nu = \sum_{i=1}^r \nu_i\delta_{x_i}$, where $\nu_i \ge 0$ and $\sum_{i=1}^r \nu_i = 1$. $\mathscr{M}(\Gamma)$ may be identified with the compact convex subset of $\mathbb{R}^r$ consisting of all vectors $\nu = (\nu_1, \ldots, \nu_r)$ which satisfy $\nu_i \ge 0$ and $\sum_{i=1}^r \nu_i = 1$. The *relative entropy*[2] of $\nu$ with respect to the measure $\rho = \sum_{i=1}^r \rho_i\delta_{x_i}$ $(\rho_i > 0)$ is defined by

$$I_\rho^{(2)}(\nu) = \sum_{i=1}^r \nu_i \log\frac{\nu_i}{\rho_i} \qquad \text{where } 0\log 0 = 0.$$

We have the following properties.

**Proposition I.4.1.** (a) *$I_\rho^{(2)}(\nu)$ is a convex function of $\nu \in \mathscr{M}(\Gamma)$.*

(b) *$I_\rho^{(2)}(\nu)$ measures the discrepancy between $\nu$ and $\rho$ in the sense that $I_\rho^{(2)}(\nu) \geq 0$ with equality if and only if $\nu = \rho$. Thus $I_\rho^{(2)}(\nu)$ attains its infimum over $\mathscr{M}(\Gamma)$ at the unique measure $\nu = \rho$.*

*Proof.* (a) $I_\rho^{(2)}(\nu)$ equals $\sum_{i=1}^r k(\nu_i/\rho_i)\rho_i$, where $h(x)$ is the convex function $x \log x$, $x \geq 0$. Let $\mu$ and $\nu$ be probability measures on $\mathscr{B}(\Gamma)$. Then for $0 \leq \lambda \leq 1$

$$\begin{aligned} I_\rho^{(2)}(\lambda\mu + (1-\lambda)\nu) &= \sum_{i=1}^r h(\lambda\mu_i/\rho_i + (1-\lambda)\nu_i/\rho_i)\rho_i \\ &\leq \lambda \sum_{i=1}^r h(\mu_i/\rho_i)\rho_i + (1-\lambda)\sum_{i=1}^r h(\nu_i/\rho_i)\rho_i \\ &= \lambda I_\rho^{(2)}(\mu) + (1-\lambda) I_\rho^{(2)}(\nu). \end{aligned}$$

(b) For any $x \geq 0$, $x \log x \geq x - 1$ with equality iff $x = 1$. Hence

$$\frac{\nu_i}{\rho_i} \log \frac{\nu_i}{\rho_i} \geq \frac{\nu_i}{\rho_i} - 1 \tag{1.14}$$

with equality iff $\nu_i = \rho_i$. Multiplying this inequality by $\rho_i$ and summing over $i$ yields

$$I_\rho^{(2)}(\nu) = \sum_{i=1}^r \nu_i \log \frac{\nu_i}{\rho_i} \geq 0.$$

$I_\rho^{(2)}(\nu)$ equals 0 iff equality holds in (1.14) for each $i$; $\nu_i$ equals $\rho_i$ for each $i$ iff $\nu$ equals $\rho$. □

We single out an important special case of relative entropy.

**Example I.4.2.** If $\rho$ is the uniform measure on $\Gamma = \{x_1, x_2, \ldots, x_r\}$ ($\rho_i = 1/r$ for each $i$), then $I_\rho^{(2)}(\nu) = \log r + \sum_{i=1}^r \nu_i \log \nu_i$. The quantity $H(\nu) = -\sum_{i=1}^r \nu_i \log \nu_i$ is called the *Shannon entropy* of $\nu$.[3] Since $-\nu_i \log \nu_i \geq 0$, $H(\nu)$ is non-negative. We show that $H(\nu)$ is a measure of the randomness in $\nu$. By Proposition I.4.1, $H(\nu) = \log r - I_\rho^{(2)}(\nu) \leq \log r$; $H(\nu) = \log r$ iff $I_\rho^{(2)}(\nu) = 0$ and this holds iff each $\nu_i = \rho_i = 1/r$. Hence $H(\nu)$ attains its maximum value of $\log r$ iff $\nu$ equals the uniform measure $\rho$. The measure $\rho$ is in a sense the most random probability measure on $\mathscr{B}(\Gamma)$. At the other extreme, $H(\nu)$ equals 0 iff one of the $\nu_i$'s, say $\nu_{i'}$, is 1 and the other $\nu_i$'s, $i \neq i'$, are 0. The corresponding measures $\delta_{x_{i'}}$ are the least random probability measures on $\mathscr{B}(\Gamma)$.

We now turn to the large deviation results. For each $\omega \in \Omega$, the empirical measure $L_n(\omega, \cdot) = n^{-1} \sum_{j=1}^n \delta_{X_j(\omega)}(\cdot)$ is a probability measure on $\mathscr{B}(\Gamma)$. Hence $L_n(\omega, \cdot)$ takes values in $\mathscr{M}(\Gamma)$. Let $Q_n^{(2)}$ be the $P_\rho$-distribution of $L_n$ on $\mathscr{M}(\Gamma)$ and define the closed set

$$A_2 = \{\nu \in \mathscr{M}(\Gamma): \max_{i=1,\dots,r} |\nu_i - \rho_i| \geq \varepsilon\} \quad \text{where } 0 < \varepsilon < \min_{i=1,\dots,r} \{\rho_i, 1-\rho_i\}. \tag{1.15}$$

The set $A_2$ is nonempty and $Q_n^{(2)}\{A_2\} = P_\rho\{\max_{i=1,\dots,r}|L_{n,i} - \rho_i| \geq \varepsilon\}$ is positive for all sufficiently large $n$. Since $A_2$ does not contain $\rho$, $Q_n^{(2)}\{A_2\} \to 0$. According to Theorem I.4.3, $Q_n^{(2)}\{A_2\}$ decays exponentially and the decay rate is given in terms of the relative entropy $I_\rho^{(2)}(\nu)$. For this reason $I_\rho^{(2)}(\nu)$ is called the *level-2 entropy function*. For level-1, let $Q_n^{(1)}$ be the $P_\rho$-distribution of $S_n/n$ on $\mathbb{R}$ and define the closed set*

$$A_1 = \{z \in \mathbb{R}: |z - m_\rho| \geq \varepsilon\} \quad \text{where } 0 < \varepsilon < \min\{m_\rho - x_1, x_r - m_\rho\}. \tag{1.16}$$

The set $A_1 \cap [x_1, x_r]$ is nonempty and $Q_n^{(1)}\{A_1\} = P_\rho\{|S_n/n - m_\rho| \geq \varepsilon\}$ is positive for all sufficiently large $n$. Since $A_1$ does not contain $m_\rho$, $Q_n^{(1)}\{A_1\} \to 0$. According to Theorem I.4.3, $Q_n^{(1)}\{A_1\}$ decays exponentially and the decay rate is given in terms of a function $I_\rho^{(1)}$ calculated from $I_\rho^{(2)}(\nu)$ by a variational formula

$$I_\rho^{(1)}(z) = \begin{cases} \min\{I_\rho^{(2)}(\nu): \nu \in \mathscr{M}(\Gamma), \sum_{i=1}^r x_i \nu_i = z\} & \text{for } z \in [x_1, x_r], \\ \infty & \text{for } z \notin [x_1, x_r]. \end{cases} \tag{1.17}$$

$I_\rho^{(1)}$ is called the *level-1 entropy function*. It is well-defined, and it measures the discrepancy between $z$ and $m_\rho$ in the sense that $I_\rho^{(1)}(z) \geq 0$ with equality if and only if $z = m_\rho$. Thus, the point $m_\rho$ is the unique minimum point of $I_\rho^{(1)}(z)$. In addition $I_\rho^{(1)}(z)$ is a continuous convex function of $z \in [x_1, x_r]$. These properties are proved in Sections VII.5 and VIII.3. In Section II.4 we give another formula for $I_\rho^{(1)}$ in terms of a Legendre–Fenchel transform [see (2.14)].

