*Chapter 1*

# Introductory discussion

Toss a fair coin $n$ times. When $n$ is small there is nothing to say beyond enumerating all the outcomes and their probabilities. With a large number of tosses patterns and order emerge from the randomness: heads appear about 50% of the time and the histogram approaches a bell curve. As the number of tosses increases, these patterns become more and more pronounced. But from time to time a random fluctuation might break the pattern: perhaps 10,000 tosses of a fair coin give 6,000 heads. In fact, we know that there is a chance of $(1/2)^{10,000}$ that all tosses yield heads. The point is that to understand the system well one cannot be satisfied with understanding only the most likely outcomes. One also needs to understand rare events. But why care about an event that has a chance of $(1/2)^{10,000}$?

Here is a simplified example to illustrate the importance of probabilities of rare events. Imagine that an insurance company collects premiums at a steady rate of $c$ per month. Let $X_k$ be the random amount that the insurance company pays out in month $k$ to cover customer claims. Let $S_n = X_1 + \cdots + X_n$ be the total payout in $n$ months. Naturally the premiums must cover the average outlays, so $c > E[X_k]$. The company stays solvent as long as $S_n \leq cn$. Quantifying the chances of the rare event $S_n > cn$ is then of obvious interest.

This is an introductory book on the methods of computing asymptotics of probabilities of rare events: the theory of large deviations. Let us start with a basic computation.

**Example 1.1.** Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables with success probability $p$ (each $X_k = 1$ with probability $p$ and 0 with probability $1 - p$). Denote the

partial sum by $S_n = X_1 + \cdots + X_n$. The strong law of large numbers says that, as $n \to \infty$, the sample mean $S_n/n$ converges to $p$ almost surely. But at any given $n$ there is a chance $p^n$ for all heads ($S_n = n$) and also a chance $(1-p)^n$ for all tails ($S_n = 0$). In fact, for any $s \in (0,1)$ there is a positive chance of a fraction of heads close to $s$. Let us compute the asymptotics of this probability.

Denote the integer part of $x \in \mathbb{R}$ by $\lfloor x \rfloor$; that is, $\lfloor x \rfloor$ is the largest integer less than or equal to $x$. From binomial probabilities

$$\begin{aligned} P\{S_n = \lfloor ns \rfloor\} &= \frac{n!}{\lfloor ns \rfloor! (n - \lfloor ns \rfloor)!}\, p^{\lfloor ns \rfloor} (1-p)^{n - \lfloor ns \rfloor} \\ &\sim \frac{n^n p^{\lfloor ns \rfloor} (1-p)^{n-\lfloor ns \rfloor}}{\lfloor ns \rfloor^{\lfloor ns \rfloor} (n - \lfloor ns \rfloor)^{n - \lfloor ns \rfloor}} \sqrt{\frac{n}{2\pi \lfloor ns \rfloor (n - \lfloor ns \rfloor)}}\,. \end{aligned}$$

We used Stirling's formula (Exercise 3.5)

$$n! \sim e^{-n} n^n \sqrt{2\pi n}\,. \tag{1.1}$$

The notation $a_n \sim b_n$ means that $a_n/b_n \to 1$. Abbreviate

$$\beta_n = \sqrt{\frac{n}{2\pi \lfloor ns \rfloor (n - \lfloor ns \rfloor)}}$$

and to get rid of integer parts, let also

$$\gamma_n = \frac{(ns)^{ns} (n - ns)^{n-ns}}{\lfloor ns \rfloor^{\lfloor ns \rfloor} (n - \lfloor ns \rfloor)^{n - \lfloor ns \rfloor}} \cdot \frac{p^{\lfloor ns \rfloor} (1-p)^{n - \lfloor ns \rfloor}}{p^{ns} (1-p)^{n-ns}}\,.$$

Then

$$P\{S_n = \lfloor ns \rfloor\} \;\sim\; \beta_n \gamma_n \exp\Big( ns \log \frac{p}{s} + n(1-s) \log \frac{1-p}{1-s} \Big).$$

***Exercise 1.2.** Show that there exists a constant $C$ such that

$$\frac{1}{C\sqrt{n}} \le \beta_n \le \frac{C}{\sqrt{n}} \quad \text{and} \quad \frac{1}{Cn} \le \gamma_n \le Cn$$

for large enough $n$. By being a little more careful you can improve the second statement to $C^{-1} \le \gamma_n \le C$.

The asymptotics above give the limit

$$\begin{aligned} &\lim_{n \to \infty} \frac{1}{n} \log P\{S_n = \lfloor ns \rfloor\} = -I_p(s) \\ &\text{with} \quad I_p(s) = s \log \frac{s}{p} + (1-s) \log \frac{1-s}{1-p}. \end{aligned} \tag{1.2}$$

Note the minus sign introduced in front of $I_p(s)$. This is a convention of large deviation theory.

It is instructive to look at the graph of $I_p$ (Figure 1.1). $I_p$ extends continuously to $[0,1]$ with values $I_p(0) = \log \frac{1}{1-p}$ and $I_p(1) = \log \frac{1}{p}$ that

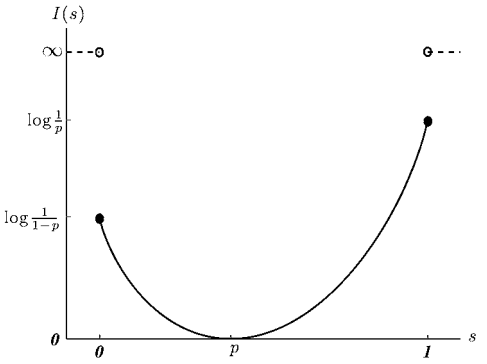


**Figure 1.1.** The rate function for coin tosses.

match the exponential decay of the probabilities of the events $\{S_n = 0\}$ and $\{S_n = n\}$. The unique zero of $I_p$ is at the law of large numbers limit $p$ which we would regard as the "typical" behavior of $S_n/n$. Increasing values of $I_p$ correspond to less likely outcomes. For $s \notin [0,1]$ it is natural to set $I_p(s) = \infty$.

The function $I_p$ in (1.2) is a *large deviation rate function*. We shall understand later that $I_p(s)$ is also the *relative entropy* of the coin with success probability $s$ relative to the one with success probability $p$. The choice of terminology is not a coincidence. This quantity is related to both information-theoretic and thermodynamic entropy.

For this reason we take a brief detour to discuss these well-known notions of entropy and to point out the link with the large deviation rate function $I_p$. The relative entropy that appears in large deviation theory will take center stage in Chapters 5–6, and again in Chapter 8 when we discuss statistical mechanics of lattice systems.

Limit (1.2) is our first large deviation result. One of the very last ones in the book is limit (16.12), which is the analogue of (1.2) for a random walk in a dynamical random environment, that is, in a setting where the success probability of the coin also fluctuates randomly.

## 1.1. Information-theoretic entropy

A coin that always comes up heads is not random at all, and the same of course for a coin that always comes up tails. On the other hand, we should probably regard a fair coin as the "most random" coin because we cannot predict whether we see more heads or tails in a sequence of tosses with better than even odds. We discuss here briefly the quantification of the degree of randomness of a sequence of coin flips. We take the point of view that the

degree of randomness of the coin is reflected in the average number of bits needed to encode a sequence of tosses. This section is inspired by Chapter 2 of Ash's book [**4**].

Let $\Omega = \{0,1\}^n$ be the space of *words* $\omega \in \Omega$ of length $n$. A *message* is a concatenation of words. The message made of words $\omega_1, \omega_2, \dots, \omega_m$ is written $\omega_1\omega_2\cdots\omega_m$. A *code* is a map $C : \Omega \to \bigcup_{\ell\ge1}\{0,1\}^\ell$ that assigns to each word $\omega \in \Omega$ a *code word* $C(\omega)$ which is a finite sequence of 0's and 1's. $|C(\omega)|$ denotes the length of code word $C(\omega)$. A concatenation of code words is a *code message*. Thus, a message is encoded by concatenating the code words of its individual words to make a code message: $C(\omega_1\cdots\omega_m) = C(\omega_1)\cdots C(\omega_m)$. A code should be *uniquely decipherable*. That is, for every finite sequence $c_1\cdots c_\ell$ of 0's and 1's there exists at most one message $\omega_1\cdots\omega_m$ such that $C(\omega_1)\cdots C(\omega_m) = c_1\cdots c_\ell$.

Now sample words at random under a probability distribution $P$ on the space $\Omega$. In this discussion we employ the base 2 logarithm $\log_2 x = \log x/\log 2$.

**Noiseless coding theorem.** *If $C$ is a uniquely decipherable code, then its average length satisfies*

$$\sum_{\omega\in\Omega} P(\omega)\,|C(\omega)| \ge -\sum_{\omega\in\Omega} P(\omega)\log_2 P(\omega) \tag{1.3}$$

*with equality if and only if $P(\omega) = 2^{-|C(\omega)|}$.*

In information theory the quantity on the right of (1.3) is called the *Shannon entropy* of the probability distribution $P$. For a simple proof of the theorem see [**4**, Theorem 2.5.1, page 37].

Consider the case where the $n$ *characters* of the word $\omega$ are chosen independently, and let $s \in [0,1]$ be the probability that a character is a 1. Then $P(\omega) = s^{N(\omega)}(1-s)^{n-N(\omega)}$, where $N(\omega)$ is the number of ones in $\omega$. (As usual, $0^0 = 1$.) By the noiseless coding theorem, the average length of a decipherable code $C$ satisfies

$$\sum_{\omega\in\Omega} |C(\omega)|\, s^{N(\omega)}(1-s)^{n-N(\omega)} \ge -\sum_{\omega\in\Omega} s^{N(\omega)}(1-s)^{n-N(\omega)}\log_2 s^{N(\omega)}(1-s)^{n-N(\omega)}.$$

Since $\sum_\omega s^{N(\omega)}(1-s)^{n-N(\omega)} = 1$ and $\sum_\omega N(\omega)s^{N(\omega)}(1-s)^{n-N(\omega)} = ns$, the right-hand side equals $nh(s)$ where

$$h(s) = -s\log_2 s - (1-s)\log_2(1-s) = 1 - \frac{I_{1/2}(s)}{\log 2},$$

and we see the large deviations rate function from (1.2) appear. Thus we have the lower bound

$$\sum_{\omega\in\Omega} |C(\omega)|\, s^{N(\omega)}(1-s)^{n-N(\omega)} \geq nh(s). \tag{1.4}$$

In other words, any uniquely decipherable code for independent and identically distributed characters with probability $s$ for a 1 must use, on average, at least $h(s)$ bits per character. In this case the Shannon entropy and the rate function $I_{1/2}$ are related by

$$-\sum_{\omega\in\Omega} P(\omega)\log_2 P(\omega) = 1 - \frac{I_{1/2}(s)}{\log 2}.$$

Here is a simplistic way to see the lower bound $nh(s)$ on the number of bits needed that makes an indirect appeal to large deviations in the sense that deviant words are ignored. With probability $s$ for symbol 1, the typical word of length $n$ has about $ns$ ones. Suppose we use code words of length $L$ to code these typical words. Then

$$2^L \geq \binom{n}{\lfloor ns \rfloor}$$

and the lower bound $L \geq nh(s) + \mathcal{O}(\log n)$ follows from Stirling's formula.

