Large Deviation Theory

J.M. Swart

March 17, 2021

Preface

The earliest origins of large deviation theory lie in the work of Boltzmann on entropy in the 1870ies and Cramér's theorem from 1938 [Cra38]. A unifying mathematical formalism was only developed starting with Varadhan's definition of a 'large deviation principle' (LDP) in 1966 [Var66].

Basically, large deviation theory centers around the observation that suitable functions F of large numbers of i.i.d. random variables $(X_1, \ldots, X_n)$ often have the property that

$$\mathbb{P}\big[F(X_1, \ldots, X_n) \in \mathrm{d}x\big] \sim e^{-s_n I(x)} \quad \text{as } n \to \infty, \qquad \text{(LDP)}$$

where s_n are real contants such that $\lim_{n\to\infty} s_n = \infty$ (in most cases simply $s_n = n$). In words, (LDP) says that the probability that $F(X_1, \ldots, X_n)$ takes values near a point x decays exponentially fast, with *speed* s_n, and *rate function* I.

Large deviation theory has two different aspects. On the one hand, there is the question of how to formalize the intuitive formula (LDP). This leads to the already mentioned definition of 'large deviation principles' and involves quite a bit of measure theory and real analysis. The most important basic results of the abstract theory were proved more or less between 1966 and 1991, when O'Brian en Verwaat [OV91] and Puhalskii [Puk91] proved that exponential tightness implies a subsequential LDP. The abstract theory of large deviation principles plays more or less the same role as measure theory in (usual) probability theory.

On the other hand, there is a much richer and much more important side of large deviation theory, which tries to identify rate functions I for various functions F of independent random variables, and study their properties. This part of the theory is as rich as the branch of probability theory that tries to prove limit theorems for functions of large numbers of random variables, and has many relations to the latter.

There exist a number of good books on large deviation theory. The oldest book that I am aware of is the one by Ellis [Ell85], which is still useful for applications of large deviation theory in statistical mechanics and gives a good intuitive feeling for the theory, but lacks some of the standard results. A modern book that gives a statistical mechanics oriented view of large deviations is the book by Rassoul-Agha and Seppäläinen [RS15].

The classical books on the topic are the ones of Deuschel and Stroock [DS89] and especially Dembo and Zeitouni [DZ98], the latter originally published in 1993.

While these are very thorough introductions to the field, they can at places be a bit hard to read due to the technicalities involved. Also, both books came a bit too early to pick the full fruit of the developement of the abstract theory.

A very pleasant book to read as a first introduction to the field is the book by Den Hollander [Hol00], which avoids many of the technicalities in favour of a clear exposition of the intuitive ideas and a rich choice of applications. A disadvantage of this book is that it gives little attention to the abstract theory, which means many results are not proved in their strongest form.

Two modern books on the topic, which each try to stress certain aspects of the theory, are the books by Dupuis and Ellis [DE97] and Puhalskii [Puh01]. These books are very strong on the abstract theory, but, unfortunately, they indulge rather heavily in the introduction of their own terminology and formalism (for example, in [DE97], replacing the large deviation principle by the almost equivalent 'Laplace principle') which makes them somewhat inaccessible, unless read from the beginning to the end. The book by Rassoul-Agha and Seppäläinen [RS15] gives a very readable account of the modern abstract theory.

A difficulty encountered by everyone who tries to teach large deviation theory is that in order to do it properly, one first needs quite a bit of abstract theory, which however is intuitively hard to grasp unless one has seen at least a few examples. I have tried to remedy this by first stating, without proof, a number of motivating examples. In the proofs, I have tried to make optimal use of some of the more modern abstract theory, while sticking with the classical terminology and formulations as much as possible.

Contents

Chapter 0

Some motivating examples

0.1 Cramér's theorem

Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. absolutely integrable (i.e., $\mathbb{E}[|X_1|] < \infty$) real random variables with mean $\rho := \mathbb{E}[X_1]$, and let

$$T_n := \frac{1}{n}\sum_{k=1}^{n} X_k \qquad (n \geq 1).$$

be their *empirical averages*. Then the *weak law of large numbers* states that

$$\mathbb{P}\big[|T_n - \rho| \geq \varepsilon\big] \underset{n\to\infty}{\longrightarrow} 0 \qquad (\varepsilon > 0).$$

In 1938, the Swedish statistician and probabilist Harald Cramér [Cra38] studied the question how fast this probability tends to zero. For laws with sufficiently light tails (as stated in the condition (0.1) below), he arrived at the following conclusion.