For coin tossing ($\Gamma = \{0, 1\}$, $\rho = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$), formula (1.17) for $I_\rho^{(1)}$ reduces to (1.9). Indeed the only measure $\nu \in \mathscr{M}(\Gamma)$ which satisfies the constraint $\sum_{i=1}^2 x_i \nu_i = z \in [0, 1]$ is $\nu = (1 - z)\delta_0 + z\delta_1$. Hence by (1.17)

$$I_\rho^{(1)}(z) = I_\rho^{(2)}((1-z)\delta_0 + z\delta_1) = (1-z)\log(2(1-z)) + z\log(2z) \quad \text{for } z \in [0, 1].$$

Formula (1.17), which relates the level-1 and level-2 entropy functions, is called a *contraction principle*. It will be seen to follow directly from (1.6), which expresses $S_n/n$ as the mean of the empirical measure $L_n$. Here is the large deviation theorem for levels-1 and 2.

**Theorem I.4.3.**

$$\lim_{n\to\infty} \frac{1}{n} \log Q_n^{(1)}\{A_1\} = \lim_{n\to\infty} \frac{1}{n} \log P_\rho\{|S_n/n - m_\rho| \geq \varepsilon\} = -\min_{z \in A_1} I_\rho^{(1)}(z), \tag{1.18}$$

*The point $m_\rho = \sum_{i=1}^r x_i \rho_i$ is in the open interval $(x_1, x_r)$ since $\rho_i > 0$, $\sum_{i=1}^r \rho_i = 1$.

$$\lim_{n\to\infty}\frac{1}{n}\log Q_n^{(2)}\{A_2\} = \lim_{n\to\infty}\frac{1}{n}\log P_\rho\{\max_{i=1,\dots,r}|L_{n,i}-\rho_i|\geq\varepsilon\} = -\min_{\nu\in A_2} I_\rho^{(2)}(\nu). \tag{1.19}$$

*Since the set $A_1$ is closed and does not contain $m_\rho$, $\min_{z\in A_1} I_\rho^{(1)}(z) > I_\rho^{(1)}(m_\rho) = 0$. Similarly the minimum of $I_\rho^{(2)}$ over $A_2$ is positive. Hence both $Q_n^{(1)}\{A_1\}$ and $Q_n^{(2)}\{A_2\}$ converge to zero exponentially fast.*

It is instructive to interpret Theorem I.4.3 with reference to the discussion in Section I.1. Think of $\{X_j; j\in\mathbb{Z}\}$ as giving the successive outcomes of a gambling game in an infinite number of plays of the game separated by a constant time interval. For level-1 the macrostates are all real numbers $z\in[x_1, x_r]$. These correspond to a macroscopic description of the game in terms of the expected value of the outcome of a single play. The microscopic $n$-sums are $\{S_n(\omega)/n\}$. The $P_\rho$-probability that $S_n/n$ is close to $z$ behaves for large $n$ like $\exp(-nI_\rho^{(1)}(z))$ [see (1.12)]. $I_\rho^{(1)}$ is the entropy function and the mean $m_\rho = \sum_{i=1}^r x_i\rho_i$ is the equilibrium state. For level-2 the macrostates are all probability measures $\nu\in\mathcal{M}(\Gamma)$. Each $\nu$ is a candidate for the distribution of the $r$ outcomes $x_1, \dots, x_r$ in each play of the game. The microscopic $n$-sums are $\{L_n(\omega,\cdot)\}$. The $P_\rho$-probability that $L_n$ is close to $\nu$ behaves for large $n$ like $\exp(-nI_\rho^{(2)}(\nu))$ [see (1.21)]. $I_\rho^{(2)}$ is the entropy function and the measure $\rho$ is the equilibrium state.

*Proof of Theorem* I.4.3. First consider level-2. For fixed $n$ and $\omega$ and $i\in\{1,\dots,r\}$, let $k_i$ be the number of times $x_i$ appears in the sequence $X_1(\omega), \dots, X_n(\omega)$. Then $L_{n,i}(\omega) = k_i/n$, and $L_n(\omega,\cdot)$ is in $A_2$ if and only if $\mathbf{k} = (k_1, \dots, k_r)$ is in the set

$$A_{2,n} = \left\{\mathbf{k} = (k_1,\dots,k_r): k_i\in\{0,1,\dots,n\}, \sum_{i=1}^r k_i = n, \max_{i=1,\dots,r}\left|\frac{k_i}{n}-\rho_i\right|\geq\varepsilon\right\}.$$

For fixed $\mathbf{k}\in A_{2,n}$, define

$$C(n,\mathbf{k}) = \frac{n!}{k_1!k_2!\dots k_r!} \qquad \text{and} \qquad \rho^{\mathbf{k}} = \rho_1^{k_1}\rho_2^{k_2}\dots\rho_r^{k_r}.$$

There are $C(n,\mathbf{k})$ points $\omega = (\omega_1,\omega_2,\dots,\omega_n)$ in the finite configuration space $\Omega_n = \Gamma^n$ for which $L_{n,i}(\omega) = k_i/n$ for each $i$. Let $\pi_n P_\rho$ be the finite product measure on $\Omega_n$ with identical one-dimensional marginals $\rho$. We have

$$Q_n^{(2)}\{A_2\} = \sum_{\mathbf{k}\in A_{2,n}} \pi_n P_\rho\{\omega\in\Omega_n: L_{n,i}(\omega) = k_i/n \text{ for each } i\} = \sum_{\mathbf{k}\in A_{2,n}} C(n,\mathbf{k})\rho^{\mathbf{k}}. \tag{1.20}$$

The next lemma is proved like Lemma I.3.2 [Problem I.8.1(b)].

**Lemma I.4.4.** *Uniformly in* $\mathbf{k} = (k_1,\dots,k_r)$,

$$\frac{1}{n}\log C(n,\mathbf{k}) = -\sum_{i=1}^r \frac{k_i}{n}\log\frac{k_i}{n} + O\left(\frac{\log n}{n}\right) \quad \text{as } n\to\infty.$$

The lemma implies that for each $\mathbf{k}$,

$$\frac{1}{n}\log(C(n,\mathbf{k})\rho^k)) = \sum_{i=1}^{r} \frac{k_i}{n}\left(\log \rho_i - \log\frac{k_i}{n}\right) + O\left(\frac{\log n}{n}\right).$$

Define the measure $\nu_{\mathbf{k}/n} = \sum_{i=1}^{r} (k_i/n)\delta_{x_i} \in \mathcal{M}(\Gamma)$. The sum in the last display is exactly $-I_\rho^{(2)}(\nu_{\mathbf{k}/n})$, where $I_\rho^{(2)}(\nu_{\mathbf{k}/n})$ is the relative entropy of $\nu_{\mathbf{k}/n}$ with respect to $\rho$. Since $L_{n,i}(\omega) = k_i/n$ for each $i$ if and only if $L_n(\omega, \cdot) = \nu_{\mathbf{k}/n}$, we see that

$$\frac{1}{n}\log Q_n^{(2)}\{\nu_{\mathbf{k}/n}\} = -I_\rho^{(2)}(\nu_{\mathbf{k}/n}) + O\left(\frac{\log n}{n}\right). \tag{1.21}$$

In the sum (1.20) for $Q_n^{(2)}\{A_2\}$ there are no more than $(n+1)^r$ terms. As in the proof for coin tossing, we conclude that

$$\frac{1}{n}\log Q_n^{(2)}\{A_2\} = \max_{\mathbf{k}\in A_{2,n}} \{-I_\rho^{(2)}(\nu_{\mathbf{k}/n})\} + O\left(\frac{\log n}{n}\right) + O\left(\frac{\log(n+1)^r}{n}\right).$$

For each $n$ the set $\{\nu \in \mathcal{M}(\Gamma): \nu = \nu_{\mathbf{k}/n} \text{ for some } \mathbf{k} \in A_{2,n}\}$ is a subset of $A_2$. Problem I.8.2 yields (1.19):

$$\lim_{n\to\infty} \frac{1}{n}\log Q_n^{(2)}\{A_2\} = -\lim_{n\to\infty} \min_{\mathbf{k}\in A_{2,n}} I_\rho^{(2)}(\nu_{\mathbf{k},n}) = -\min_{\nu\in A_2} I_\rho^{(2)}(\nu).$$

The level-1 limit is proved by expressing it in terms of a level-2 limit. Since $S_n/n$ equals $\sum_{i=1}^{r} x_i L_{n,i}$, $S_n/n$ is in the set $A_1$ if and only if $L_n$ is in the set of measures $B_2 = \{\nu \in M(\Gamma): \left|\sum_{i=1}^{r} x_i \nu_i - m_\rho\right| \geq \varepsilon\}$. Hence $Q_n^{(1)}\{A_1\}$ equals $Q_n^{(2)}\{B_2\}$. The level-2 argument just given for the set $A_2$ can be easily modified for the set $B_2$, and we find

$$\lim_{n\to\infty} \frac{1}{n}\log Q_n^{(1)}\{A_1\} = \lim_{n\to\infty} \frac{1}{n}\log Q_n^{(2)}\{B_2\} = -\min_{\nu\in B_2} I_\rho^{(2)}(\nu).$$