The values $h(0) = h(1) = 0$ make asymptotic sense. For example, if $s = 0$, then a word of any length $n$ is all zeroes and can be encoded by a single bit, which in the $n \to \infty$ limit gives 0 bits per character. This is the case of complete order. At the other extreme of complete disorder is the case $s = 1/2$ of fair coin tosses where all $n$ bits are needed because all words of a given length are equally likely. For $s \neq 1/2$ a 1 is either more or less likely than a 0 and by exploiting this bias one can encode with less than 1 bit per character on average.

David A. Huffman [**48**], while a Ph.D. student at MIT, developed an optimal decipherable code, that is, a code $C$ whose average length cannot be improved upon. As $n \to \infty$, the average length of the code generated by this algorithm is exactly $h(s)$ per character and so the lower bound (1.4) is achieved asymptotically. We illustrate the algorithm through an example. For a proof of its optimality and asymptotic average length see page 42 of [**4**].

**Example 1.3** (Huffman's algorithm)**.** Consider the case $n = 3$ and $s = 1/4$. There are 8 words. Word 111 comes with probability $1/4^3$, words 110, 101, and 011 come each with probability $3/4^3$, words 100, 010, and 001 come with probability $3^2/4^3$ each, and word 000 comes with probability $(3/4)^3$. These 8 words are the terminal leaves of a binary tree that we build.

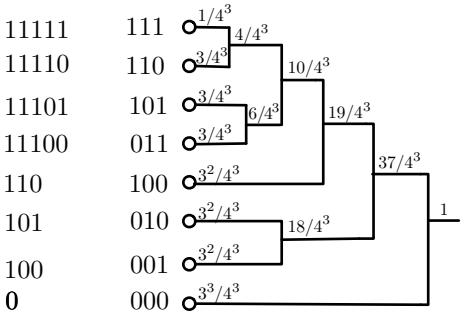


**Figure 1.2.** The tree for Huffman's algorithm in the case $n = 3$ and $s = 1/4$. The leftmost column shows the resulting codes.

First, find the two leaves with the smallest probabilities. Ties can be resolved arbitrarily. Give these two leaves $a$ and $b$ a common ancestor labeled with a probability that is the sum of the probabilities of $a$ and $b$. In our example, leaves 111 and 110 are given a common parent labeled with probability $4/4^3$.

Now leaves $a$ and $b$ are done with, and their parent is regarded as a new leaf. Repeat the step. Continue until there is one leaf left. In our example, the second step gives a common ancestor to leaves 101 and 011. This new node is lab, and so on. Figure 1.2 presents the final tree.

To produce the code of a word, start at the root and follow the tree to the leaf of that word. At each fork encode a down step with a 0 and an up step with a 1 (in our figure). For instance, word 101 is reached from the root by three successive up steps followed by a single down step and then another up step. Thus word 101 is encoded as 11101.

The average length of the code is

$$\frac{5\times 1+5\times 3+5\times 3+5\times 3+3\times 3^2+3\times 3^2+3\times 3^2+1\times 3^3}{4^3}=\frac{158}{64}.$$

This is $158/192 \approx 0.8229$ bits per character. As the number of characters $n$ grows, the average length of the encoding per character will converge to the information-theoretic entropy $h(1/4) \approx 0.811$.

## 1.2. Thermodynamic entropy

The next discussion of thermodynamics is inspired by Schrödinger's lectures [**70**]. After some preliminary computations we use the first and second laws of thermodynamics to derive an expression for entropy. In the simplest case of a system with two energy levels this expression can be related to the rate function (1.2). The reader should be aware that this section is not mathematically rigorous.

Let $A$ denote a physical system whose possible energy levels are $\{\varepsilon_\ell : \ell \in \mathbb{N}\}$. Then consider a larger system $A_n$ made up of $n$ identical physically independent copies of $A$. By physically independent we mean that these components of $A_n$ do not communicate with each other. Each component can be at any energy level $\varepsilon_\ell$. Immerse the whole system in a large heat bath at fixed absolute temperature $T$ which gives the system total energy $\mathcal{E} = nU$.

Let $a_\ell$ be the number of components in state $\varepsilon_\ell$. These numbers must satisfy the constraints

$$\sum_\ell a_\ell = n \quad \text{and} \quad \sum_\ell a_\ell \varepsilon_\ell = \mathcal{E}. \tag{1.5}$$

For given values $a_\ell$, the total number of possible arrangements of the components at different energy levels is $\frac{n!}{a_1! \cdots a_\ell! \cdots}$. When $n$ is large, it is reasonable to assume that the values $a_\ell$ that appear are the ones that maximize the number of arrangements, subject to the constraints (1.5).

To find these optimal $a_\ell$ values, maximize the logarithm of the number of arrangements and introduce Lagrange multipliers $\alpha$ and $\beta$. Thus we wish to differentiate

$$\log \frac{n!}{a_1! \cdots a_\ell! \cdots} \; - \; \alpha\Big(\sum_\ell a_\ell - n\Big) \; - \; \beta\Big(\sum_\ell a_\ell \varepsilon_\ell - \mathcal{E}\Big) \tag{1.6}$$

with respect to $a_\ell$ and set the derivative equal to zero. To use calculus, pretend that the unknowns $a_\ell$ are continuous variables and use Stirling's formula (1.1) in the form $\log n! \sim n(\log n - 1)$. We arrive at

$$\log a_\ell + \alpha + \beta\varepsilon_\ell = 0 \quad \text{for all } \ell.$$

Thus $a_\ell = Ce^{-\beta\varepsilon_\ell}$. Since the total number of components is $n = \sum a_\ell$,

$$a_\ell = \frac{ne^{-\beta\varepsilon_\ell}}{\sum_j e^{-\beta\varepsilon_j}}\,. \tag{1.7}$$

The second constraint gives

$$\mathcal{E} = \frac{n\sum_\ell \varepsilon_\ell\, e^{-\beta\varepsilon_\ell}}{\sum_\ell e^{-\beta\varepsilon_\ell}}\,.$$

These equations should be understood to hold only asymptotically. Divide both equations by $n$ and take $n \to \infty$. We interpret the limit as saying that when a typical system $A$ is immersed in a heat bath at temperature $T$ the system takes energy $\varepsilon_\ell$ with probability

$$p_\ell = \frac{e^{-\beta\varepsilon_\ell}}{\sum_j e^{-\beta\varepsilon_j}} \tag{1.8}$$

and then has average energy

$$U = \frac{\sum_\ell \varepsilon_\ell \, e^{-\beta\varepsilon_\ell}}{\sum_\ell e^{-\beta\varepsilon_\ell}} \,. \tag{1.9}$$

Expression (1.6) suggests that $\beta$ is a function of $\{\varepsilon_\ell\}$ and $U$. We argue with physical reasoning that $\beta$ is in fact a universal function of $T$ alone.

Consider another system $B$ with energy levels $\{\bar{\varepsilon}_m\}$. Let $B_n$ denote a composite system of $n$ identical and independent copies of $B$, also physically independent of $A_n$. Immersing $A_n$ in a heat bath with temperature $T$ specifies a value of $\beta$ for it. Since $\beta$ can a priori depend on $\{\varepsilon_\ell\}$, which is a characteristic of system $A$, we denote this value by $\beta_A$. Similarly, immersing $B_n$ in the same heat bath leads to value $\beta_B$.

We can also immerse $A_n$ and $B_n$ together in the heat bath and consider them together as consisting of $n$ independent and identical copies of a system $AB$. This system acquires its own value $\beta_{AB}$ which depends on the temperature $T$ and on the energies a system $AB$ can take. Since $A$ and $B$ are physically independent, $AB$ can take energies in the set $\{\varepsilon_\ell + \bar{\varepsilon}_m : \ell, m \in \mathbb{N}\}$.

Let $a_{\ell,m}$ be the number of $AB$-components whose $A$-part is at energy level $\varepsilon_\ell$ and whose $B$-part is at energy level $\bar{\varepsilon}_m$, when $A_n$ and $B_n$ are immersed together in the heat bath. Solving the Lagrange multipliers problem for the $AB$-system gives

$$a_{\ell,m} = \frac{n e^{-\beta_{AB}(\varepsilon_\ell + \bar{\varepsilon}_m)}}{\sum_{i,j} e^{-\beta_{AB}(\varepsilon_j + \bar{\varepsilon}_i)}} = n \cdot \frac{e^{-\beta_{AB}\varepsilon_\ell}}{\sum_j e^{-\beta_{AB}\varepsilon_j}} \cdot \frac{e^{-\beta_{AB}\bar{\varepsilon}_m}}{\sum_i e^{-\beta_{AB}\bar{\varepsilon}_i}}.$$

To obtain $a_\ell$, the number of $A$-components at energy $\varepsilon_\ell$, sum over $m$:

$$a_\ell = \sum_m a_{\ell,m} = \frac{n e^{-\beta_{AB}\varepsilon_\ell}}{\sum_j e^{-\beta_{AB}\varepsilon_j}}.$$

Since $A_n$ and $B_n$ do not interact, this must agree with the earlier outcome (1.7):

$$a_\ell = \frac{n e^{-\beta_A \varepsilon_\ell}}{\sum_j e^{-\beta_A \varepsilon_j}} = \frac{n e^{-\beta_{AB}\varepsilon_\ell}}{\sum_j e^{-\beta_{AB}\varepsilon_j}} \quad \text{for all } \ell \in \mathbb{N}.$$

It is reasonable to assume that system $A$ can take at least two different energies $\varepsilon_\ell \neq \varepsilon_{\ell'}$ for otherwise the discussion is trivial. Then the above gives $e^{-\beta_A(\varepsilon_\ell - \varepsilon_{\ell'})} = e^{-\beta_{AB}(\varepsilon_\ell - \varepsilon_{\ell'})}$ and so $\beta_A = \beta_{AB}$. Switching the roles of $A$ and $B$ leads to $\beta_B = \beta_{AB} = \beta_A$. Since system $B$ was arbitrary, we conclude that $\beta$ is a universal function of $T$.

We regard $\beta$ as the more fundamental quantity and view $T$ as a universal function of $\beta$. The state of the system is then determined by the energy levels $\{\varepsilon_\ell\}$ and $\beta$ by equations (1.8) and (1.9).

Next we derive the precise formula for the dependence of $T$ on $\beta$. Working with fixed energies $\varepsilon_\ell$ and considering $\beta$ to be the only variable will not help since we can replace $\beta$ by any monotone function of it and nothing in the above reasoning changes. We need to make energies $\varepsilon_\ell$ vary, which leads to the notion of work done by the system.