Theorem 0.1 (Cramér's theorem) *Assume that*

$$Z(\lambda) := \mathbb{E}[e^{\lambda X_1}] < \infty \qquad (\lambda \in \mathbb{R}). \tag{0.1}$$

Then

$$\begin{aligned} &\text{(i)} \quad \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[T_n \geq y] = -I(y) \qquad (y > \rho), \\ &\text{(ii)} \quad \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[T_n \leq y] = -I(y) \qquad (y < \rho), \end{aligned} \tag{0.2}$$

where I is defined by

$$I(y) := \sup_{\lambda\in\mathbb{R}} \big[\lambda y - \log Z(\lambda)\big] \qquad (y \in \mathbb{R}). \tag{0.3}$$

The function Z in (0.1) is called the *moment generating function* or *cumulant generating function*, and its logarithm is consequently called the *logarithmic moment generating function* (or *logarithmic cumulant generating function* of the law of X_1. In the context of large deviation theory, $\log Z(\lambda)$ is also called the *free energy function*, see [Ell85, Section II.4].

The function I defined in (0.3) is called the *rate function*. In order to see what Cramér's theorem tells us exactly, we need to know some elementary properties of this function. Note that (0.1) implies that $\mathbb{E}[|X_1|^2] < \infty$. To avoid trivial cases, we assume that the X_k are not a.s. constant, i.e., $\mathrm{Var}(X_1) > 0$.

Below, $\mathrm{int}(A)$ denotes the interior of a set A, i.e., the largest open set contained in A. We recall that for any finite measure μ on $\mathbb{R}$, $\mathrm{support}(\mu)$ is the smallest closed set such that μ is concentrated on $\mathrm{support}(\mu)$.

Lemma 0.2 (Properties of the rate function) *Let μ be the law of X_1, let $\rho := \langle\mu\rangle$ and $\sigma^2 := \mathrm{Var}(\mu)$ denote its mean and variance, and assume that $\sigma > 0$. Let $y_- := \inf(\mathrm{support}(\mu))$, $y_+ := \sup(\mathrm{support}(\mu))$. Let I be the function defined in (0.3) and set*

$$\mathcal{D}_I := \{y \in \mathbb{R} : I(y) < \infty\} \quad \textit{and} \quad \mathcal{U}_I := \mathrm{int}(\mathcal{D}_I).$$

Then:

(i) *I is convex.*

(ii) *I is lower semi-continuous.*

(iii) *$0 \leq I(y) \leq \infty$ for all $y \in \mathbb{R}$.*

(iv) *$I(y) = 0$ if and only if $y = \rho$.*

(v) *$\mathcal{U}_I = (y_-, y_+)$.*

(vi) *I is infinitely differentiable on $\mathcal{U}_I$.*

(vii) *$\lim_{y\downarrow y_-} I'(y) = -\infty$ and $\lim_{y\uparrow y_+} I'(y) = \infty$.*

(viii) *$I'' > 0$ on $\mathcal{U}_I$ and $I''(\rho) = 1/\sigma^2$.*

(ix) *If $-\infty < y_-$, then $I(y_-) = -\log\mu(\{y_-\})$, and if $y_+ < \infty$, then $I(y_+) = -\log\mu(\{y_+\})$.*