We evaluate this minimum in two steps:

$$\begin{aligned}\min_{\nu\in B_2} I_\rho^{(2)}(\nu) &= \min_{z\in A_1\cap[x_1,x_r]} \min\left\{I_\rho^{(2)}(\nu): \nu\in\mathcal{M}(\Gamma), \sum_{i=1}^{r} x_i\nu_i = z\right\} \\ &= \min_{z\in A_1\cap[x_1,x_r]} I_\rho^{(1)}(z).\end{aligned}$$

Since $I_\rho^{(1)}(z) = \infty$ for $z \notin [x_1, x_r]$, it follows that

$$\lim_{n\to\infty} \frac{1}{n}\log Q_n\{A_1\} = -\min_{z\in A_1} I_\rho^{(1)}(z). \qquad \square$$

## I.5. Level-3: Empirical Pair Measure

Level-2 focuses on the empirical measure $L_n(\omega, \cdot)$, which is defined in terms of the empirical frequencies $\{L_{n,i}(\omega)\}$. We can generalize level-2 by considering the empirical frequencies of pairs of outcomes. Fix $\omega \in \Omega$ and $n \in \{2, 3, \ldots\}$ and define $Y_\beta^{(n)}(\omega)$ to be the ordered pair $(X_\beta(\omega), X_{\beta+1}(\omega))$ if

$\beta \in \{1, 2, \ldots, n-1\}$ and to be the cyclic ordered pair $(X_n(\omega), X_1(\omega))$ if $\beta = n$. For each subset $\{x_i, x_j\}$ of $\Gamma^2$, let $M_{n,i,j}(\omega)$ be $1/n$ times the number of pairs $\{Y_\beta^{(n)}(\omega)\}$ for which $Y_\beta^{(n)}(\omega) = (x_i, x_j)$; thus

$$M_{n,i,j}(\omega) = \frac{1}{n} \sum_{\beta=1}^{n} \delta_{Y_\beta^{(n)}(\omega)}\{x_i, x_j\}.$$

For each $\omega$, the numbers $\{M_{n,i,j}(\omega)\}$ define a probability measure $M_n(\omega, \cdot)$ on the set of all subsets of $\Gamma^2$. For $A$ a subset of $\Gamma^2$,

$$M_n(\omega, A) = \sum_{\{x_i, x_j\} \in A} M_{n,i,j}(\omega) = \frac{1}{n} \sum_{\beta=1}^{n} \delta_{Y_\beta^{(n)}(\omega)}\{A\}. \tag{1.22}$$

The measure $M_n(\omega, \cdot)$ is called the *empirical pair measure* corresponding to $X_1(\omega), \ldots, X_n(\omega)$. This measure is consistent with the empirical measure $L_n(\omega, \cdot)$ in the sense that both of the one-dimensional marginals of $M_n(\omega, \cdot)$ equal $L_n(\omega, \cdot)$:

$$L_{n,i}(\omega) = \sum_{j=1}^{r} M_{n,i,j}(\omega) = \sum_{k=1}^{r} M_{n,k,i}(\omega) \qquad \text{for each } i = 1, \ldots, r. \tag{1.23}$$

In fact, since $M_n(\omega, \cdot)$ considers the cyclic pair $Y_n^{(n)}(\omega)$, the number of times $x_i$ appears in the sequence $X_1(\omega), \ldots, X_n(\omega)$ equals the number of times $x_i$ appears as a left-hand member of a pair $Y_\beta^{(n)}(\omega)$. This gives the first equality in (1.23), and the second is proved similarly.

With respect to $P_\rho$ the asymptotic behavior of $M_{n,i,j}(\omega)$ is determined by the ergodic theorem [Theorem A.9.3]. Since

$$M_{n,i,j}(\omega) = \frac{1}{n} \sum_{\beta=1}^{n-1} \delta_{X_\beta(\omega)}\{x_i\} \cdot \delta_{X_{\beta+1}(\omega)}\{x_j\} + \frac{1}{n} \delta_{X_n(\omega)}\{x_i\} \cdot \delta_{X_1(\omega)}\{x_j\}, \tag{1.24}$$

$\lim_{n\to\infty} M_{n,i,j}(\omega)$ equals the limit of the sum. Since $P_\rho$ is ergodic,

$$\begin{aligned} \lim_{n\to\infty} M_{n,i,j}(\omega) &= \int_\Omega \delta_{X_1(\omega)}\{x_i\} \delta_{X_2(\omega)}\{x_j\} P_\rho(d\omega) \\ &= P_\rho\{\omega \in \Omega : X_1(\omega) = x_i, X_2(\omega) = x_j\} = \rho_i \rho_j \qquad P_\rho\text{-a.s.} \end{aligned} \tag{1.25}$$

We now formulate a large deviation problem connected with the limit (1.25). Let $\mathcal{B}(\Gamma^2)$ denote the set of all subsets of $\Gamma^2$ and $\mathcal{M}_s(\Gamma^2)$ the set of probability measures on $\mathcal{B}(\Gamma^2)$ with equal one-dimensional marginals. As we have seen, $M_n(\omega, \cdot)$ takes values $\mathcal{M}_s(\Gamma^2)$ for each $\omega$. Any $\tau \in \mathcal{M}_s(\Gamma^2)$ has the form $\tau = \sum_{i,j=1}^r \tau_{ij} \delta_{\{x_i, x_j\}}$, where $\tau_{ij} \geq 0$, $\sum_{i,j=1}^r \tau_{ij} = 1$, and $\sum_{i=1}^r \tau_{ij} = \sum_{i=1}^r \tau_{ki}$ for each $i$. $\mathcal{M}_s(\Gamma^2)$ may be identified with the compact convex subset of $\mathbb{R}^{r^2}$ consisting of all vectors $\tau = \{\tau_{ij}; i, j = 1, \ldots, r\}$ which satisfy $\tau_{ij} \geq 0$, $\sum_{i,j=1}^r \tau_{ij} = 1$, and $\sum_{j=1}^r \tau_{ij} = \sum_{k=1}^r \tau_{ki}$ for each $i$.

Let $Q_{n,2}^{(3)}$ be the $P_\rho$-distribution of $M_n$ on $\mathcal{M}_s(\Gamma^2)$ and define the closed set

$$\begin{aligned} A_3 = \{\tau \in \mathcal{M}_s(\Gamma^2) : \max_{i,j=1,\ldots,r} |\tau_{ij} - \rho_i \rho_j| \geq \varepsilon\}, \\ \text{where } 0 < \varepsilon < \min_{i,j=1,\ldots,r} \{\rho_i \rho_j, 1 - \rho_i \rho_j\}. \end{aligned} \tag{1.26}$$

The set $A_3$ is nonempty and $Q_{n,2}^{(3)}\{A_3\} = P_\rho\{\max_{i,j=1,\dots,r}|M_{n,i,j} - \rho_i\rho_j| \geq \varepsilon\}$ is positive for all sufficiently large $n$. Let $\pi_2 P_\rho$ be the product measure in $\mathscr{M}_s(\Gamma^2)$ with one-dimensional marginals $\rho$ ($\pi_2 P_\rho\{x_i, x_j\} = \rho_i\rho_j$). Since $A_3$ does not contain $\pi_2 P_\rho$, the ergodic limit (1.25) implies that $Q_{n,2}^{(3)}\{A_3\} \to 0$. A level-3 large deviation problem is to determine the decay rate of this probability. In the next section, we will consider other level-3 problems which involve the empirical frequencies of strings $\{x_{i_1}, x_{i_2}, \dots, x_{i_k}\}$ of arbitrary length $k = 3, 4, \dots$.