The first law of thermodynamics states that if the parameters of the system (i.e. its energies $\varepsilon_\ell$) change, it will absorb an average amount of heat $dQ = d\mathcal{E} + dW$, where $dW$ is the work done by the system. If the energies change by $d\varepsilon_\ell$, then $dW = -\sum_\ell a_\ell \, d\varepsilon_\ell$ and

$$dQ = d\mathcal{E} - \sum_\ell a_\ell \, d\varepsilon_\ell.$$

Let $nS$ be the *entropy* of the system $A_n$. By the second law of thermodynamics

$$dQ = nT \, dS.$$

Define the *free energy* $F = \log \sum e^{-\beta\varepsilon_j}$. Divide the two displays above by $n$ to obtain

$$\begin{aligned} dS &= \frac{1}{T}\Big(dU - \sum p_\ell \, d\varepsilon_\ell\Big) \\ &= \frac{1}{T\beta}\Big(d(\beta U) - U \, d\beta - \beta \sum p_\ell \, d\varepsilon_\ell\Big) \\ &= \frac{1}{T\beta}\Big(d(\beta U) + \frac{\partial F}{\partial \beta} \, d\beta + \sum \frac{\partial F}{\partial \varepsilon_\ell} \, d\varepsilon_\ell\Big) \\ &= \frac{1}{T\beta} \, d(\beta U + F). \end{aligned} \tag{1.10}$$

Abbreviate $G = \beta U + F$, which, by the display above, has to be a function $f(S)$ such that $f'(S) = T\beta$.

Recall that the three systems $A$, $B$, and $AB$ acquire the same $\beta$ when immersed in the heat bath. Consequently $F_A + F_B = F_{AB}$. Since $U = -\frac{\partial F}{\partial \beta}$, the same additivity holds for the function $G$, and so

$$f(S_A) + f(S_B) = f(S_{AB}).$$

Then by (1.10), since $T$ is a universal function of $\beta$, $dS_{AB} = dS_A + dS_B$, which implies $S_{AB} = S_A + S_B + c$. Now we have

$$f(S_A) + f(S_B) = f(S_A + S_B + c).$$

Differentiate in $S_A$ and $S_B$ to see that $f'(S_A) = f'(S_B)$. Since the system $B$ was chosen arbitrarily, entropy $S_B$ can be made equal to any number regardless of the value of temperature $T$. Therefore $f'(S)$ must be a universal constant, which we call $1/k$. (This constant cannot be zero because $T$ and $\beta$ vary with each other.) This implies $\beta = \frac{1}{kT}$ and $G = k^{-1}S$. The constant $k$ is called *Boltzmann's constant*. If $k < 0$, (1.8) would imply that as $T \to 0$

the system chooses the highest energy state, which goes against physical sense. Hence $k > 0$.

Let us compute $S$ for a system with two energy levels $\varepsilon_0$ and $\varepsilon_1$. By symmetry, recentering, and a change of units, we can assume that $\varepsilon_0 = 0$ and $\varepsilon_1 = 1$. The system takes energy 0 with probability $p_0$ and energy 1 with probability $p_1$. The average energy $U = p_1$ and from (1.8) $p_1 = e^{-\beta}/(1 + e^{-\beta})$. Then

$$\begin{aligned} S = kG &= k(\beta U + F) = k\big(p_1(\beta + F) + (1 - p_1)F\big) \\ &= -k\big(p_1 \log p_1 + (1 - p_1)\log(1 - p_1)\big) \\ &= k \log 2 - kI_{1/2}(p_1). \end{aligned}$$

Thus rate function $I_{1/2}$ of Example 1.1 is, up to a universal positive multiplicative factor and an additive constant, the negative thermodynamic entropy of a two-energy system. In the previous section we saw that $-I_{1/2}$ is a linear function (with positive slope) of information-theoretic entropy. Together these observations imply that the thermodynamic entropy of a physical system represents the amount of information needed to describe the system or, equivalently, the amount of uncertainty remaining in it.

The identity $(k\beta)^{-1}S = U + \beta^{-1}F$ expresses an energy-entropy balance and reappears several times later in various guises. It can be found in Exercise 5.19, as equation (7.8) for the Curie-Weiss model, and in Section 8.3 as part (c) of the Dobrushin-Lanford-Ruelle variational principle for lattice systems.

## 1.3. Large deviations as useful estimates

The subject of large deviations is about controlling probabilities of atypical events. There are two somewhat different forms of this activity.

(i) Proofs of limit theorems in probability require estimates to rule out atypical behavior. Such estimates could be called "ad-hoc large deviations".

(ii) Precise limits of vanishing probabilities on an exponential scale are stated as large deviation principles.

The subject of this book is the second kind of large deviations. The next chapter begins a systematic development of large deviation principles. Before that, let us look at two textbook examples to illustrate the use of independence in the derivation of estimates to prove limit theorems.

**Example 1.4.** Let $\{X_n\}$ be an i.i.d. sequence with $E[X] = 0$. (Common device: $X$ is a random variable that has the same distribution as all the $X_n$'s.) We wish to show that, under a suitable hypothesis,

$$S_n/n^p \to 0 \quad P\text{-almost surely (a.s.), for any } p > 1/2. \tag{1.11}$$

In order to illustrate a method, we make a strong assumption. Assume the existence of $\delta > 0$ such that $E[e^{\theta X}] < \infty$ for $|\theta| \le \delta$. When $p \ge 1$ limit (1.11) follows from the strong law of large numbers. So let us assume $p \in (1/2, 1)$.

For $t \ge 0$ Chebyshev's inequality gives

$$P\{S_n \ge \varepsilon n^p\} \le E[e^{tS_n - \varepsilon t n^p}] = \exp\{-\varepsilon t n^p + n \log E[e^{tX}]\}.$$

The exponential moment assumption implies that $E[|X|^k]t^k/k!$ is summable for $t \in [0, \delta]$. Recalling that $E[X] = 0$,

$$\begin{aligned} E[e^{tX}] = E[e^{tX} - tX] &\le 1 + \sum_{k=2}^{\infty} \frac{t^k}{k!} E[|X|^k] \\ &\le 1 + t^2\delta^{-2} \sum_{k=2}^{\infty} \frac{\delta^k}{k!} E[|X|^k] \le 1 + ct^2 \quad \text{for } t \in [0, \delta]. \end{aligned}$$

Then, taking $t = \frac{\varepsilon n^p}{2nc}$ and $n$ large enough,

$$\begin{aligned} P\{S_n \ge \varepsilon n^p\} &\le \exp\{-\varepsilon t n^p + n \log(1 + ct^2)\} \\ &\le \exp\{-\varepsilon t n^p + nct^2\} = \exp\Big\{ -\frac{\varepsilon^2}{4c} n^{2p-1} \Big\}. \end{aligned} \tag{1.12}$$

Applying this to the sequence $\{-X_n\}$ gives the matching bound on the left:

$$P\{S_n \le -\varepsilon n^p\} \le \exp\Big\{ -\frac{\varepsilon^2}{4c} n^{2p-1} \Big\}. \tag{1.13}$$

Inequalities (1.12)–(1.13) can be regarded as large deviation estimates. (Although later we see that since the scale is $n^p$ for $1/2 < p < 1$, technically these are called moderate deviations. But that distinction is not relevant here.) These estimates imply the summability

$$\sum_n P\{|S_n| \ge \varepsilon n^p\} < \infty.$$

The Borel-Cantelli lemma implies that for any $\varepsilon > 0$

$$P\{\exists n_0 : n \ge n_0 \Rightarrow |S_n/n^p| \le \varepsilon\} = 1.$$

A countable intersection over $\varepsilon = 1/k$ for $k \in \mathbb{N}$ gives

$$P\{\forall k \ \exists n_0 : n \ge n_0 \Rightarrow |S_n/n^p| \le 1/k\} = 1,$$

which says that $S_n/n^p \to 0$, $P$-a.s.

We used an unnecessarily strong assumption to illustrate the exponential Chebyshev method. We can achieve the same result with martingales under

the assumption $E[|X|^2] < \infty$. Since $S_n$ is a martingale (relative to the filtration $\sigma(X_1, \dots, X_n)$), Doob's inequality (Theorem 5.4.2 of [**27**] or (8.26) of [**54**]) gives

$$P\Big\{ \max_{k\le n} |S_k| \ge \varepsilon n^p \Big\} \le \frac{1}{\varepsilon^2 n^{2p}} E[|S_n|^2] = \frac{nE[|X|^2]}{\varepsilon^2 n^{2p}} = \frac{c}{\varepsilon^2} n^{-(2p-1)}.$$

Pick $r > 0$ such that $r(2p-1) > 1$. Then,

$$P\Big\{ \max_{k\le m^r} |S_k| \ge \varepsilon m^{pr} \Big\} \le \frac{c_1}{m^{r(2p-1)}}.$$

Hence, $P\{\max_{k\le m^r} |S_k| \ge \varepsilon m^{pr}\}$ is summable over $m$ and the Borel-Cantelli lemma implies that $m^{-rp} \max_{k\le m^r} |S_k| \to 0$ $P$-a.s. as $m \to \infty$.

To get the result for the full sequence pick $m_n$ such that $(m_n - 1)^r \le n < m_n^r$. Then,

$$n^{-p} \max_{k\le n} |S_k| \le \Big(\frac{m_n^r}{n}\Big)^p m_n^{-rp} \max_{k\le m_n^r} |S_k| \longrightarrow 0 \quad \text{as } n \to \infty$$

because $m_n^r/n \to 1$.

**Example 1.5** (Longest run of heads)**.** Let $\{X_n\}$ be an i.i.d. sequence of Bernoulli random variables with success probability $p$. For each $n \ge 1$ let $R_n$ be the length of the longest success run among $(X_1, \dots, X_n)$. We derive estimates to prove a result of Rényi [**66**] that

$$P\Big\{ \lim_{n\to\infty} \frac{R_n}{\log n} = -\frac{1}{\log p} \Big\} = 1. \tag{1.14}$$

Fix $b > 1$ and $r$ such that $r(b-1) > 1$. Let $\ell_m = \lceil -br \log m / \log p \rceil$. ($\lceil x \rceil$ is the smallest integer larger than or equal to $x$.) If $R_{m^r} \ge \ell_m$, then there is an $i \le m^r$ such that $X_i = X_{i+1} = \cdots = X_{i+\ell_m - 1} = 1$. Therefore

$$P\{R_{m^r} \ge \ell_m\} \le m^r p^{\ell_m} \le 1/m^{r(b-1)}. \tag{1.15}$$

By the Borel-Cantelli lemma, with probability one, $R_{m^r} \le \ell_m$ for large enough $m$ (though how large $m$ needs to be is random). Consequently, with probability one,

$$\varlimsup_{m\to\infty} \frac{R_{m^r}}{\log m^r} \le -\frac{b}{\log p}.$$

Given $n$, let $m_n$ be such that $m_n^r \le n < (m_n+1)^r$. Then $R_n \le R_{(m_n+1)^r}$ and

$$\varlimsup_{n\to\infty} \frac{R_n}{\log n} \le \varlimsup_{n\to\infty} \frac{\log(m_n+1)^r}{\log m_n^r} \cdot \frac{R_{(m_n+1)^r}}{\log(m_n+1)^r} \le -\frac{b}{\log p}.$$

Taking $b \searrow 1$ along a sequence shows that

$$P\Big\{ \varlimsup_{n\to\infty} \frac{R_n}{\log n} \le -\frac{1}{\log p} \Big\} = 1.$$

We have the upper bound for the goal (1.14).