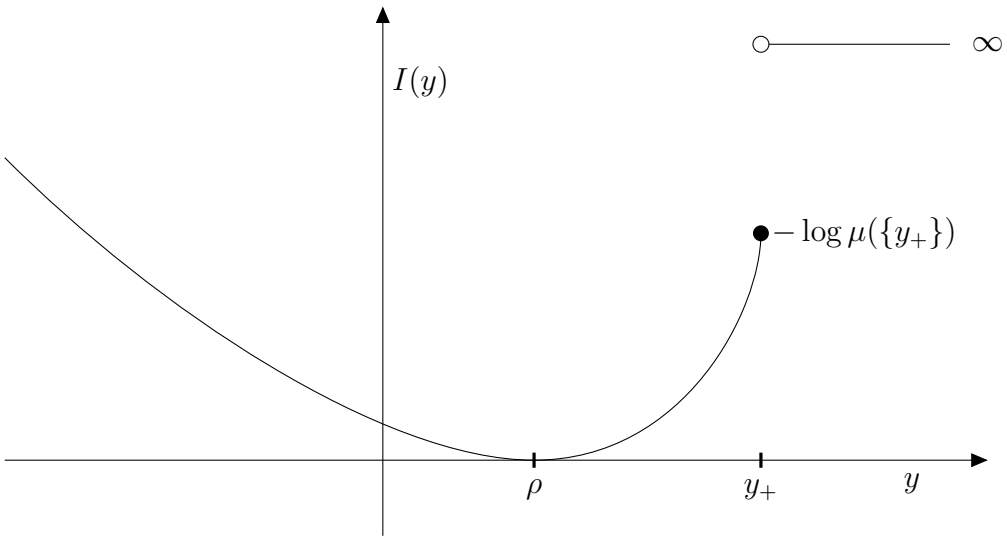

Figure 1: A typical example of a rate function.

See Figure 1 for a picture. Here, if E is any metric space (e.g. $E = \mathbb{R}$), then we say that a function $f : E \to [-\infty, \infty]$ is *lower semi-continuous* if one (and hence both) of the following equivalent conditions are satisfied:

(i) $\liminf_{n\to\infty} f(x_n) \geq f(x)$ whenever $x_n \to x$.

(ii) For each $-\infty \leq a \leq \infty$, the *level set* $\{x \in E : I(x) \leq a\}$ is a closed subset of E.

In view of Lemma 0.2, Theorem 0.1 tells us that the probability that the empirical average T_n deviates by any given constant from its mean decays exponentially fast in n. More precisely, formula (0.2) (i) says that

$$\mathbb{P}[T_n \geq y] = e^{-nI(y) + o(n)} \qquad \text{as} \quad n \to \infty \qquad (y > \rho),$$

were, as usual, $o(n)$ denotes any function such that

$$o(n)/n \to 0 \qquad \text{as} \quad n \to \infty.$$

Note that formulas (0.2) (i) and (ii) only consider one-sided deviations of T_n from its mean ρ. Nevertheless, the limiting behavior of two-sided deviations can easily be derived from Theorem 0.1. Indeed, for any $y_- < \rho < y_+$,

$$\begin{aligned}\mathbb{P}[T_n \leq y_- \text{ or } T_n \geq y_+] &= e^{-nI(y_-) + o(n)} + e^{-nI(y_+) + o(n)} \\ &= e^{-n\min\{I(y_-), I(y_+)\} + o(n)} \qquad \text{as} \quad n \to \infty.\end{aligned}$$

In particular,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[|T_n - \rho| \geq \varepsilon\big] = \min\{I(\rho-\varepsilon), I(\rho+\varepsilon)\} \qquad (\varepsilon > 0).$$

Exercise 0.3 Use Theorem 0.1 and Lemma 0.2 to deduce that, under the assumptions of Theorem 0.1,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[T_n > y\big] = -I_{\rm up}(y) \qquad (y \geq \rho),$$

where $I_{\rm up}$ is the upper semi-continuous modification of I, i.e., $I_{\rm up}(y) = I(y)$ for $y \neq y_-, y_+$ and $I_{\rm up}(y_-) = I_{\rm up}(y_+) := \infty$.

0.2 Moderate deviations

As in the previous section, let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. absolutely integrable real random variables with mean $\rho := \mathbb{E}[|X_1|]$ and assume that (0.1) holds. Let

$$S_n := \sum_{k=1}^n X_k \qquad (n \geq 1).$$

be the *partial sums* of the first n random variables. Then Theorem 0.1 says that

$$\mathbb{P}\big[S_n - \rho n \geq yn\big] = e^{-nI(\rho+y)+o(n)} \quad \text{as } n\to\infty \qquad (y > 0).$$

On the other hand, by the central limit theorem, we know that

$$\mathbb{P}\big[S_n - \rho n \geq y\sqrt{n}\big] \underset{n\to\infty}{\longrightarrow} \Phi(y/\sigma) \qquad (y \in \mathbb{R}),$$

where Φ is the distribution function of the standard normal distribution and

$$\sigma^2 = \mathrm{Var}(X_1),$$

which we assume to be positive. One may wonder what happens at in-between scales, i.e., how does $\mathbb{P}[S_n - \rho n \geq y_n]$ decay to zero if $\sqrt{n} \ll y_n \ll n$? This is the question of *moderate deviations*. We will only consider the case $y_n = yn^\alpha$ with $\frac{1}{2} < \alpha < 1$, even though other timescales (for example in connection with the law of the iterated logarithm) are also interesting.