The probabilities $Q_{n,2}^{(3)}\{A_3\}$ decay exponentially and the decay rate is given in terms of a function $I_{\rho,2}^{(3)}$ which is a natural extension of the level-2 function $I_\rho^{(2)}$. The latter determines the decay rate of level-2 probabilities. For $\tau \in \mathscr{M}_s(\Gamma^2)$, set $(v_\tau)_i = \sum_{j=1}^r \tau_{ij}$ and define

$$I_{\rho,2}^{(3)}(\tau) = \sum \tau_{ij} \log \frac{\tau_{ij}}{(v_\tau)_i \rho_j}, \tag{1.27}$$

where the sum runs over all $i$ and $j$ for which $(v_\tau)_i > 0$. $I_{\rho,2}^{(3)}(\tau)$ is well-defined ($0 \log 0 = 0$) and equals the relative entropy of $\tau$ with respect to $\{(v_\tau)_i \rho_j\}$. We have the following properties [Problems IX.6.1–IX.6.3]. $I_{\rho,2}^{(3)}(\tau)$ is a convex function of $\tau$. $I_{\rho,2}^{(3)}(\tau)$ measures the discrepancy between $\tau \in \mathscr{M}_s(\Gamma^2)$ and $\pi_2 P_\rho$ in the sense that $I_{\rho,2}^{(3)}(\tau) \geq 0$ with equality if and only if $\tau = \pi_2 P_\rho$. Thus the measure $\pi_2 P_\rho$ is the unique minimum point of $I_{\rho,2}^{(3)}$ on $\mathscr{M}_s(\Gamma^2)$.

Here is a level-3 large deviation theorem for the empirical pair measure. We sketch the proof in the case where $\Gamma$ consists of two points ($r = 2$).

**Theorem I.5.1.**

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log Q_{n,2}^{(3)}\{A_3\} &= \lim_{n\to\infty} \frac{1}{n} \log P_\rho\{\max_{i,j=1,\dots,r} |M_{n,i,j} - \rho_i\rho_j| \geq \varepsilon\} \\ &= -\min_{\tau \in A_3} I_{\rho,2}^{(3)}(\tau). \end{aligned} \tag{1.28}$$

*Since the set $A_3$ is closed and does not contain $\pi_2 P_\rho$, $\min_{\tau\in A_3} I_{\rho,2}^{(3)}(\tau) > I_{\rho,2}^{(3)}(\pi_2 P_\rho) = 0$. Hence $Q_{n,2}^{(3)}\{A_3\}$ converges to zero exponentially fast.*

*Proof for $r = 2$.* For fixed $n$ and $\omega$, define $N_{ij} = nM_{n,i,j}(\omega)$. $M_n(\omega, \cdot)$ is in $A_3$ if and only if $N = \{N_{ij}; i,j = 1,2\}$ is in the set

$$A_{3,n} = \Big\{ N = \{N_{ij}; i,j = 1,2\} : N_{ij} \in \{0, 1, \dots, n\}, \sum_{i,j=1}^{2} N_{ij} = n, \\ \sum_{j=1}^{2} N_{ij} = \sum_{k=1}^{2} N_{ki} \quad \text{for each } i, \quad \max_{i,j=1,2} |N_{ij}/n - \rho_i\rho_j| \geq \varepsilon \Big\}.$$

Now fix $N \in A_{3,n}$. Let $k_i = \sum_{j=1}^2 N_{ij}$ and define $\gamma(n, N)$ to be the number of points $\omega$ in the finite configuration space $\Omega_n = \Gamma^n$ for which $M_{n,i,j}(\omega) = N_{ij}/n$ for each $i$ and $j$. Then

$$\begin{aligned} Q_{n,2}^{(3)}\{A_3\} &= \sum_{N\in A_{3,n}} \pi_n P_\rho\{\omega \in \Omega_n : M_{n,i,j}(\omega) = N_{ij}/n \quad \text{for each } i \text{ and } j\} \\ &= \sum_{N\in A_{3,n}} \gamma(n, N) \rho^k, \end{aligned}$$

where $\rho^{\mathbf{k}} = \rho_1^{k_1}\rho_2^{k_2}$. The asymptotics of $\gamma(n, N)$ are given by the next lemma. Define $D_n = \{N: \gamma(n, N) > 0\}$.

**Lemma I.5.2.** *If $N_{12} > 0$, then $N \in D_n$. Uniformly in $N \in D_n$,*

$$\frac{1}{n}\log\gamma(n,N) = \sum_{i=1}^{2}\frac{k_i}{n}\log\frac{k_i}{n} - \sum_{i,j=1}^{2}\frac{N_{ij}}{n}\log\frac{N_{ij}}{n} + O\left(\frac{\log n}{n}\right) \quad \text{as } n \to \infty. \tag{1.29}$$

Define the measure $\tau_{N/n} = \sum_{i,j=1}^{2}(N_{ij}/n)\delta_{\{x_i,x_j\}} \in \mathscr{M}_s(\Gamma^2)$. Since $M_{n,i,j}(\omega) = N_{ij}/n$ for each $i$ and $j$ if and only if $M_n(\omega, \cdot) = \tau_{N/n}$, the lemma implies that for $N \in D_n$

$$\frac{1}{n}\log Q_{n,2}^{(3)}\{\tau_{N/n}\} = \frac{1}{n}\log(\gamma(n,N)\rho^{\mathbf{k}}) = -I_{\rho,2}^{(3)}(\tau_{N/n}) + O\left(\frac{\log n}{n}\right).$$

As in the proof of Theorem I.4.3, we have

$$\lim_{n\to\infty}\frac{1}{n}\log Q_{n,2}^{(3)}\{A_3\} = -\lim_{n\to\infty}\min_{N\in A_{3,n}\cap D_n} I_{\rho,2}^{(3)}(\tau_{N/n}).$$

For each $n$ the set $\{\tau \in \mathscr{M}_s(\Gamma^2): \tau = \tau_{N/n} \text{ for some } N \in A_{3,n} \cap D_n\}$ is a subset of $A_3$. Problem I.8.2 yields (1.28):

$$\lim_{n\to\infty}\frac{1}{n}\log Q_{n,2}^{(3)}\{A_3\} = -\min_{\tau\in A_3} I_{\rho,2}^{(3)}(\tau). \qquad \square$$

Formula (1.31) below is due to J. K. Percus (private communication).

*Proof of Lemma* I.5.2. Suppose that $\Gamma = \{x_1, x_2\}$. For fixed $\omega \in \Omega_n$, let $N_{ij} = nM_{n,i,j}(\omega)$ for $i, j = 1, 2$. The point $\omega$ has the form $(x_{i_1}, x_{i_2}, \ldots, x_{i_n})$, where each $i_j = 1$ or 2. We introduce the product $B(\omega) = B_{i_1i_2}B_{i_2i_3}\ldots B_{i_{n-1}i_n}B_{i_ni_1}$, where $B_{11}$, $B_{12}$, $B_{21}$, $B_{22}$ are four positive real variables. In the product defining $B(\omega)$, each $B_{ij}$ appears exactly $N_{ij}$ times, so that $B(\omega) = B_{11}^{N_{11}}B_{12}^{N_{12}}B_{21}^{N_{21}}B_{22}^{N_{22}}$. The sum of all $B(\omega)$ for $\omega \in \Omega_n$ can be written in terms of the $2 \times 2$ matrix $B = \{B_{ij}; i, j = 1, 2\}$. In fact,*

$$\begin{aligned}\operatorname{Tr} B^n &= \sum_{i_1,\ldots,i_n=1}^{2} B_{i_1i_2}B_{i_2i_3}\ldots B_{i_{n-1}i_n}B_{i_ni_1} \\ &= \sum_{\omega\in\Omega_n} B(\omega) = \sum\gamma(n,N)B_{11}^{N_{11}}B_{12}^{N_{12}}B_{21}^{N_{21}}B_{22}^{N_{22}},\end{aligned} \tag{1.30}$$

where the last sum runs over all $N = \{N_{ij}; i, j = 1, 2\}$ such that $N_{ij} \in \{0, 1, \ldots, n\}$, $\sum_{i,j=1}^{2} N_{ij} = n$, and $\sum_{j=1}^{2} N_{ij} = \sum_{k=1}^{2} N_{ki}$ for each $i$. If $0 < B_{11} + B_{12} + B_{21} + B_{22} < 1$, then the sum $\sum_{n=0}^{\infty} B^n$ converges and equals the inverse matrix $(I - B)^{-1}$. We find

$$\begin{aligned}\sum_{n=0}^{\infty}\operatorname{Tr} B^n &= \operatorname{Tr}\sum_{n=0}^{\infty} B^n = \operatorname{Tr}(I-B)^{-1} \\ &= \frac{2 - B_{11} - B_{22}}{1 - (B_{11} + B_{22} + B_{12}B_{21} - B_{11}B_{22})}.\end{aligned}$$

*Tr denotes trace.