Fix $a \in (0,1)$ and let $\ell_n = \lfloor -a \log n / \log p \rfloor$. Let $A_i$ be the event that $X_{i\ell_n+1} = \cdots = X_{(i+1)\ell_n} = 1$. Then

$$\{R_n < \ell_n\} \subset \bigcap_{i=0}^{\lfloor n/\ell_n \rfloor - 1} A_i^c.$$

By the independence of the $A_i$'s

$$P\{R_n < \ell_n\} \le (1 - p^{\ell_n})^{\frac{n}{\ell_n}-1} \le e^{-p^{\ell_n}(\frac{n}{\ell_n}-1)} \le e^{1-n^{1-a}/\ell_n}. \tag{1.16}$$

Once again, by the Borel-Cantelli lemma, $R_n < \ell_n$ happens only finitely often, with probability one, and thus $\underline{\lim}_{n\to\infty} R_n / \log n \ge -a/\log p$. Taking $a \nearrow 1$ proves that

$$P\Big\{ \varliminf_{n\to\infty} \frac{R_n}{\log n} \ge -\frac{1}{\log p} \Big\} = 1.$$

Looking back, the proof relied again on a right-tail estimate (1.15) and a left-tail estimate (1.16). It might be a stretch to call (1.15) a large deviation bound since it is not exponential, but (1.16) can be viewed as a large deviation bound.

**Remark 1.6.** Combining the limit theorem above with the fact that the variance of $R_n$ remains bounded as $n \to \infty$ (see [**10**], [**42**]) provides a very accurate test of the hypothesis that the sequence $\{X_n\}$ is i.i.d. Bernoulli with probability of success $p$.

*Chapter 2*

# The large deviation principle

## 2.1. Precise asymptotics on an exponential scale

Since the 1960s a standard formalism has been employed to express limits of probabilities of rare events on an exponential scale. The term for these statements is *large deviation principle* (LDP). We introduce this in a fairly abstract setting and then return to the Bernoulli example.

There is a sequence $\{\mu_n\}$ of probability measures whose asymptotics we are interested in. These measures exist on some measurable space $(\mathcal{X}, \mathscr{B})$. Throughout our general discussion we take $\mathcal{X}$ to be a Hausdorff topological space, unless further assumptions are placed on it. $\mathscr{B} = \mathscr{B}_{\mathcal{X}}$ is the Borel $\sigma$-algebra of $\mathcal{X}$, and $\mathcal{M}_1(\mathcal{X})$ is the space of probability measures on the measurable space $(\mathcal{X}, \mathscr{B}_{\mathcal{X}})$. Thus $\{\mu_n\}$ is a sequence in the space $\mathcal{M}_1(\mathcal{X})$. In Example 1.1 $\mathcal{X} = \mathbb{R}$ and $\mu_n$ is the probability distribution of $S_n/n$: $\mu_n(A) = P\{S_n/n \in A\}$ for Borel subsets $A \subset \mathbb{R}$.

*Remark on mathematical generality.* A reader not familiar with point-set topology can assume without any harm that $\mathcal{X}$ is a metric space. Even taking $\mathcal{X} = \mathbb{R}$ or $\mathbb{R}^d$ will do for a while. However, later we will study large deviations on spaces of probability measures, and the more abstract point of view becomes a necessity. If the notion of a Borel set is not familiar, it is safe to think of Borel sets as "all the reasonable sets for which a probability can be defined".

To formulate a general large deviation statement, let us look at result (1.2) of Example 1.1 for guidance. The first ingredient of interest in (1.2) is the normalization $n^{-1}$ in front of the logarithm. Obviously this can change

in a different example. Thus we should consider probabilities $\mu_n(A)$ that decay roughly like $e^{-r_n C(A)}$ for some normalization $r_n \nearrow \infty$ and a constant $C(A) \in [0, \infty]$ that depends on the event $A$.

In (1.2) we identified a rate function. How should the constant $C(A)$ relate to a rate function? Consider a finite set $A = \{x_1, \ldots, x_n\}$. Then asymptotically

$$r_n^{-1} \log \mu_n(A) = r_n^{-1} \log \sum_i \mu_n\{x_i\} \approx \max_i r_n^{-1} \log \mu_n\{x_i\}$$

so that $C(A) = \min_i C(x_i)$. This suggests that in general $C(A)$ should be the infimum of a rate function $I$ over $A$.

The final technical point is that it is in general unrealistic to expect $r_n^{-1} \log \mu_n(A)$ to actually converge on account of boundary effects, even if $A$ is a nice set. A reasonable goal is to expect statements in terms of limsup and liminf.

From these considerations we arrive at the following tentative formulation of a large deviation principle: for Borel subsets $A$ of the space $\mathcal{X}$,

$$(2.1) \quad -\inf_{x \in A^\circ} I(x) \le \varliminf_{n\to\infty} \frac{1}{r_n} \log \mu_n(A) \le \varlimsup_{n\to\infty} \frac{1}{r_n} \log \mu_n(A) \le -\inf_{x \in \overline{A}} I(x),$$

where $A^\circ$ and $\overline{A}$ are, respectively, the topological interior and closure of $A$. This statement is basically what we want, except that we need to address the uniqueness of the rate function.

**Example 2.1.** Let us return to the i.i.d. Bernoulli sequence $\{X_n\}$ of Example 1.1. We claim that probability measures $\mu_n(A) = P\{S_n/n \in A\}$ satisfy (2.1) with normalization $r_n = n$ and rate $I_p$ of (1.2). This follows from (1.2) with a small argument.

For an open set $G$ and $s \in G \cap [0,1]$, $\lfloor ns \rfloor / n \in G$ for large enough $n$. So

$$\varliminf_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in G\} \ge \lim_{n\to\infty} \frac{1}{n} \log P\{S_n = \lfloor ns \rfloor\} = -I_p(s).$$

This holds also for $s \in G \setminus [0,1]$ because $I_p(s) = \infty$. Taking the supremum over $s \in G$ on the right gives the inequality

$$\varliminf_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in G\} \ge \sup_{s\in G}\bigl(-I_p(s)\bigr) = -\inf_{s\in G} I_p(s).$$

With $G = A^\circ$ this gives the lower bound in (2.1).

Split a closed set $F$ into $F_1 = F \cap (-\infty, p]$ and $F_2 = F \cap [p, \infty)$. First prove the upper bound in (2.1) for $F_1$ and $F_2$ separately. Let $a = \sup F_1 \le p$ and $b = \inf F_2 \ge p$. (If $F_1$ is empty, then $a = -\infty$, and if $F_2$ is empty, then

$b = \infty$.) Assume first that $a \geq 0$. Then

$$\frac{1}{n} \log P\{S_n/n \in F_1\} \leq \frac{1}{n} \log P\{S_n/n \in [0,a]\} = \frac{1}{n} \log \sum_{k=0}^{\lfloor na \rfloor} P\{S_n = k\}.$$

***Exercise 2.2.** Prove that $P\{S_n = k\}$ increases with $k \leq \lfloor na \rfloor$.

By the exercise above,

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log P\{S_n/n \in F_1\} \leq \lim_{n\to\infty} \frac{1}{n} \log(\lfloor na \rfloor + 1) P\{S_n = \lfloor na \rfloor\} = -I_p(a).$$

This formula is still valid even when $a < 0$ because the probability vanishes. A similar upper bound works for $F_2$. Next write

$$\begin{aligned} \frac{1}{n} \log P\{S_n/n \in F\} &\leq \frac{1}{n} \log \Big( P\{S_n/n \in F_1\} + P\{S_n/n \in F_2\} \Big) \\ &\leq \frac{1}{n} \log 2 + \max \Big( \frac{1}{n} \log P\{S_n/n \in F_1\},\ \frac{1}{n} \log P\{S_n/n \in F_2\} \Big). \end{aligned}$$

$I_p$ is decreasing on $[0,p]$ and increasing on $[p,1]$. Hence, $\inf_{F_1} I_p = I_p(a)$, $\inf_{F_2} I_p = I_p(b)$, and $\inf_F I_p = \min(I_p(a), I_p(b))$. Finally,

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log P\{S_n/n \in F\} \leq -\min(I_p(a), I_p(b)) = -\inf_F I_p.$$

If we now take $F = \overline{A}$, the upper bound in (2.1) follows.

We have shown that (2.1) holds with $I_p$ defined in (1.2). This is our first example of a full-fledged large deviation principle.

**Remark 2.3.** The limsup for closed sets and liminf for open sets in (2.1) remind us of weak convergence of probability measures where the same boundary issue arises. Section B.4 gives the definition of weak convergence.

The following exercises contain other instances where the rate function can be derived by hand.

**Exercise 2.4.** Prove (2.1) for the distribution of the sample mean of an i.i.d. sequence of real-valued normal random variables. Identifying $I$ is part of the task.

**Hint:** The density of $S_n/n$ can be written down explicitly. This suggests $I(x) = (x-\mu)^2/(2\sigma^2)$, where $\mu$ is the mean and $\sigma^2$ is the variance of $X_1$.

**Exercise 2.5.** Prove (2.1) for the distribution of the sample mean of an i.i.d. sequence of exponential random variables and compute the rate function explicitly.

**Hint:** Use Stirling's formula.

## 2.2. Lower semicontinuous and tight rate functions

We continue with some general facts and then in Definition 2.12 state precisely what is meant by a large deviation principle. We recall the definition of a *lower semicontinuous* function.

**Definition 2.6.** A function $f : \mathcal{X} \to [-\infty, \infty]$ is lower semicontinuous if $\{f \le c\} = \{x \in \mathcal{X} : f(x) \le c\}$ is a closed subset of $\mathcal{X}$ for all $c \in \mathbb{R}$.

***Exercise 2.7.** Prove that if $\mathcal{X}$ is a metric space, then $f$ is lower semicontinuous if and only if $\underline{\lim}_{y\to x} f(y) \ge f(x)$ for all $x$.

An important transformation produces a lower semicontinuous function $f_{\rm lsc}$ from an arbitrary function $f : \mathcal{X} \to [-\infty, \infty]$. This *lower semicontinuous regularization* of $f$ is defined by

$$f_{\rm lsc}(x) = \sup\Big\{ \inf_{y\in G} f(y) : G \ni x \text{ and } G \text{ is open}\Big\}. \tag{2.2}$$

This turns out to be the maximal lower semicontinuous minorant of $f$.

**Lemma 2.8.** *$f_{\rm lsc}$ is lower semicontinuous and $f_{\rm lsc}(x) \le f(x)$ for all $x$. If $g$ is lower semicontinuous and satisfies $g(x) \le f(x)$ for all $x$, then $g(x) \le f_{\rm lsc}(x)$ for all $x$. In particular, if $f$ is lower semicontinuous, then $f = f_{\rm lsc}$.*

**Proof.** $f_{\rm lsc} \le f$ is clear. To show $f_{\rm lsc}$ is lower semicontinuous, let $x \in \{f_{\rm lsc} > c\}$. Then there is an open set $G$ containing $x$ such that $\inf_G f > c$. Hence by the supremum in the definition of $f_{\rm lsc}$, $f_{\rm lsc}(y) \ge \inf_G f > c$ for all $y \in G$. Thus $G$ is an open neighborhood of $x$ contained in $\{f_{\rm lsc} > c\}$. So $\{f_{\rm lsc} > c\}$ is open.