Theorem 0.4 (Moderate deviations) *Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. absolutely integrable real random variables with mean $\rho := \mathbb{E}[|X_1|]$, variance $\sigma^2 = \mathrm{Var}(X_1) > 0$, and $\mathbb{E}[e^{\lambda X_1}] < \infty$ $(\lambda \in \mathbb{R})$. Then*

$$\lim_{n\to\infty} \frac{1}{n^{2\alpha-1}} \log \mathbb{P}[S_n - \rho n \geq y n^\alpha] = -\tfrac{1}{2\sigma^2} y^2 \qquad (y > 0,\ \tfrac{1}{2} < \alpha < 1). \tag{0.4}$$

Remark Setting $y_n := y n^{\alpha-1}$ and naively applying Cramér's theorem, pretending that y_n is a constant, using Lemma 0.2 (viii), we obtain

$$\begin{aligned} \log \mathbb{P}[S_n - \rho n \geq y n^\alpha] &= \log \mathbb{P}[S_n - \rho n \geq y_n n] \\ &\approx -nI(y_n) \approx -n \tfrac{1}{2\sigma^2} y_n^2 = -\tfrac{1}{2\sigma^2} y^2 n^{2\alpha-1}. \end{aligned}$$

Dividing both sides of this equation by $n^{2\alpha-1}$ yields formula (0.4) (although this derivation is not correct). There does not seem to be a good basic reference for moderate deviations. Some more or less helpful references are [DB81, Led92, Aco02, EL03].

0.3 Relative entropy

Imagine that we throw a dice n times, and keep record of how often each of the possible outcomes $1, \ldots, 6$ comes up. Let $N_n(x)$ be the number of times outcome x has turned up in the first n throws, let $M_n(x) := N_n(x)/x$ be the relative frequency of x, and set

$$\Delta_n := \max_{1\leq x\leq 6} M_n(x) - \min_{1\leq x\leq 6} M_n(x).$$

By the strong law of large numbers, we know that $M_n(x) \to 1/6$ a.s. as $n \to \infty$ for each $x \in \{1, \ldots, 6\}$, and therefore $\mathbb{P}[\Delta_n \geq \varepsilon] \to 0$ as $n \to \infty$ for each $\varepsilon > 0$. It turns out that this convergence happens exponentially fast.

Proposition 0.5 (Deviations from uniformity) *There exists a continuous, strictly increasing function $I : [0,1] \to \mathbb{R}$ with $I(0) = 0$ and $I(1) = \log 6$, such that*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[\Delta_n \geq \varepsilon\big] = -I(\varepsilon) \qquad (0 \leq \varepsilon \leq 1). \tag{0.5}$$

Proposition 0.5 follows from a more general result that was already discovered by the physicist Boltzmann in 1877. A much more general version of this result for random variables that do not need to take values in a finite space was proved by

the Russian mathematician Sanov [San61]. We will restrict ourselves to finite state spaces for the moment. To state the theorem, we first need a few definitions.

Let S be a finite set and let $\mathcal{M}_1(S)$ be the set of all probability measures on S. Since S is finite, we may identify $\mathcal{M}_1(S)$ with the set

$$\mathcal{M}_1(S) := \big\{\pi \in \mathbb{R}^S : \pi(x) \geq 0 \ \forall x \in S, \ \sum_{x\in S} \pi(1) = 1\big\},$$

where $\mathbb{R}^S$ denotes the space of all functions $\pi : S \to \mathbb{R}$. Note that $\mathcal{M}_1(S)$ is a compact, convex subset of the $(|S|-1)$-dimensional space $\{\pi \in \mathbb{R}^S : \sum_{x\in S} \pi(1) = 1\}$.