The last term can be expanded in a power series in $B_{11}$, $B_{12}$, $B_{21}$, $B_{22}$. Comparing this series with (1.30), one calculates [Problem I.8.5(a)]

$$\gamma(n,N) = \frac{k_1!k_2!}{N_{11}!N_{12}!N_{21}!N_{22}!}\frac{nN_{12}}{k_1k_2} \qquad \text{for } k_i = N_{i1} + N_{i2} \geq 1 \ (i = 1,2). \tag{1.31}$$

Formula (1.29) follows from Stirling's approximation. □

## I.6. Level-3: Empirical Process

The previous sections have considered three levels of large deviations. Level-1 studies $S_n/n$, level-2 the empirical measure $L_n$, and level-3 the empirical pair measure $M_n$. For each level there is an entropy function $I_\rho^{(1)}$, $I_\rho^{(2)}$, and $I_{\rho,2}^{(3)}$, respectively. In this section we formulate another level-3 problem which includes as special cases all the results in the previous sections. Associated with this problem is another level-3 entropy function $I_\rho^{(3)}$. The three entropy functions encountered already can be obtained from $I_\rho^{(3)}$ by contraction principles.

A subset $\Sigma$ of $\Omega$ is said to be a *cylinder set* if it has the form

$$\Sigma = \{\omega \in \Omega : (\omega_{m+1}, \ldots, \omega_{m+k}) \in F\}, \tag{1.32}$$

where $m$ and $k$ are integers with $k \geq 1$ and $F$ is a subset of $\Gamma^k$. For now, we consider cylinder sets of the form

$$\Sigma = \{\omega \in \Omega : \omega_{m+1} = x_{i_1}, \omega_{m+2} = x_{i_2}, \ldots, \omega_{m+k} = x_{i_k}\},$$

where $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ are elements of $\Gamma$ (not necessarily distinct). We will define a random quantity $R_n(\omega, \cdot)$ which for each $\omega$ is a strictly stationary probability measure on $\Omega$. $R_n(\omega, \Sigma)$ gives the empirical frequency of the string $\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$ in the sequence $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$. If $k = 1$, then $R_n(\omega, \Sigma)$ reduces to the empirical frequency $L_{n,i_1}(\omega)$ while if $k = 2$, then it reduces to the empirical pair frequency $M_{n,i_1,i_2}(\omega)$. In general, with respect to the measure $P_\rho$, $R_n(\omega, \Sigma)$ converges almost surely to the value $P_\rho\{\Sigma\} = \rho_{i_1}\rho_{i_2}\ldots\rho_{i_k}$ as $n \to \infty$. A level-3 large deviation problem is to determine the decay rate of the $P_\rho$-probability that $R_n(\omega, \Sigma)$ differs from $P_\rho\{\Sigma\}$ for finitely many cylinder sets $\Sigma$. This level-3 problem extends the level-2 problem, which studies fluctuations of $L_n(\omega, \cdot)$ away from the measure $\rho$, as well as the level-3 pair problem, which studies fluctuations of $M_n(\omega, \cdot)$ away from the product measure $\pi_2 P_\rho$. We will see that level-3 large deviation probabilities decay exponentially and that the decay rate is determined by a function $I_\rho^{(3)}$ which is a natural extension of $I_\rho^{(2)}$ and $I_{\rho,2}^{(3)}$.

We first define strict stationarity. The set $\Gamma$ is topologized by the discrete topology and the set $\Omega = \Gamma^{\mathbb{Z}}$ by the product topology. The $\sigma$-field generated by the open sets of the product topology is called the *Borel $\sigma$-field* of $\Omega$

and is denoted by $\mathscr{B}(\Omega)$. $\mathscr{B}(\Omega)$ coincides with the $\sigma$-field generated by the cylinder sets $\Sigma$ in (1.32) [Propositions A.3.2 and A.3.5(b)]. Let $T$ be the mapping from $\Omega$ onto $\Omega$ defined by

$$(T\omega)_j = \omega_{j+1} \qquad \text{for } j \in \mathbb{Z}.$$

$T$ is called the *shift mapping* on $\Omega$. For each $B \in \mathscr{B}(\Omega)$, $T^{-1}B$ is also in $\mathscr{B}(\Omega)$ so that $T$ is measurable. A probability measure $P$ on $\mathscr{B}(\Omega)$ is said to be *strictly stationary*, or *translation invariant*, if the shift $T$ preserves $P$: $P\{B\} = P\{T^{-1}B\}$ for all $B \in \mathscr{B}(\Omega)$ or equivalently $P\{\Sigma\} = P\{T^{-1}\Sigma\}$ for all cylinder sets $\Sigma$. We denote by $\mathscr{M}_s(\Omega)$ the set of strictly stationary probability measures on $\mathscr{B}(\Omega)$.

**Example I.6.1.** (a) [See Example A.7.3(a)] *Infinite product measure.* Let $\nu$ be a probability measure on the set $\Gamma = \{x_1, x_2, \ldots, x_r\}$. A cylinder set of the form $\Sigma = \{\omega \in \Omega: \omega_{m+1} \in F_1, \ldots, \omega_{m+k} \in F_k\}$, $F_1, \ldots, F_k$ subsets of $\Gamma$, is called a *product cylinder set*. The class of product cylinder sets generates $\mathscr{B}(\Omega)$. For such a set, we define $P\{\Sigma\} = \prod_{j=1}^{k} \nu\{F_j\}$. Clearly $P\{T^{-1}\Sigma\} = P\{\Sigma\}$. $P$ can be extended uniquely to a strictly stationary probability measure on $\mathscr{B}(\Omega)$. The extension is the infinite product measure with identical one-dimensional marginals $\nu$ and is written $P_\nu$.

(b) [See Example A.7.3(b)] *Markov chain.* Let $\nu$ be a probability measure on the set $\Gamma = \{x_1, x_2, \ldots, x_r\}$ and let $\{\gamma_{ij}\}$ be a non-negative $r \times r$ matrix with row sums 1. Assume that $\sum_{i=1}^{r} \nu_i \gamma_{ij} = \nu_j$ for each $j$. Let $\Sigma$ be the cylinder set $\{\omega \in \Omega: (\omega_{m+1}, \ldots, \omega_{m+k}) \in F\}$, where $F$ is a subset of $\Gamma^k$. We define $P\{\Sigma\} = \nu\{F\}$ for $k = 1$ and

$$P\{\Sigma\} = \sum_{(x_{i_1}, \ldots, x_{i_k}) \in F} \nu_{i_1} \gamma_{i_1 i_2} \cdots \gamma_{i_{k-1} i_k} \qquad \text{for } k \geq 2.$$

Clearly $P\{T^{-1}\Sigma\} = P\{\Sigma\}$. $P$ can be extended uniquely to a strictly stationary probability measure on $\mathscr{B}(\Omega)$. The extension is a Markov chain with transition matrix $\{\gamma_{ij}\}$; $\nu$ is an invariant measure for the chain.

We now state a level-3 large deviation problem. Given a positive integer $n$, repeat the sequence $(X_1(\omega), X_2(\omega), \ldots, X_n(\omega))$ periodically into a doubly infinite sequence, obtaining a point

$$X(n, \omega) = (\ldots X_n(\omega), X_1(\omega), X_2(\omega), \ldots, X_n(\omega), X_1(\omega), X_2(\omega), \ldots, X_n(\omega), X_1(\omega), \ldots)$$

in $\Omega$; $(X(n, \omega))_1 = X_1(\omega)$, $(X(n, \omega))_2 = X_2(\omega)$, etc. For each $\omega \in \Omega$, define a probability measure on $\mathscr{B}(\Omega)$ by

$$R_n(\omega, \cdot) = \frac{1}{n} \sum_{k=0}^{n-1} \delta_{T^k X(n,\omega)}(\cdot), \tag{1.33}$$

where $T^0$ is the identity mapping and $T^k = T(T^{k-1})$ for $k = 2, \ldots, n-1$. For

each Borel subset $B$ of $\Omega$, $R_n(\omega, B)$ is the relative frequency with which $X(n,\omega)$, $TX(n,\omega)$, ..., $T^{n-1}X(n,\omega)$ is in $B$. Since $X(n,\omega)$ is periodic of period $n$, $R(n,\cdot)$ is for each $\omega$ a strictly stationary probability measure. It is called the *empirical process* corresponding to $X_1(\omega)$, ..., $X_n(\omega)$.