To show $g \le f_{\rm lsc}$ one just needs to show that $g_{\rm lsc} = g$. For then

$$\begin{aligned} g(x) &= \sup\Big\{ \inf_G g : x \in G \text{ and } G \text{ is open}\Big\} \\ &\le \sup\Big\{ \inf_G f : x \in G \text{ and } G \text{ is open}\Big\} = f_{\rm lsc}(x). \end{aligned}$$

We already know that $g_{\rm lsc} \le g$. To show the other direction let $c$ be such that $g(x) > c$. Then, $G = \{g > c\}$ is an open set containing $x$ and $\inf_G g \ge c$. Thus $g_{\rm lsc}(x) \ge c$. Now increase $c$ to $g(x)$. □

The above can be reinterpreted in terms of *epigraphs*. The epigraph of a function $f$ is the set $\operatorname{epi} f = \{(x,t) \in \mathcal{X} \times \mathbb{R} : f(x) \le t\}$. For the next lemma we endow $\mathcal{X} \times \mathbb{R}$ with its product topology.

**Lemma 2.9.** *The epigraph of $f_{\rm lsc}$ is the closure of* $\operatorname{epi} f$.

**Proof.** Note that the epigraph of $f_{\rm lsc}$ is closed. That it contains the epigraph of $f$ (and thus also the closure of the epigraph of $f$) is immediate because

$f_{\text{lsc}} \le f$. For the other inclusion we need to show that any open set outside the epigraph of $f$ is also outside the epigraph of $f_{\text{lsc}}$. Let $A$ be such a set and let $(x,t) \in A$. By the definition of the product topology, there is an open neighborhood $G$ of $x$ and an $\varepsilon > 0$ such that $G \times (t-\varepsilon, t+\varepsilon) \subset A$. So for any $y \in G$ and any $s \in (t-\varepsilon, t+\varepsilon)$, $s < f(y)$. In particular, $t + \varepsilon/2 \le \inf_G f \le f_{\text{lsc}}(x)$. So $(x,t)$ is outside the epigraph of $f_{\text{lsc}}$. □

Lower semicontinuous regularization can also be expressed in terms of pointwise alterations of the values of $f$.

**Exercise 2.10.** Assume $\mathcal{X}$ is a metric space. Show that if $x_n \to x$, then $f_{\text{lsc}}(x) \le \varliminf f(x_n)$. Prove that for each $x \in \mathcal{X}$ there is a sequence $x_n \to x$ such that $f(x_n) \to f_{\text{lsc}}(x)$. (The constant sequence $x_n = x$ is allowed here.) This gives the alternate definition $f_{\text{lsc}}(x) = \min(f(x), \varliminf_{y \to x} f(y))$.

Now we apply this to large deviation rate functions. The next lemma shows that rate functions can be assumed to be lower semicontinuous.

**Lemma 2.11.** *Suppose $I$ is a function such that* (2.1) *holds for all measurable sets $A$. Then* (2.1) *continues to hold if $I$ is replaced by $I_{\text{lsc}}$.*

**Proof.** $I_{\text{lsc}} \le I$ and the upper bound is immediate. For the lower bound observe that $\inf_G I_{\text{lsc}} = \inf_G I$ when $G$ is open. □

Due to Lemma 2.11 we will call a $[0,\infty]$-valued function $I$ a *rate function* only when it is lower semicontinuous. Here is the precise definition of a *large deviation principle (LDP)* for the remainder of the text.

**Definition 2.12.** Let $I : \mathcal{X} \to [0,\infty]$ be a lower semicontinuous function and $r_n \nearrow \infty$ a sequence of positive real constants. A sequence of probability measures $\{\mu_n\} \subset \mathcal{M}_1(\mathcal{X})$ is said to satisfy a *large deviation principle with rate function $I$ and normalization $r_n$* if the following inequalities hold:

$$\varlimsup_{n\to\infty} \frac{1}{r_n} \log \mu_n(F) \le -\inf_{x \in F} I(x) \qquad \forall \text{ closed } F \subset \mathcal{X}, \tag{2.3}$$

$$\varliminf_{n\to\infty} \frac{1}{r_n} \log \mu_n(G) \ge -\inf_{x \in G} I(x) \qquad \forall \text{ open } G \subset \mathcal{X}. \tag{2.4}$$

We will abbreviate $\text{LDP}(\mu_n, r_n, I)$ if all of the above hold. When the sets $\{I \le c\}$ are compact for all $c \in \mathbb{R}$, we say $I$ is a *tight* rate function.

Lower semicontinuity makes a rate function unique. For this we assume of $\mathcal{X}$ a little bit more than Hausdorff. A topological space is *regular* if points and closed sets can be separated by disjoint open neighborhoods. In particular, metric spaces are regular topological spaces.

**Theorem 2.13.** *If $\mathcal{X}$ is a regular topological space, then there is at most one (lower semicontinuous) rate function satisfying the large deviation bounds* (2.3) *and* (2.4).

**Proof.** We show that $I$ satisfies

$$I(x) = \sup\Big\{ -\varliminf \frac{1}{r_n} \log \mu_n(B) : B \ni x \text{ and } B \text{ is open} \Big\}.$$

One direction is easy: for all open $B \ni x$

$$-\varliminf \frac{1}{r_n} \log \mu_n(B) \le \inf_B I \le I(x).$$

For the other direction, fix $x$ and let $c < I(x)$. One can separate $x$ from $\{I \le c\}$ by disjoint neighborhoods. Thus, there exists an open set $G$ containing $x$ and such that $\overline{G} \subset \{I > c\}$. (Note that this is true also for $c < 0$, which is relevant in case $I(x) = 0$.) Then

$$\begin{aligned} &\sup\Big\{ -\varliminf \frac{1}{r_n} \log \mu_n(B) : B \ni x \text{ and } B \text{ is open} \Big\} \\ &\quad \ge -\varliminf \frac{1}{r_n} \log \mu_n(G) \ge -\varlimsup \frac{1}{r_n} \log \mu_n(\overline{G}) \ge \inf_{\overline{G}} I \ge c. \end{aligned}$$

Increasing $c$ to $I(x)$ concludes the proof. □

**Remark 2.14.** Tightness of a rate function is a very useful property, as illustrated by the two exercises below. In a large part of the large deviation literature a rate function $I$ is called *good* when the sets $\{I \le c\}$ are compact for $c \in \mathbb{R}$. We prefer the term tight as more descriptive and because of the connection with exponential tightness: see Theorem 2.19 below.

***Exercise 2.15.** Suppose $\mathcal{X}$ is a Hausdorff topological space and let $\mathcal{E} \subset \mathcal{X}$ be a closed set. Assume that the relative topology on $\mathcal{E}$ is metrized by the metric $d$. Let $I : \mathcal{E} \to [0, \infty]$ be a tight rate function and fix an arbitrary closed set $F \subset \mathcal{E}$. Prove that

$$\lim_{\varepsilon \searrow 0} \inf_{F^\varepsilon} I = \inf_F I,$$

where $F^\varepsilon = \{x \in \mathcal{E} : \exists y \in F \text{ such that } d(x, y) < \varepsilon\}$.

***Exercise 2.16.** $\mathcal{X}$ and $\mathcal{E}$ as in the exercise above. Suppose $\xi_n$ and $\eta_n$ are $\mathcal{E}$-valued random variables defined on $(\Omega, \mathscr{F}, P)$, and for any $\delta > 0$ there exists an $n_0 < \infty$ such that $d(\xi_n(\omega), \eta_n(\omega)) < \delta$ for all $n \ge n_0$ and $\omega \in \Omega$.

(a) Show that if the distributions of $\xi_n$ satisfy the lower large deviation bound (2.4) with some rate function $I : \mathcal{E} \to [0, \infty]$, then so do the distributions of $\eta_n$.

(b) Show that if the distributions of $\xi_n$ satisfy the upper large deviation bound (2.3) with some tight rate function $I : \mathcal{E} \to [0, \infty]$, then so do the distributions of $\eta_n$.

## 2.3. Weak large deviation principle

It turns out that it is sometimes difficult to satisfy the upper bound (2.3) for all closed sets. A useful weakening of the LDP requires the upper bound only for compact sets.

**Definition 2.17.** A sequence of probability measures $\{\mu_n\} \subset \mathcal{M}_1(\mathcal{X})$ satisfies a *weak large deviation principle* with lower semicontinuous rate function $I : \mathcal{X} \to [0, \infty]$ and normalization $\{r_n\}$ if the lower large deviation bound (2.4) holds for all open sets $G \subset \mathcal{X}$ and the upper large deviation bound (2.3) holds for all compact sets $F \subset \mathcal{X}$.

With enough control on the tails of the measures $\mu_n$, a weak LDP is sufficient for the full LDP.

**Definition 2.18.** We say $\{\mu_n\} \subset \mathcal{M}_1(\mathcal{X})$ is *exponentially tight* with normalization $r_n$ if for each $0 < b < \infty$ there exists a compact set $K_b$ such that

$$\overline{\lim_{n\to\infty}} \frac{1}{r_n} \log \mu_n(K_b^c) \leq -b. \tag{2.5}$$

**Theorem 2.19.** *Assume the upper bound* (2.3) *holds for compact sets and* $\{\mu_n\}$ *is exponentially tight with normalization* $r_n$*. Then the upper bound* (2.3) *holds for all closed sets with the same rate function* $I$*.*

*If the weak* LDP$(\mu_n, r_n, I)$ *holds and* $\{\mu_n\}$ *is exponentially tight with normalization* $r_n$*, then the full* LDP$(\mu_n, r_n, I)$ *holds and* $I$ *is a tight rate function.*

**Proof.** Let $F$ be a closed set.

$$\begin{aligned}
\overline{\lim_{n\to\infty}} \frac{1}{r_n} \log \mu_n(F) &\leq \overline{\lim_{n\to\infty}} \frac{1}{r_n} \log\bigl(\mu_n(F \cap K_b) + \mu_n(K_b^c)\bigr) \\
&\leq \max\Bigl(-b\,,\ \overline{\lim_{n\to\infty}} \frac{1}{r_n} \log \mu_n(F \cap K_b)\Bigr) \\
&\leq \max\bigl(-b\,, -\inf_{F\cap K_b} I\,\bigr) \\
&\leq \max\bigl(-b\,, -\inf_{F} I\,\bigr).
\end{aligned}$$

Letting $b \nearrow \infty$ proves the upper large deviation bound (2.3).

The weak LDP already contains the lower large deviation bound (2.4) and so we have both bounds. From the lower bound and exponential tightness it follows that

$$\inf_{K_{b+1}^c} I \geq -\varliminf_{n\to\infty} \frac{1}{r_n} \log \mu_n(K_{b+1}^c) \geq b+1.$$

This implies that $\{I \leq b\} \subset K_{b+1}$. As a closed subset of a compact set $\{I \leq b\}$ is compact. □

The connection between a tight rate function and exponential tightness is an equivalence if we assume a little more of the space. To prove the other implication in Theorem 2.21 below we give an equivalent reformulation of exponential tightness in terms of open balls. In a metric space $(\mathcal{X}, d)$, $B(x,r) = \{y \in \mathcal{X} : d(x,y) < r\}$ is the open $r$-ball centered at $x$.