Let $\mu, \nu \in \mathcal{M}_1(S)$ and assume that $\mu(x) > 0$ for all $x \in S$. Then we define the *relative entropy* of ν with respect to μ by

$$H(\nu|\mu) := \sum_{x\in S} \nu(x) \log \frac{\nu(x)}{\mu(x)} = \sum_{x\in S} \mu(x) \frac{\nu(x)}{\mu(x)} \log \frac{\nu(x)}{\mu(x)},$$

where we use the conventions that $\log(0) := -\infty$ and $0 \cdot \infty := 0$. Note that since $\lim_{z\downarrow 0} z \log z = 0$, the second formula shows that $H(\nu|\mu)$ is continuous in ν. The function $H(\nu|\mu)$ is also known as the *Kullback-Leibler distance* or *divergence*.

Lemma 0.6 (Properties of the relative entropy) *Assume that $\mu \in \mathcal{M}_1(S)$ and assume that $\mu(x) > 0$ for all $x \in S$. Then the function $\nu \mapsto H(\nu|\mu)$ has the following properties.*

(i) $0 \leq H(\nu|\mu) < \infty$ *for all* $\nu \in \mathcal{M}_1(S)$.

(ii) $H(\mu|\mu) = 0$.

(iii) $H(\nu|\mu) > 0$ *for all* $\nu \neq \mu$.

(iv) $\nu \mapsto H(\nu|\mu)$ *is convex and continuous on* $\mathcal{M}_1(S)$.

(v) $\nu \mapsto H(\nu|\mu)$ *is infinitely differentiable on the interior of* $\mathcal{M}_1(S)$.

Assume that $\mu \in \mathcal{M}_1(S)$ satisfies $\mu(x) > 0$ for all $x \in S$ and let $(X_k)_{k\geq 1}$ be an i.i.d. sequence with common law $\mathbb{P}[X_1 = x] = \mu(x)$. As in the example of the dice throws, we let

$$M_n(x) := \frac{1}{n} \sum_{k=1}^{n} 1_{\{X_k = x\}} \qquad (x \in S, \ n \geq 1). \tag{0.6}$$

Note that M_n is a $\mathcal{M}_1(S)$-valued random variable. We call M_n the *empirical distribution*.

Theorem 0.7 (Boltzmann-Sanov) *Let C be a closed subset of $\mathcal{M}_1(S)$ such that C is the closure of its interior. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[M_n \in C] = -\min_{\nu\in C} H(\nu|\mu). \tag{0.7}$$

Note that (0.7) says that

$$\mathbb{P}[M_n \in C] = e^{-nI_C + o(n)} \text{ as } n \to \infty \quad \text{where} \quad I_C = \min_{\nu\in C} H(\nu|\mu). \tag{0.8}$$

This is similar to what we have already seen in Cramér's theorem: if I is the rate function from Theorem 0.1, then $I(y) = \min_{y'\in[y,\infty)} I(y')$ for $y > \rho$ and $I(y) = \min_{y'\in(-\infty,y]} I(y')$ for $y < \rho$. Likewise, as we have seen in (0.1), the probability that $T_n \in (-\infty, y_-] \cup [y_+, \infty)$ decays exponentially with rate $\min_{y'\in(-\infty,y_-]\cup[y_+,\infty)} I(y')$.

The proof of Theorem 0.7 will be delayed till later, but we will show here how Theorem 0.7 implies Proposition 0.5.

Proof of Proposition 0.5 We set $S := \{1, \dots, 6\}$, $\mu(x) := 1/6$ for all $x \in S$, and apply Theorem 0.7. For each $0 \le \varepsilon < 1$, the set

$$C_\varepsilon := \big\{\nu \in \mathcal{M}_1(S) : \max_{x\in S} \nu(x) - \min_{x\in S} \nu(x) \ge \varepsilon\big\}$$

is a closed subset of $\mathcal{M}_1(S)$ that is the closure of its interior. (Note that the last statement fails for $\varepsilon = 1$.) Therefore, Theorem 0.7 implies that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[\Delta_n \ge \varepsilon\big] = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[M_n \in C_\varepsilon\big] = -\min_{\nu\in C_\varepsilon} H(\nu|\mu) =: -I(\varepsilon). \tag{0.9}$$