Let $k$ be a positive integer less than $n$. In order to interpret formula (1.33), define $Y_\beta^{(n)}(\omega)$ to be the $k$-tuple $(X_\beta(\omega), X_{\beta+1}(\omega), \ldots, X_{\beta+k-1}(\omega))$ if $\beta \in \{1,2,\ldots,n-k+1\}$ and to be the cyclic $k$-tuple $(X_\beta(\omega),\ldots,X_n(\omega),X_1(\omega),\ldots,X_{\beta+k-1-n}(\omega))$ if $\beta \in \{n-k+2,\ldots,n\}$. If $\Sigma$ is the cylinder set $\{\bar\omega \in \Omega: \bar\omega_1 = x_{i_1}, \bar\omega_2 = x_{i_2}, \ldots, \bar\omega_k = x_{i_k}\}$,* then

$$R_n(\omega,\Sigma) = \frac{1}{n}\sum_{\beta=1}^{n} \delta_{Y_\beta^{(n)}(\omega)}\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}.$$

This is the empirical frequency of the string $\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$ in the sequence $X_1(\omega)$, ..., $X_n(\omega)$ (with cyclic counting). The contribution of the cyclic terms $Y_\beta^{(n)}(\omega)$, $\beta \in \{n-k+2,\ldots,n\}$, to $R_n(\omega,\Sigma)$ is $o(n)$ as $n \to \infty$.

$R_n(\omega,\cdot)$ is a natural generalization of the empirical measure $L_n(\omega,\cdot)$ and of the empirical pair measure $M_n(\omega,\cdot)$. If $\Sigma$ is the one-dimensional cylinder set $\{\bar\omega \in \Omega: \bar\omega_1 = x_i\}$, then

$$R_n(\omega,\Sigma) = L_n(\omega,\{x_i\}) = L_{n,i}(\omega). \tag{1.34}$$

Thus $L_n(\omega,\cdot)$ is the one-dimensional marginal of $R_n(\omega,\cdot)$. If $\Sigma$ is the two-dimensional cylinder set $\{\bar\omega \in \Omega: \bar\omega_1 = x_i, \bar\omega_2 = x_j\}$, then

$$R_n(\omega,\Sigma) = M_n(\omega,\{x_i,x_j\}) = M_{n,i,j}(\omega). \tag{1.35}$$

Thus $M_n(\omega,\cdot)$ is the two-dimensional marginal of $R_n(\omega,\cdot)$.

If $\Sigma$ is any cylinder set, then the ergodic theorem implies that

$$\lim_{n\to\infty} R_n(\omega,\Sigma) = P_\rho\{\Sigma\} \qquad P_\rho\text{-a.s.}$$

[Theorems A.9.2(c) and Corollary A.9.8]. Let $\Sigma_1, \ldots, \Sigma_N$ be cylinder sets such that $0 < P_\rho\{\Sigma_k\} < 1$, $k = 1, \ldots, N$, and define the closed set†

$$B_3 = \{P \in \mathcal{M}_s(\Omega): \max_{k=1,\ldots,N} |P\{\Sigma_k\} - P_\rho\{\Sigma_k\}| \geq \varepsilon\}. \tag{1.36}$$

Let $Q_n^{(3)}$ be the $P_\rho$-distribution of $R_n$ on $\mathcal{M}_s(\Omega)$. For all sufficiently small $\varepsilon > 0$ the set $B_3$ is nonempty, and for all sufficiently large $n$

$$Q_n^{(3)}\{B_3\} = P_\rho\{\omega \in \Omega: \max_{k=1,\ldots,N} |R_n(\omega,\Sigma_k) - P_\rho\{\Sigma_k\}| \geq \varepsilon\} > 0.$$

Since $B_3$ does not contain $P_\rho$, the ergodic theorem implies that $Q_n^{(3)}\{B_3\} \to 0$. In fact, the probabilities decay exponentially, and the decay rate is given in terms of a function $I_\rho^{(3)}$, which we now define.

Let $\alpha$ be a positive integer and $\pi_\alpha$ the projection of $\Gamma^{\mathbb{Z}}$ onto $\Gamma^\alpha$ defined by

*We use $\bar\omega$ since $\omega$ labels the empirical process.

†The topology on $\mathcal{M}_s(\Omega)$ is the topology of weak convergence [Sections A.8–A.9].

$\pi_\alpha\omega = (\omega_1, \ldots, \omega_\alpha)$. If $P$ is a measure in $\mathscr{M}_s(\Omega)$, then define a probability measure $\pi_\alpha P$ on $\mathscr{B}(\Gamma^\alpha)$ by requiring

$$\pi_\alpha P\{F\} = P\{\pi_\alpha^{-1}F\} = P\{\omega : (\omega_1, \ldots, \omega_\alpha) \in F\}$$

for subsets $F$ of $\Gamma^\alpha$. The measure $\pi_\alpha P$ is called the *$\alpha$-dimensional marginal* of $P$. We consider the quantity

$$I^{(2)}_{\pi_\alpha P_\rho}(\pi_\alpha P) = \sum_{\omega \in \Gamma^\alpha} \pi_\alpha P\{\omega\} \log \frac{\pi_\alpha P\{\omega\}}{\pi_\alpha P_\rho\{\omega\}},$$

which is the relative entropy of $\pi_\alpha P$ with respect to $\pi_\alpha P_\rho$. In Chapter IX, we prove that the limit

$$I^{(3)}_\rho(P) = \lim_{\alpha \to \infty} \frac{1}{\alpha} I^{(2)}_{\pi_\alpha P_\rho}(\pi_\alpha P) \tag{1.37}$$

exists, that $I^{(3)}_\rho(P)$ is an affine function of $P \in \mathscr{M}_s(\Omega)$, and that $I^{(3)}_\rho(P) \geq 0$ with equality if and only if $P = P_\rho$. Thus $I^{(3)}_\rho(P)$ measures the discrepancy between $P$ and $P_\rho$. It is called the *mean relative entropy* of $P$ with respect to $P_\rho$. Here are some examples of mean relative entropy.

**Example I.6.2.** Let $\rho$ be the uniform measure on $\Gamma = \{x_1, x_2, \ldots, x_r\}$ ($\rho_i = 1/r$ for each $i$). For $P \in \mathscr{M}_s(\Omega)$

$$I^{(2)}_{\pi_\alpha P_\rho}(\pi_\alpha P) = \alpha \log r + \sum_{\omega \in \Gamma^\alpha} \pi_\alpha P\{\omega\} \cdot \log \pi_\alpha P\{\omega\}$$

since $\pi_\alpha P_\rho\{\omega\} = r^{-\alpha}$ for each $\omega \in \Gamma^\alpha$. According to Note 2 of Chapter IX, the limit

$$h(P) = -\lim_{\alpha \to \infty} \frac{1}{\alpha} \sum_{\omega \in \Gamma^\alpha} \pi_\alpha P\{\omega\} \log \pi_\alpha P\{\omega\}$$

exists; $h(P)$ is called the *mean entropy*[4] of $P$. It follows that the limit $I^{(3)}_\rho(P)$ in (1.37) exists and $I^{(3)}_\rho(P) = \log r - h(P)$. By properties of $I^{(3)}_\rho(P)$ mentioned above, $h(P) \leq \log r$ and $h(P) = \log r$ if and only if $P = P_\rho$. Mean entropy $h(P)$ generalizes Shannon entropy, which was discussed in Example I.4.2. Accordingly, $h(P)$ can be interpreted as a measure of the randomness in $P$ per unit time.

(a) For the infinite product measure $P_\nu$ [Example I.6.1(a)], $I^{(3)}_\rho(P_\nu) = \log r - h(P_\nu)$, where $h(P_\nu) = -\sum_{i=r}^{r} \nu_i \log \nu_i$. The latter is the Shannon entropy of $\nu$, $H(\nu)$.

(b) Let $P$ be a Markov chain with transition matrix $\{\gamma_{ij}\}$ and invariant measure $\nu$ [Example I.6.1 (b)]. Then $I^{(3)}_\rho(P) = \log r - h(P)$, where $h(P) = -\sum_{i,j=1}^{r} \nu_i \gamma_{ij} \log \gamma_{ij}$.

For any measure $\rho$ in $\mathscr{M}(\Gamma)$ with each $\rho_i > 0$ and for any $P \in \mathscr{M}_s(\Omega)$, one can express $I^{(3)}_\rho(P)$ as an expectation involving relative entropy [see (2.20)]. Here is the level-3 large deviation theorem. It will be proved in Chapter IX.