**Lemma 2.20.** *Let $\{\mu_n\}$ be a sequence of probability measures on a Polish space $\mathcal{X}$. (A Polish space is a complete and separable metric space.) Then $\{\mu_n\}$ is exponentially tight if and only if for every $b < \infty$ and $\delta > 0$ there exist finitely many $\delta$-balls $B_1, \ldots, B_m$ such that*

$$\mu_n\Big(\Big[\bigcup_{i=1}^m B_i\Big]^c\Big) \leq e^{-r_n b} \quad \forall\, n \in \mathbb{N}.$$

**Proof.** Ulam's theorem (page 282) says that on a Polish space an individual probability measure $\nu$ is tight, which means that $\forall \varepsilon > 0$ there exists a compact set $A$ such that $\nu(A^c) < \varepsilon$. Consequently on such a space exponential tightness is equivalent to the stronger statement that for all $b < \infty$ there exists a compact set $K_b$ such that $\mu_n(K_b^c) \leq e^{-r_n b}$ for all $n \in \mathbb{N}$.

Since a compact set can be covered by finitely many $\delta$-balls, the ball condition is a consequence of this stronger form of exponential tightness.

Conversely, assume the ball condition and let $1 \leq b < \infty$. We need to produce the compact set $K_b$. For each $k \in \mathbb{N}$, find $m_k$ balls $B_{k,1}, \ldots, B_{k,m_k}$ of radius $k^{-1}$ such that

$$\mu_n\Big(\Big[\bigcup_{i=1}^{m_k} B_{k,i}\Big]^c\Big) \leq e^{-2kr_n b} \quad \forall\, n \in \mathbb{N}.$$

Let $K = \bigcap_{k=1}^\infty \bigcup_{i=1}^{m_k} \overline{B}_{k,i}$. As a closed subset of $\mathcal{X}$, $K$ is complete. By its construction $K$ is totally bounded. This means that for any $\varepsilon > 0$ it can be covered by finitely many $\varepsilon$-balls. Completeness and total boundedness are equivalent to compactness in a metric space [**26**, Theorem 2.3.1]. By

explicitly evaluating the geometric series and some elementary estimation,

$$\mu_n(K^c) \le \sum_{k=1}^{\infty} e^{-2kr_nb} \le e^{-r_nb}$$

as long as $r_n \ge 1$. Exponential tightness has been verified. □

**Theorem 2.21.** *Suppose $\mathcal{X}$ is a Polish space. Assume probability measures $\{\mu_n\}$ satisfy the upper large deviation bound* (2.3) *with a tight rate function $I$. Then $\{\mu_n\}$ is exponentially tight.*

**Proof.** Let $\{x_i\}_{i\in\mathbb{N}}$ be a countable dense set in $\mathcal{X}$. Suppose we can show that for every $b < \infty$ and $\varepsilon > 0$ there exists $m \in \mathbb{N}$ such that

$$\varlimsup_{n\to\infty} r_n^{-1} \log \mu_n\Big(\Big[\bigcup_{i=1}^{m} B(x_i,\varepsilon)\Big]^c\Big) \le -b. \tag{2.6}$$

This is sufficient for exponential tightness by Lemma 2.20. (See Exercise 2.22 below.)

To show (2.6), take $m$ large enough so that the compact set $\{I \le b\}$ is covered by $G = B(x_1,\varepsilon) \cup \cdots \cup B(x_m,\varepsilon)$. (Since $\{x_i\}$ is dense, the entire space is covered by $\bigcup_{i\ge1} B(x_i,\varepsilon)$, and by compactness $\{I \le b\}$ has a finite subcover.) By the upper large deviation bound,

$$\varlimsup_{n\to\infty} r_n^{-1} \log \mu_n(G^c) \le -\inf_{x\in G^c} I(x) \le -b. \qquad \square$$

Here is the missing detail from the proof.

***Exercise 2.22.** Show that the condition of Lemma 2.20 follows from the condition established in the proof above. The fact that the balls $B(x_i,\varepsilon)$ cover the entire space is again crucial.

The results of this section offer a strategy for proving an LDP. First prove a weak LDP and then verify exponential tightness. A weak LDP may be easier to prove because it reduces entirely to analyzing asymptotics of $r_n^{-1} \log \mu_n(B(x,\varepsilon))$ for small neighborhoods. This idea already appeared in the proof of Example 2.1 where we reduced the proof to asymptotics of point probabilities. Here is an example where this method applies.

**Exercise 2.23.** Prove the large deviation principle for the distribution of the sample mean $S_n/n$ of an i.i.d. sequence of $\mathbb{R}^d$-valued normal random variables with mean $m$ and nonsingular covariance matrix $A$.

**Hint:** The density of $S_n/n$ suggests $I(x) = \frac{1}{2}(x-m)\cdot A^{-1}(x-m)$. Note that this is different from the one-dimensional case in Exercise 2.4 because one cannot use monotonicity of $I$ and split closed sets $F$ into a part below $m$ and a part above $m$.

We end the section with an important theoretical exercise.

***Exercise 2.24.** For $x \in \mathcal{X}$, define upper and lower local rate functions by

$$\overline{\kappa}(x) \;=\; -\inf_{G \subset \mathcal{X}:\ G \text{ open},\ x \in G} \ \varlimsup_{n\to\infty} \frac{1}{r_n} \log \mu_n(G) \tag{2.7}$$

and

$$\underline{\kappa}(x) \;=\; -\inf_{G \subset \mathcal{X}:\ G \text{ open},\ x \in G} \ \varliminf_{n\to\infty} \frac{1}{r_n} \log \mu_n(G). \tag{2.8}$$

Show that if $\overline{\kappa} = \underline{\kappa} = \kappa$, then the weak LDP holds with rate function $\kappa$. Note that, by monotonicity, the same infimum in (2.7) and (2.8) can be taken over any base of open neighborhoods at $x$.

## 2.4. Aspects of Cramér's theorem

Cramér's theorem is the LDP for the sample mean

$$S_n/n = (X_1 + \cdots + X_n)/n$$

of i.i.d. random variables $\{X_n\}$ with values in $\mathbb{R}$ or $\mathbb{R}^d$. Discussion around this theorem raises several basic themes of large deviation theory: moment generating functions, compactness, convexity, minimax theorems, and the change of measure argument. We prove partial results here and formulate many statements as exercises with hints for hands-on practice. The important themes appear again later, so this section can be skipped, though we would recommend that the reader at least skim the main points. A complete proof of Cramér's theorem in $\mathbb{R}^d$ is given in Section 4.3.

We start by stating the one-dimensional theorem. Let $\{X_n\}$ be i.i.d. real-valued random variables, and let $X$ be another random variable with the same distribution. The moment generating function is $M(\theta) = E[e^{\theta X}]$ for $\theta \in \mathbb{R}$. $M(\theta) > 0$ always and $M(\theta) = \infty$ is possible. Define

$$I(x) = \sup_{\theta \in \mathbb{R}} \{\theta x - \log M(\theta)\}. \tag{2.9}$$

Since $M(0) = 1$, $I : \mathbb{R} \to [0, \infty]$ is a well-defined function.

**Cramér's theorem on $\mathbb{R}$.** *Let $\{X_n\}$ be a sequence of i.i.d. real-valued random variables. Let $\mu_n$ be the distribution of the sample mean $S_n/n$. Then the large deviation principle* $\mathrm{LDP}(\mu_n, n, I)$ *is satisfied with $I$ defined in* (2.9).

A proof of this general one-dimensional Cramér theorem that applies to all i.i.d. sequences can be found in [**15**]. The case where $M$ is finite in a neighborhood of 0 is covered by our multidimensional Cramér theorem in Section 4.3. Here we develop the upper bound and some related facts as a

series of exercises. Then we turn to discuss parts of the multidimensional Cramér theorem under stronger assumptions.

Using Chebyshev's inequality

$$P\{S_n \geq nb\} \leq e^{-n\theta b} E[e^{\theta S_n}] = e^{-n\theta b} M(\theta)^n \quad \text{for } \theta \geq 0 \tag{2.10}$$

and

$$P\{S_n \leq na\} \leq e^{-n\theta a} E[e^{\theta S_n}] = e^{-n\theta a} M(\theta)^n \quad \text{for } \theta \leq 0. \tag{2.11}$$

From the above we get immediately the upper bounds

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log P\{S_n \geq nb\} \leq -\sup_{\theta \geq 0}\{\theta b - \log M(\theta)\}$$

and

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log P\{S_n \leq na\} \leq -\sup_{\theta \leq 0}\{\theta a - \log M(\theta)\}.$$

***Exercise 2.25.** Suppose $X$ has a finite mean $\bar{x} = E[X]$. Prove that if $a \leq \bar{x} \leq b$, then

$$\sup_{\theta \geq 0}\{\theta b - \log M(\theta)\} = \sup_{\theta \in \mathbb{R}}\{\theta b - \log M(\theta)\}$$

and

$$\sup_{\theta \leq 0}\{\theta a - \log M(\theta)\} = \sup_{\theta \in \mathbb{R}}\{\theta a - \log M(\theta)\}.$$

**Hint:** Use Jensen's inequality to show that $\theta b - \log M(\theta) \leq 0$ for $\theta < 0$ and $\theta a - \log M(\theta) \leq 0$ for $\theta > 0$.

**Definition 2.26.** A subset $A$ of a vector space $\mathcal{X}$ is *convex* if for all $x, y \in A$ and $t \in [0,1]$, $tx + (1-t)y \in A$. A function $f : \mathcal{X} \to [-\infty, \infty]$ is *convex* if $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ for all $x, y \in \mathcal{X}$ and $t \in [0,1]$ such that the right-hand side of the inequality is well-defined (that is, not $\infty - \infty$).

***Exercise 2.27.** Prove that $I$ is lower semicontinuous and convex and that if $\bar{x} = E[X]$ is finite, then $I$ achieves its minimum at $\bar{x}$ with $I(\bar{x}) = 0$.

**Hint:** $I$ is a supremum of lower semicontinuous convex functions. $I(x) \geq 0$ for all $x$, but by Jensen's inequality $I(\bar{x}) \leq 0$.

***Exercise 2.28.** Suppose $M(\theta) < \infty$ in some open neighborhood around the origin. Show that then $\bar{x}$ is the *unique* zero of $I$: that is, $x \neq \bar{x}$ implies $I(x) > 0$.

**Hint:** For any $x > \bar{x}$, $(\log M(\theta))' < x$ for $\theta$ in some interval $(0, \delta)$.

**Exercise 2.29.** Check that the rate functions found in Example 1.1 and Exercises 2.4 and 2.5 match (2.9).

Exercise 2.27 together with the earlier observations shows that when $\bar{x}$ is finite $I(x)$ is nonincreasing for $x < \bar{x}$ and nondecreasing for $x > \bar{x}$. In particular, if $a \le \bar{x} \le b$, then $I(a) = \inf_{x \le a} I(x)$ and $I(b) = \inf_{x \ge b} I(x)$. This proves the upper bound for the sets $F = (-\infty, a]$ and $F = [b, \infty)$ in the case where the mean is finite.