The fact that I is continuous and satisfies $I(0) = 0$ follows easily from the properties of $H(\nu|\mu)$ listed in Lemma 0.6. To see that I is strictly increasing, fix $0 \le \varepsilon_1 < \varepsilon_2 < 1$. Since $H(\,\cdot\,|\mu)$ is continuous and the C_{ε_2} are compact, we can find a ν_* (not necessarily unique) such that $H(\,\cdot\,|\mu)$ assumes its minimum over C_{ε_2} in ν_*. Now by the fact that $H(\,\cdot\,|\mu)$ is convex and assumes its unique minimum in μ, we see that $\nu' := \frac{\varepsilon_1}{\varepsilon_2}\nu_* + (1 - \frac{\varepsilon_1}{\varepsilon_2})\mu \in C_{\varepsilon_1}$ and therefore $I(\varepsilon_1) \le H(\nu'|\mu) < H(\nu_*|\mu) = I(\varepsilon_2)$.

Finally, by the continuity of $H(\,\cdot\,|\mu)$, we see that

$$I(\varepsilon) \uparrow \min_{\nu\in C_1} H(\nu|\mu) = H(\delta_1|\mu) = \log 6 \qquad \text{as } \varepsilon \uparrow 1.$$

To see that (0.5) also holds for $\varepsilon = 1$ (which does not follow directly from Theorem 0.7 since C_1 is not the closure of its interior), it suffices to note that $\mathbb{P}[\Delta_n = 1] = (\frac{1}{6})^{n-1}$. ∎

Remark 1 It is quite tricky to calculate the function I from Proposition 0.5 explicitly. For ε sufficiently small, it seems that the minimizers of the entropy $H(\,\cdot\,|\mu)$ on the set C_ε are (up to permutations of the coordinates) of the form $\nu(1) = \frac{1}{6} - \frac{1}{2}\varepsilon$, $\nu(2) = \frac{1}{6} + \frac{1}{2}\varepsilon$, and $\nu(3), \ldots, \nu(6) = \frac{1}{6}$. For $\varepsilon > \frac{1}{3}$, this solution is of course no longer well-defined and the minimizer must look differently.

Remark 2 I do not know whether the function I is convex.

0.4 Non-exit probabilities

In this section we move away from the i.i.d. setting and formulate a large deviation result for Markov processes. To keep the technicalities to a minimum, we restrict ourselves to Markov processes with a finite state space. We recall that a continuous-time, time-homogeneous Markov process $X = (X_t)_{t\geq 0}$ taking value in a finite set S is uniquely characterized (in law) by its initial law $\mu(x) := \mathbb{P}[X_0 = x]$ and its *transition probabilities* $P_t(x, y)$. Indeed, X has piecewise constant, right-continuous sample paths and its finite-dimensional distributions are characterized by

$$\mathbb{P}\big[X_{t_1} = x_1, \ldots, X_{t_n} = x_n\big] = \sum_{x_0} \mu(x_0) P_{t_1}(x_0, x_1) P_{t_2 - t_1}(x_1, x_2) \cdots P_{t_n - t_{n-1}}(x_n, x_n)$$

$(t_1 < \cdots < t_n,\ x_1, \ldots, x_n \in S)$. The transition probabilities are continuous in t, have $P_0(x, y) = 1_{\{x=y\}}$ and satisfy the Chapman-Kolmogorov equation

$$\sum_y P_s(x, y) P_t(y, z) = P_{s+t}(x, z) \qquad (s, t \geq 0,\ x, z \in S).$$

As a result, they define a semigroup $(P_t)_{t\geq 0}$ of linear operators $P_t : \mathbb{R}^S \to \mathbb{R}^S$ by

$$P_t f(x) := \sum_y P_t(x, y) f(y) = \mathbb{E}^x[f(X_t)],$$

where $\mathbb{E}^x$ denotes expectation with respect to the law $\mathbb{P}^x$ of the Markov process with initial state $X_0 = x$. One has

$$P_t = e^{Gt} = \sum_{n=0}^{\infty} \frac{1}{n!} G^n t^n,$$

where $G : \mathbb{R}^S \to \mathbb{R}^S$, called the *generator* of the semigroup $(P_t)_{t\geq 0}$, is an operator of the form