**Theorem I.6.3.** *For all sufficiently small* $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{n} \log Q_n^{(3)}\{B_3\} = \lim_{n\to\infty} \frac{1}{n} \log P_\rho\{\omega \in \Omega: \max_{k=1,\dots,N} |R_n(\omega, \Sigma_k) - P_\rho\{\Sigma_k\}| \geq \varepsilon\}$$

$$= -\min_{P\in B_3} I_\rho^{(3)}(P). \tag{1.38}$$

*Since the set* $B_3$ *is closed and does not contain* $P_\rho$, $\min_{P\in B_3} I_\rho^{(3)}(P) > I_\rho^{(3)}(P_\rho) = 0$. *Hence* $Q_n^{(3)}\{B_3\}$ *converges to zero exponentially fast.*

The theorem can be interpreted as follows. For level-3, the macrostates are all strictly stationary probability measures $P \in \mathscr{M}_s(\Omega)$. Each $P$ is a candidate for describing the probabilistic structure of a gambling game on $\Omega$ [see page 16]. The microscopic $n$-sums are the empirical processes $\{R_n(\omega, \cdot)\}$. The $P_\rho$-probability that $R_n(\omega, \cdot)$ is close to $P \in \mathscr{M}_s(\Omega)$ behaves for large $n$ like $\exp(-nI_\rho^{(3)}(P))$. $I_\rho^{(3)}$ is the entropy function and the measure $P_\rho$ is the equilibrium state.

We end this chapter by discussing the relationship between Theorem I.6.3 and the previous large deviation theorems. Theorem I.5.1 treated the empirical pair measure $M_n(\omega, \cdot)$. If $P$ is a measure in $\mathscr{M}_s(\Omega)$, then the two-dimensional marginal $\pi_2 P$ is a probability measure on $\Gamma^2$, and since $P$ is strictly stationary, $\pi_2 P$ has equal one-dimensional marginals. Hence $\pi_2 P$ belongs to $\mathscr{M}_s(\Gamma^2)$. In the set $B_3$ in (1.36), let $\{\Sigma_k\}$ run through all two-dimensional cylinder sets $\{\bar{\omega} \in \Omega: \bar{\omega}_1 = x_i, \bar{\omega}_2 = x_j\}$, $i, j = 1, \dots, r$. Then

$$\pi_2 B_3 = \{\tau \in \mathscr{M}_s(\Gamma^2): \tau = \pi_2 P \text{ for some } P \in B_3\}$$

$$= \{\tau \in \mathscr{M}_s(\Gamma^2): \max_{i,j=1,\cdots,r} |\tau_{ij} - \rho_i \rho_j| \geq \varepsilon\}.$$

The latter is the set $A_3$ in (1.26). Since $\pi_2 R_n(\omega, \cdot) = M_n(\omega, \cdot)$ for each $\omega$, it follows that

$$P_\rho\{\omega \in \Omega: M_n(\omega, \cdot) \in A_3\} = P_\rho\{\omega \in \Omega: R_n(\omega, \cdot) \in B_3\}.$$

By Theorem I.5.1 the decay rate of the first probability is $-\min_{\tau\in A_3} I_{\rho,2}^{(3)}(\tau)$. This must equal the decay rate of $P_\rho\{\omega \in \Omega: R_n(\omega, \cdot) \in B_3\}$, which by Theorem I.6.3 equals $-\min\{I_\rho^{(3)}(P): P \in B_3\}$. The latter can be rewritten as

$$-\min_{\tau\in B_3} \min\{I_\rho^{(3)}(P): P \in \mathscr{M}_s(\Omega), \pi_2 P = \tau\}.$$

Thus, one expects that for $\tau \in \mathscr{M}_s(\Gamma^2)$

$$I_{\rho,2}^{(3)}(\tau) = \sum \tau_{ij} \log \frac{\tau_{ij}}{(\nu_\tau)_i \rho_j} = \min\{I_\rho^{(3)}(P): P \in \mathscr{M}_s(\Omega), \pi_2 P = \tau\}, \tag{1.39}$$

where $(\nu_\tau)_i = \sum_{j=1}^r \tau_{ij}$. We shall prove (1.39) in Chapter IX. It is the contraction principle relating $I_{\rho,2}^{(3)}$ and $I_\rho^{(3)}$.

The connection with Theorem I.4.3 is similar. If $P$ is a measure in $\mathscr{M}_s(\Omega)$, then the one-dimensional marginal $\pi_1 P$ is an element of $\mathscr{M}(\Gamma)$. In the set $B_3$

in (1.36), let $\{\Sigma_k\}$ run through all one-dimensional cylinder sets $\{\bar{\omega}\in\Omega: \bar{\omega}_1 = x_i\}, i = 1, \ldots, r$. Then

$$\pi_1 B_3 = \{\nu\in\mathscr{M}(\Gamma): \nu = \pi_1 P \text{ for some } P\in B_3\}$$
$$= \{\nu\in\mathscr{M}(\Gamma): \max_{i=1,\ldots,r} |\nu_i - \rho_i| \geq \varepsilon\}.$$

The latter is the set $A_2$ in (1.15). Since $\pi_1 R_n(\omega, \cdot) = L_n(\omega, \cdot)$ for each $\omega$, it follows that

$$P_\rho\{\omega\in\Omega: L_n(\omega, \cdot)\in A_2\} = P_\rho\{\omega\in\Omega: R_n(\omega, \cdot)\in B_3\}.$$

Again, comparing decay rates, we have

$$\min_{\nu\in A_2} I_\rho^{(2)}(\nu) = \min_{P\in B_3} I_\rho^{(3)}(P) = \min_{\nu\in A_2} \min\{I_\rho^{(3)}(P): P\in\mathscr{M}_s(\Omega), \pi_1 P = \nu\}.$$

In Chapter IX, we prove for $\nu\in\mathscr{M}(\Gamma)$ the following contraction principle which is consistent with the last display:

$$(1.40)\qquad I_\rho^{(2)}(\nu) = \sum_{i=1}^{r} \nu_i \log\frac{\nu_i}{\rho_i} = \min\{I_\rho^{(3)}(P): P\in\mathscr{M}_s(\Omega), \pi_1 P = \nu\}.$$

Finally, for level-1, recall the contraction principle (1.17) relating $I_\rho^{(1)}$ and $I_\rho^{(2)}$: for $z\in[x_1, x_r]$

$$I_\rho^{(1)}(z) = \min\left\{I_\rho^{(2)}(\nu): \nu\in\mathscr{M}(\Gamma), \sum_{i=1}^{r} x_i\nu_i = z\right\}.$$

Comparing this with (1.40), we conclude that for $z\in[x_1, x_r]$,

$$(1.41)\qquad I_\rho^{(1)}(z) = \min\left\{I_\rho^{(3)}(P): P\in\mathscr{M}_s(\Omega), \sum_{i=1}^{r} x_i(\pi_1 P)_i = z\right\}.$$

This completes our discussion of large deviations for i.i.d. random variables with a finite outcome space. An interesting feature of this chapter was the use of combinatorics to calculate explicit formulas for the entropy functions $I_\rho^{(1)}$, $I_\rho^{(2)}$, and $I_{\rho,2}^{(3)}$. In the next chapter, the three levels of large deviations will be generalized to random vectors taking values in $\mathbb{R}^d$. The theory will be applied to statistical mechanics in later parts of the book.

## I.7. Notes

1 (page 3). Wehrl (1978) discusses the historical and physical backgrounds of entropy together with modern developments. He lists many references.

2 (page 13). Relative entropy $I_\rho^{(2)}(\nu)$, introduced by Kullback and Leibler in 1951, is also known as the Kullback–Leibler information number. It plays an important role in statistics, especially in large sample theories of estimation and testing [Kullback (1959), Bahadur (1967, 1971)]. $I_\rho^{(2)}(\nu)$ measures the statistical distance between $\nu$ and $\rho$. The smaller this distance, the harder

it is to discriminate between $\nu$ and $\rho$. Applications of large deviations to statistics are discussed by Bahadur (1971), Chernoff (1972), and individual articles in "Grandes deviations et applications statistiques," Astérisque *68*, Sociéte Mathématique de France, Paris, 1979.

3 (page 14). Shannon entropy $H(\nu)$ was first defined by Shannon (1948) and independently by Wiener (1948). The form of $H(\nu)$ can be derived from a set of axioms which a reasonable measure of randomness should satisfy; see, e.g., Khinchin (1957, pages 9–13).