**Exercise 2.30.** Prove that the sample mean $S_n/n$ of i.i.d. real-valued random variables satisfies the upper large deviation bound (2.3) with normalization $n$ and rate $I$ defined in (2.9), with no further assumptions on the distribution.

**Hint:** The case of finite mean is almost done above. Then consider separately the cases where the mean is infinite and where the mean does not exist.

While Cramér's theorem is valid in general, it does not give much information unless the variables have exponentially decaying tails. This point is explored in the next exercise.

**Exercise 2.31.** Let $\{X_i\}$ be an i.i.d. real-valued sequence. Assume $E[X_1^2] < \infty$ but, for any $\varepsilon > 0$, $P\{X_1 > b\} > e^{-\varepsilon b}$ for all large enough $b$. Show that

(a) $\lim_{n\to\infty} \frac{1}{n} \log P\{S_n/n > E[X_1] + \delta\} = 0$ for any $\delta > 0$,

(b) the rate function is identically 0 on $[E(X_1), \infty)$.

**Hint:** For (a), deduce

$$P\{S_n/n \ge E[X_1] + \delta\} \ge P\{S_{n-1} \ge (n-1)E[X_1]\}P\{X_1 \ge n\delta + E[X_1]\}$$

and apply the central limit theorem. For (b), first find $M(\theta)$ for $\theta > 0$. Then observe that for $\theta \le 0$ and $x \ge E[X_1]$,

$$\theta x - \log M(\theta) \le \theta(x - E[X_1]) \le 0.$$

**Exercise 2.32.** Let $\{X_i\}$ be an i.i.d. real-valued sequence. Prove that the closure of the set $\{I < \infty\}$ is the same as the closure of the convex hull of the support of the distribution of $X$. (The *convex hull* of a set is the intersection of all convex sets containing it.)

**Hint:** Let $K$ be the latter set and let $y \notin K$. To show that $I(y) = \infty$, find $\theta \in \mathbb{R}$ such that $\theta y - \varepsilon > \sup_{x \in K} x\theta$. For the other direction, take $y$ in the interior of $\{I = \infty\}$. To get $y \notin K$, show first that there exists a sequence $\theta_n$ converging to either $\infty$ or $-\infty$ such that $\phi_y(\theta_n) = \theta_n y - \log M(\theta_n)$ converges to infinity. Assume $\theta_n \to \infty$. Show that for some $\varepsilon$, $|x - y| \le \varepsilon$ implies $\phi_x(\theta) \to \infty$ as $\theta \to \infty$. Then, for $\theta > 0$, $\theta(y - \varepsilon) - \log M(\theta) \le -\log \mu\{x : |x - y| \le \varepsilon\}$ where $\mu$ is the distribution of $X$. Let $\theta \to \infty$.

Cramér's theorem is quite crude because only the exponentially decaying terms of a full expansion affect the result. In some cases one can derive much more precise asymptotics.

**Exercise 2.33.** Prove that if the $\{X_k\}$ are i.i.d. standard normal, then for any $k \in \mathbb{N}$ and $a > 0$

$$\log P\{S_n \geq an\} \sim -\frac{a^2 n}{2} - \frac{1}{2}\log(2\pi n a^2) + \log\Big(1 - \frac{1}{a^2 n} + \frac{1 \cdot 3}{a^4 n^2} - \cdots + (-1)^k \frac{1 \cdot 3 \cdot 5 \cdots (2k-1)}{a^{2k} n^k}\Big).$$

**Hint:** Observe that

$$\frac{d}{dx}\Big(e^{-x^2/2} \sum_{k=0}^{n} (-1)^k (1 \cdot 3 \cdots (2k-1)) x^{-2k-1}\Big) \begin{cases} < -e^{-x^2/2} & \text{if } n \text{ is even,} \\ > -e^{-x^2/2} & \text{if } n \text{ is odd.} \end{cases}$$

**Exercise 2.34.** Continuing Exercise 2.29, derive Cramér rate functions for further basic distributions.

(a) For real $\alpha > 0$, the rate $\alpha$ exponential distribution has density $f(x) = \alpha e^{-\alpha x}$ on $\mathbb{R}_+$. Derive the Cramér rate function

$$I(x) = \alpha x - 1 - \log \alpha x \qquad \text{for } x > 0.$$

(b) For real $\lambda > 0$, the mean $\lambda$ Poisson distribution has probability mass function $p(k) = e^{-\lambda}\lambda^k/k!$ for $k \in \mathbb{Z}_+$. Derive the Cramér rate function

$$I(x) = x \log(x/\lambda) - x + \lambda \qquad \text{for } x \geq 0.$$

We turn to Cramér's theorem in multiple dimensions. When the $\{X_n\}$ are $\mathbb{R}^d$-valued, the moment generating function is given by $M(\theta) = E[e^{\theta \cdot X}]$ for $\theta \in \mathbb{R}^d$. Again, $M(\theta) \in (0, \infty]$. Define

$$I(x) = \sup_{\theta \in \mathbb{R}^d} \{\theta \cdot x - \log M(\theta)\}. \tag{2.12}$$

**Exercise 2.35.** Check that Exercises 2.27 and 2.28 apply to the multi-dimensional case as well.

Hölder's inequality implies that $\log M(\theta)$ is a convex function: with $t \in (0,1)$, $p = 1/t$, and $q = 1/(1-t)$,

$$\begin{aligned} M(t\theta_1 + (1-t)\theta_2) &= E[e^{t\theta_1 \cdot X} e^{(1-t)\theta_2 \cdot X}] \\ &\leq E[e^{\theta_1 \cdot X}]^t E[e^{\theta_2 \cdot X}]^{1-t} = M(\theta_1)^t M(\theta_2)^{1-t}. \end{aligned} \tag{2.13}$$

The full LDP of the one-dimensional Cramér theorem does not generalize to multiple dimensions without an additional assumption. Counterexamples appear in [**20**].

**Cramér's theorem on $\mathbb{R}^d$.** *Let $\{X_n\}$ be a sequence of i.i.d. $\mathbb{R}^d$-valued random variables and let $\mu_n$ be the distribution of the sample mean $S_n/n$. Then without further assumptions weak* LDP$(\mu_n, n, I)$ *holds with $I$ defined in* (2.12). *If, moreover, $M(\theta) < \infty$ in a neighborhood of* 0, *then* LDP$(\mu_n, n, I)$ *holds and $I$ is a tight rate function.*

At this point we prove the upper bound for compact sets without assumptions on $M$ and then exponential tightness assuming that $M$ is finite near the origin. Then we give a proof of the lower bound under the restrictive assumption

$$M(\theta) < \infty \text{ for all } \theta \in \mathbb{R}^d \text{ and } |\theta|^{-1} \log M(\theta) \to \infty \text{ as } |\theta| \to \infty. \tag{2.14}$$

Both proofs introduce important techniques. Assumption (2.14) ensures that the supremum in (2.12) is achieved. This is precisely the issue that needs to be overcome when no assumptions on $M$ are present. In Section 4.3 we revisit the theorem and prove its final version.

**Proof of the upper bound for compact sets and exponential tightness.** For any Borel set $C$ and $\theta \in \mathbb{R}^d$,

$$\begin{aligned} P\{S_n/n \in C\} &= E[\mathbb{1}\{S_n/n \in C\}] \le e^{-\inf_{y\in C} n\theta\cdot y} E[e^{\theta\cdot S_n}] \\ &= e^{-n\inf_{y\in C}\theta\cdot y} M(\theta)^n. \end{aligned}$$

This shows that

$$\frac{1}{n}\log P\{S_n/n \in C\} \le -\sup_{\theta}\inf_{y\in C}\{\theta\cdot y - \log M(\theta)\}. \tag{2.15}$$

We would like to interchange the sup and the inf to find $I(y)$ on the right-hand side. This can be done if $C$ is a compact convex set.

**Minimax theorem on $\mathbb{R}^d$.** *Let $C \subset \mathbb{R}^d$ be compact and convex. Let $D \subset \mathbb{R}^d$ be convex. Let $f : C \times D \to \mathbb{R}$ be such that for each $\theta \in D$, $f(y,\theta)$ is convex and continuous in $y \in C$, and for each $y \in C$, $f(y,\theta)$ is concave in $\theta \in D$. Then*

$$\sup_{\theta\in D}\inf_{y\in C} f(y,\theta) = \inf_{y\in C}\sup_{\theta\in D} f(y,\theta).$$

This theorem is a special case of the more general minimax theorem proved in Section A.5 of Appendix A. To have a feeling for the theorem above think of a horse saddle in $\mathbb{R}^3$. We have a smooth function that is convex in one direction and concave in the other. Taking sup in the concave direction and inf in the convex direction will result in the saddle point regardless of the order.

The set $D = \{\theta : M(\theta) < \infty\}$ is convex by (2.13), $C$ is a compact convex set by assumption, and $f(y,\theta) = \theta\cdot y - \log M(\theta)$ satisfies the assumptions of the minimax theorem. Thus the sup and the inf can be switched in (2.15) to give

$$\frac{1}{n}\log P\{S_n/n \in C\} \le -\inf_{y\in C}\sup_{\theta}\{\theta\cdot y - \log M(\theta)\} = -\inf_{y\in C} I(y). \tag{2.16}$$

We have the upper bound with rate function $I$ of (2.12) for compact convex sets, even without taking the $n \to \infty$ limit.

We extend the upper bound to an arbitrary compact set $K$. Let $\alpha < \inf_K I$. Since $I$ is lower semicontinuous, $\{I > \alpha\}$ is open. For each $x \in K \subset \{I > \alpha\}$ pick a compact ball $C_x$ centered at $x$ with nonempty interior and such that $C_x \subset \{I > \alpha\}$. Cover $K$ with a finite collection $C_{x_1}, \ldots, C_{x_N}$ of such balls. The upper bound for compact convex sets gives

$$P\{S_n/n \in K\} \le \sum_{i=1}^{N} P\{S_n/n \in C_{x_i}\} \le \sum_{i=1}^{N} e^{-n \inf_{C_{x_i}} I} \le N e^{-n\alpha}.$$

Taking $n \nearrow \infty$ and then $\alpha \nearrow \inf_K I$ gives the upper bound (2.3) in weak $\mathrm{LDP}(\mu_n, n, I)$.

Last, we verify exponential tightness under the assumption that $M$ is finite near the origin. Theorem 2.19 then implies the upper bound for closed sets. To this end, from (2.10) and (2.11) it follows that for any $b > 0$ we can find a large enough $a = a(b) > 0$ such that

$$P\{|S_n^{(i)}| \ge na\} \le e^{-bn} \quad \text{for } i = 1, 2, \ldots, d \text{ and all } n \in \mathbb{N}.$$

Here $y^{(i)}$ denotes the $i$th coordinate of a vector $y \in \mathbb{R}^d$. Definition 2.18 of exponential tightness is satisfied with $r_n = n$ and $K_b = \{y : |y^{(i)}| \le a(b) \text{ for all } i = 1, \ldots, d\}$. □

**Exercise 2.36.** The minimax theorem was used above to turn (2.15) into the nonasymptotic upper bound (2.16) for compact convex sets. This was done to illustrate the minimax trick and because bounds that are valid for finite $n$ are useful. However, we can proceed directly from (2.15) to the upper large deviation bound for a general compact set $K$. Fill in the details in the following outline. With notation as above, for each $x \in K$ find $\theta_x$ such that $\theta_x \cdot x - \log M(\theta_x) > \alpha$. Pick a compact convex ball $U_x$ centered at $x$ and with nonempty interior such that $\theta_x \cdot y - \log M(\theta_x) > \alpha - \varepsilon$ for $y \in U_x$. Proceed as in the proof above.

**Proof of Cramér's lower bound under** (2.14)**.** We introduce the classical *change of measure* argument for the lower bound. Let our random variables $\{X_k\}$ be defined on a probability space $(\Omega, \mathscr{F}, P)$.

On any open set where $M(\theta)$ is finite it is differentiable and $\nabla M(\theta) = E[Xe^{\theta \cdot X}]$. This is by dominated convergence. Thus $\theta \cdot x - \log M(\theta)$ is a concave differentiable function of $\theta$ that, by (2.14), achieves its maximum $I(x)$ at some $\theta_x$. Then $\nabla M(\theta_x) = xM(\theta_x)$.

Define the probability measure $\nu_x$ on $\mathbb{R}^d$ by

$$\nu_x(B) = \frac{1}{M(\theta_x)} E[e^{\theta_x \cdot X} \mathbb{1}\{X \in B\}], \quad B \in \mathscr{B}_{\mathbb{R}^d}.$$

The mean of $\nu_x$ is

$$\int_{\mathbb{R}^d} y\,\nu_x(dy) = \frac{E[Xe^{\theta_x\cdot X}]}{M(\theta_x)} = \frac{\nabla M(\theta_x)}{M(\theta_x)} = x.$$

Let $Q_{x,n}$ be the probability measure on $\Omega$ defined by

$$Q_{x,n}(A) = \frac{E[\mathbb{1}_A \cdot e^{\theta_x\cdot S_n}]}{E[e^{\theta_x\cdot S_n}]} \quad \text{for } A \in \mathscr{F}.$$

Now for the lower bound. Take an open set $G \subset \mathbb{R}^d$, $x \in G$, and $\varepsilon > 0$ such that $\{y : |y - x| < \varepsilon\} \subset G$. Then

$$\begin{aligned} P\{S_n/n \in G\} &\geq P\{|S_n - nx| < \varepsilon n\} \\ &\geq e^{-n\theta_x\cdot x - n\varepsilon|\theta_x|} E[e^{\theta_x\cdot S_n}\mathbb{1}\{|S_n - nx| < \varepsilon n\}] \\ &= e^{-n\theta_x\cdot x - n\varepsilon|\theta_x|} M(\theta_x)^n Q_{x,n}\{|S_n - nx| < \varepsilon n\}. \end{aligned}$$

The key observation is that under $Q_{x,n}$ the variables $X_1, X_2, \ldots, X_n$ are i.i.d. $\nu_x$-distributed: for $B_1, \ldots, B_n \in \mathscr{B}_{\mathbb{R}^d}$,

$$Q_{x,n}\Big(\bigcap_{k=1}^n \{X_k \in B_k\}\Big) = \prod_{k=1}^n \frac{E[\mathbb{1}_{B_k}(X)e^{\theta_x\cdot X}]}{E[e^{\theta_x\cdot X}]} = \prod_{k=1}^n \nu_x(B_k).$$

By the law of large numbers $Q_{x,n}\{|S_n - nx| < \varepsilon n\} \to 1$, and we get the bound

$$\varliminf_{n\to\infty} \frac{1}{n} \log P\{S_n/n \in G\} \geq -I(x) - \varepsilon|\theta_x|.$$

Taking $\varepsilon \to 0$ and sup over $x \in G$ on the right proves the lower bound (2.4). □

The measure $\nu_x$ is called the *tilted* measure. The dependence on $n$ in $Q_{x,n}$ is an artifact we can eliminate by using a single infinite product measure on a sequence space. This is what we do in Section 5.2 on Sanov's theorem.

The change of measure argument replaced the original measure $P$ by a new measure $Q_{x,n}$ under which outcome $x$ became typical rather than rare. In the proof this appears to be merely a trick, but we shall see later that there is more to it. Namely, to produce the deviation $S_n \approx nx$ the process $\{X_k\}$ actually behaves like an i.i.d. $\nu_x$-sequence. This is an interesting conclusion. A priori one could also imagine that the system prefers to deviate a small number of variables while letting most $X_k$'s behave in a typical fashion. (See Exercises 2.38 and 6.19 and the related maximum entropy principle in Section 5.3.) A lesson of large deviation theory is that a deviation is not produced in an arbitrary manner, but rather in the most probable way, and this can be captured by the rate function.

**Exercise 2.37.** Let $\{X_n\}$ be i.i.d. Bernoulli random variables with success probability $p \in [0,1]$. Show that for $s \in [0,1]$ the measure $\nu_s$ in the proof above is the Bernoulli measure with success probability $s$. Investigate $\nu_x$ for your other favorite distributions.

**Exercise 2.38.** Let $S_n = X_1 + \cdots + X_n$ be a simple symmetric random walk on $\mathbb{Z}$. That is, the $\{X_k\}$ are i.i.d. with distribution $P(X_k = \pm 1) = 1/2$. Let $a \in [0,1]$. With elementary calculation find the limit of the process $\{X_k\}$ conditioned on $|S_n - \lfloor na \rfloor| \le 1$, as $n \to \infty$.

**Hint:** Fix $x_1, \ldots, x_m \in \{\pm 1\}$, write the probability $P(X_1 = x_1, \ldots, X_m = x_m \,|\, |S_n - \lfloor na \rfloor| \le 1)$ in terms of factorials, and observe the asymptotics. Note that the conditioning event cannot always be written $S_n = \lfloor na \rfloor$ because $S_n$ must have the parity of $n$.

## 2.5. Limits, deviations, and fluctuations

Let $\{Y_n\}$ be a sequence of random variables with values in a metric space $(\mathcal{X}, d)$ and let $\mu_n$ be the distribution of $Y_n$, that is, $\mu_n(B) = P\{Y_n \in B\}$ for $B \in \mathscr{B}_{\mathcal{X}}$. Naturally an LDP for the sequence $\{\mu_n\}$ is related to the asymptotic behavior of $Y_n$. Suppose $\mathrm{LDP}(\mu_n, r_n, I)$ holds and $Y_n \to \bar{y}$ in probability. Then the limit $\bar{y}$ does not represent a deviation. The rate function $I$ recognizes this with the value $I(\bar{y}) = 0$ that follows from the upper bound. For any open neighborhood $G$ of $\bar{y}$ we have $\mu_n(G) \to 1$. Consequently for the closure

$$0 \le \inf_{\overline{G}} I \le -\overline{\lim}\, r_n^{-1} \log \mu_n(\overline{G}) = 0.$$

Let $G$ shrink down to $\bar{y}$. Lower semicontinuity forces $I(\bar{y}) = 0$.

Every LDP satisfies $\inf I = 0$, as can be seen by taking $F = \mathcal{X}$ in the upper bound (2.3). But the zero set of $I$ does not necessarily represent limit values. It may simply be that the probability of a deviation decays slower than exponentially in $r_n$, which leads to $I = 0$.

**Exercise 2.39.** In case the reader prefers an off-the-shelf example rather than playing with his or her own examples, here is one. Fix a sequence $0 < a_n \nearrow \infty$, let $m$ denote Lebesgue measure, and define $\{\mu_n\}$ on $\mathbb{R}$ by

$$\mu_n(A) = (1 - a_n^{-1})\mathbb{1}_A(0) + a_n^{-1} m(A \cap (0,1]).$$

Clearly $\mu_n \to \delta_0$ weakly or, equivalently, if $Y_n$ has distribution $\mu_n$, then $Y_n \to 0$ in probability. Given any $c \in [0, \infty]$, show that by an appropriate choice of $r_n$ we can have the LDP with rate function

$$I(x) = \begin{cases} 0, & x = 0, \\ c, & x \in (0,1], \\ \infty, & x \notin [0,1]. \end{cases}$$

Returning to the general discussion, an LDP can imply convergence of the random variables if the rate function has good properties. Assume that $I$ is a tight rate function and has a unique zero $I(\bar{y}) = 0$. Let $A = \{y : d(y, \bar{y}) \geq \varepsilon\}$. Compactness and lower semicontinuity ensure that the infimum $u = \inf_A I$ is achieved. Since $\bar{y} \notin A$, it must be that $u > 0$. Then, for $n$ large enough, the upper large deviation bound (2.3) implies

$$P\{d(Y_n, \bar{y}) \geq \varepsilon\} \leq e^{-r_n(\inf_A I - u/2)} = e^{-r_n u/2}.$$

Thus, $Y_n \to \bar{y}$ in probability. If, moreover, $r_n$ grows fast enough so that $\sum e^{-cr_n} < \infty \ \forall c > 0$, then the Borel-Cantelli lemma implies that $Y_n \to \bar{y}$ almost surely.

For i.i.d. variables Cramér's theorem should also be understood in relation to the central limit theorem (CLT). Consider the case where $M(\theta)$ is finite in a neighborhood of the origin so that $X$ has finite mean $\bar{x} = E[X]$ and finite variance $\sigma^2$, and $I(x) > 0$ for $x \neq \bar{x}$ (Exercise 2.28). Then for each $\delta > 0$ we have the large deviation bound

$$P\{S_n/n - \bar{x} \geq \delta\} \leq e^{-nI(\bar{x}+\delta)}. \tag{2.17}$$

(Recall (2.10) and Exercise 2.25.)

By contrast, the CLT tells us that small deviations of order $n^{-1/2}$ converge to a limit distribution: for $r \in \mathbb{R}$,

$$P\{S_n/n - \bar{x} \geq rn^{-1/2}\} \underset{n\to\infty}{\longrightarrow} \int_r^\infty \frac{e^{-s^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\, ds.$$

This distinction is sometimes expressed by saying that the CLT describes *fluctuations* as opposed to deviations. There is a significant qualitative difference between Cramér's theorem and the CLT. The CLT is an example of *universality*: the Gaussian limit is valid for all distributions with finite variance. The Cramér rate function $I$ on the other hand depends on the entire distribution. (From convex analysis we will learn that $I$ determines $M$.)

There are also results on *moderate deviations* that fall between large deviations and CLT fluctuations. For example, if $d = 1$ and $M$ is finite in a neighborhood of 0, then for any $\alpha \in (0, 1/2)$

$$n^{-2\alpha} \log P\{|S_n/n - \bar{x}| \geq \delta n^{-1/2+\alpha}\} \underset{n\to\infty}{\longrightarrow} -\frac{\delta^2}{2\sigma^2}.$$

Note that this limit picks the leading exponential factor from the Gaussian. In Chapter 11 we discuss refinements to Cramér's theorem and moderate deviations.