4 (page 24). In information theory, the mean entropy $h(P)$, $P \in \mathcal{M}_s(\Omega)$, measures the amount of information per symbol in a message which is generated according to $P$ [Khinchin (1957), MeEliece (1977)]. Large deviation bounds, known as Chernoff bounds [see Problem VII.8.9], are widely used [Wozencraft and Jacobs (1965), Gallager (1968)]. In ergodic theory, $h(P)$ is the Kolmogorov–Sinai invariant of the dynamical system $(\Omega, \mathcal{F}, P, T)$ [Martin and England (1981)].

## I.8. Problems

**I.8.1.** (a) Given $n \in \{2, 3, \ldots, \}$, prove that $n \log n - n + 1 < \log(n!) < (n+1)\log(n+1) - (n+1) + 1$ by considering the area under the graph of $\log x$, $x \geq 1$. Deduce the weak form of Stirling's approximation: $\log(n!) = n \log n - n + O(\log n)$, $n \geq 2$.

(b) Prove Lemma I.4.4.

**I.8.2.** This problem shows how to complete the proofs of Theorems I.3.1, I.4.3, and I.5.1. Let $A$ be a compact subset of $\mathbb{R}^d$ and $f$ a real-valued function on $A$ which is continuous relative to $A$. Let $\{A_n; n = 1, 2, \ldots\}$ be closed subsets of $A$ such that for any $a \in A$ there exists a sequence $a_n \in A_n$ with $a_n \to a$ as $n \to \infty$. Prove that $\lim_{n\to\infty} \min_{x \in A_n} f(x) = \min_{x \in A} f(x)$.

**I.8.3.** Let $p = \{p_{ij}; i, j = 1, \ldots, r\}$ be a set of positive numbers such that $\sum_{i,j=1}^{r} p_{ij} = 1$. Define $\nu_i = \sum_{j=1}^{r} p_{ij}$ and $\mu_j = \sum_{i=1}^{r} p_{ij}$. Let $\rho = \{\rho_i; i = 1, \ldots, r\}$ be a sequence of positive numbers such that $\sum_{i=1}^{r} \rho_i = 1$. Prove that

$$I^{(2)}_{\{\rho_i \rho_j\}}(\{p_{ij}\}) = \sum_{i,j=1}^{r} p_{ij} \log \frac{p_{ij}}{\rho_i \rho_j} \geq I^{(2)}_{\rho}(\nu) + I^{(2)}_{\rho}(\mu)$$

$$= \sum_{i=1}^{r} \nu_i \log \frac{\nu_i}{\rho_i} + \sum_{j=1}^{r} \mu_j \log \frac{\mu_j}{\rho_j}$$

with equality if and only if $p_{ij} = \nu_i \mu_j$ for each $i$ and $j$.

If each $\rho_i = 1/r$, then we conclude that the randomness in $p$ (as measured by Shannon entropy) is no greater than the sum of the randomness in $\nu$ and the randomness in $\mu$ and that equality holds if and only if $p$ is product measure.

**I.8.4.** (a) Prove Jensen's inequality: if $w_1, w_2, \ldots, w_r$ and $y_1, y_2, \ldots, y_r$ are nonnegative numbers such that $\sum_{j=1}^r w_j = 1$, then

$$\left(\sum_{j=1}^r w_j y_j\right) \log\left(\sum_{j=1}^r w_j y_j\right) \leq \sum_{j=1}^r w_j y_j \log y_j.$$

[*Hint:* Let $h(x) = x \log x$. If $a = \sum_{j=1}^r w_j y_j > 0$, then $h(y_j) \geq h(a) + h'(a) \cdot (y_j - a)$.]

(b) [Rényi (1970b, page 556)]. Let $v = \{v_j; j = 1, \ldots, r\}$ be a sequence of non-negative numbers such that $\sum_{j=1}^r v_j = 1$. Let $\gamma = \{\gamma_{jk}\}$ be an $r \times r$ doubly stochastic matrix ($\gamma_{jk} \geq 0, \sum_{j=1}^r \gamma_{jk} = 1$ for each $k$, $\sum_{k=1}^r \gamma_{jk} = 1$ for each $j$) and define $\mu_k = \sum_{j=1}^r v_j \gamma_{jk}$. Set $\rho_i = 1/r$ for $i = 1, \ldots, r$. Prove that

$$\sum_{k=1}^r \mu_k \log \mu_k \leq \sum_{j=1}^r v_j \log v_j \quad \text{and} \quad I_\rho^{(2)}(\mu) \leq I_\rho^{(2)}(v).$$

See Voight (1981) for generalizations.

**I.8.5** [J. K. Percus (private communication)]. (a) Derive formula (1.31). [*Hint:* Expand

$$\frac{2 - B_{11} - B_{22}}{(1-c)(1 - B_{12}B_{21}/(1-c))} = (1 - B_{11} + 1 - B_{22}) \sum_{n \geq 0} (B_{12}B_{21})^n (1-c)^{-n-1},$$

where $1 - c = 1 - (B_{11} + B_{22} - B_{11}B_{22}) = (1 - B_{11})(1 - B_{22})$.]

(b) For $r \geq 2$, let $k_i = \sum_{j=1}^r N_{ij}$. Show that

$$\gamma(n, N) = \frac{\prod_{i=1}^r k_i!}{\prod_{i,j=1}^r N_{ij}!} \cdot \sum_{\alpha=1}^r \left[\delta_{ij} - \frac{N_{ij}}{k_j}\right]_{\alpha\alpha} \qquad \text{for all } N_{ij} \geq 1,$$

where $[\delta_{ij} - N_{ij}/k_j]_{\alpha\alpha}$ is the $\alpha\alpha$-cofactor of the $r \times r$ matrix $\{\delta_{ij} - N_{ij}/k_j\}$. [*Hint:* If $B = \{B_{ij}\}$ is an $r \times r$ positive matrix with $0 < \sum_{i,j=1}^r B_{ij} < 1$, then $\mathrm{Tr}(I - B)^{-1} = \sum_{\alpha=1}^r [I - B]_{\alpha\alpha}/\det(I - B)$, where $[I - B]_{\alpha\alpha}$ is the $\alpha\alpha$-cofactor of $I - B$. Use the formula (which holds in the sense of distribution theory)

$$\frac{1}{\det(I - B)} = \int_{\mathbb{R}^r} \int_{\mathbb{R}^r} \exp[i(\langle x, y\rangle - \langle x, By\rangle)] dx dy \cdot \frac{1}{(2\pi)^r}.]$$

**I.8.6.** Let $B = \{B_{ij}\}$ be a positive $2 \times 2$ matrix and $\lambda(B)$ the largest eigenvalue of $B$ in absolute value. By Lemma IX.4.1, $\lambda(B)$ is positive and $\log \lambda(B) = \lim_{n \to \infty} n^{-1} \log \mathrm{Tr}\, B^n$. As in the proof of Lemma I.5.2,

$$\mathrm{Tr}\, B^n = \Sigma \gamma(n, N) \prod_{i,j=1}^2 B_{ij}^{N_{ij}}. \tag{1.42}$$

where the sum runs over all $N = \{N_{ij}; i,j = 1,2\}$ such that $N_{ij} \in \{0, 1, \ldots, n\}$, $\sum_{i,j=1}^{2} N_{ij} = n$, and $\sum_{j=1}^{2} N_{ij} = \sum_{k=1}^{2} N_{ki}$ for each $i$. Using Lemma I.5.2, prove that

$$\log \lambda(B) = \max \left\{ \sum \tau_{ij} \log \frac{B_{ij}(v_\tau)_i}{\tau_{ij}} : \tau \in \mathscr{M}_{s,2} \right\}, \tag{1.43}$$

where $(v_\tau)_i = \sum_{j=1}^{2} \tau_{ij}$ and $\mathscr{M}_{s,2}$ is the set of $\tau = \{\tau_{ij}; i,j = 1,2\}$ satisfying $\tau_{ij} \geq 0$, $\sum_{i,j=1}^{2} \tau_{ij} = 1$, and $\sum_{j=1}^{2} \tau_{ij} = \sum_{k=1}^{2} \tau_{ki}$ for each $i$. The sum in (1.43) runs over all $i$ and $j$ for which $(v_\tau)_i > 0$. Theorem IX.4.4 and Problem IX.6.4 are generalizations.

**I.8.7.** Verify the calculations in Examples I.6.2(a), (b).