$$Gf(x) = \sum_{y:\, y\neq x} r(x,y)\big(f(y) - f(x)\big) \qquad (f \in \mathbb{R}^S,\ x \in S),$$

where $r(x,y)$ $(x, y \in S,\ x \neq y)$ are nonnegative contants. We call $r(x,y)$ the *rate* of jumps from x to y. Indeed, since $P_t = 1 + tG + O(t^2)$ as $t \to 0$, we have that

$$\mathbb{P}^x[X_t = y] = \begin{cases} tr(x,y) + O(t^2) & \text{if } x \neq y, \\ 1 - t\sum_{z:\, z\neq x} r(x,z) + O(t^2) & \text{if } x = y. \end{cases}$$

Let $U \subset S$ be some strict subset of S and assume that $X_0 \in U$ a.s. We will be interested in the probability that X_t stays in U for a long time. Let us say that the transition rates $r(x,y)$ are *irreducible* on U if for each $x, z \in U$ we can find $y_0, \ldots, y_n$ such that $y_0 = x$, $y_n = z$, and $r(y_{k-1}, y_k) > 0$ for each $k = 1, \ldots, n$. Note that this says that it is possible for the Markov process to go from any point in U to any other point in U without leaving U.

Theorem 0.8 (Non-exit probability) *Let X be a Markov process with finite state space S, transition rates $r(x,y)$ $(x, y \in S,\ x \neq y)$, and generator G. Let $U \subset S$ and assume that the transition rates are irreducible on U. Then there exists a function f, unique up to a multiplicative constant, and a constant $\lambda \geq 0$, such that*

(i) $f > 0$ *on* U,

(ii) $f = 0$ *on* $S\backslash U$,

(iii) $Gf(x) = -\lambda f(x) \qquad (x \in U)$.

Moreover, the process X started in any initial law such that $X_0 \in U$ a.s. satisfies

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}\big[X_s \in U\ \forall 0 \leq s \leq t\big] = -\lambda. \tag{0.10}$$

0.5 Outlook

Our aim will be to prove Theorems 0.1, 0.4, 0.7 and 0.8, as well as similar and more general results in a *unified framework*. Therefore, in the next chapter, we will give a formal definition of when a sequence of probability measures satisfies a *large deviation principle* with a given *rate function*. This will allow us to formulate our theorems in a unified framework that is moreover powerful enough to deal

with generalizations such as a multidimensional version of Theorem 0.1 or a generalization of Theorem 0.7 to continuous spaces. We will see that large deviation principles satisfy a number of abstract principles such as the *contraction principle* which we have already used when we derived Proposition 0.5 from Theorem 0.7. Once we have set up the general framework in Chapter 1, in the following chapters, we set out to prove Theorems 0.1, 0.7, and 0.8, as well as similar and more general results, and show how these are related.

Chapter 1

Large deviation principles

1.1 Weak convergence on Polish spaces

Recall that a topological space is a set E equipped with a collection $\mathcal{O}$ of subsets of E that are called *open* sets, such that

(i) If $(O_\gamma)_{\gamma\in\Gamma}$ is any collection of (possibly uncountably many) sets $O_\gamma \in \mathcal{O}$, then $\bigcup_{\gamma\in\Gamma} O_\gamma \in \mathcal{O}$.

(ii) If $O_1, O_2 \in \mathcal{O}$, then $O_1 \cap O_2 \in \mathcal{O}$.

(iii) $\emptyset, E \in \mathcal{O}$.

Any such collection of sets is called a *topology*. It is fairly standard to also assume the *Hausdorff* property

(iv) For each $x_1, x_2 \in E$, $x_1 \neq x_2$ $\exists O_1, O_2 \in \mathcal{O}$ s.t. $O_1 \cap O_2 = \emptyset$, $x_1 \in O_1$, $x_2 \in O_2$.

A sequence of points $x_n \in E$ converges to a limit x in a given topology $\mathcal{O}$ if for each $O \in \mathcal{O}$ such that $x \in O$ there is an n such that $x_m \in O$ for all $m \geq n$. (If the topology is Hausdorff, then such a limit is unique, i.e., $x_n \to x$ and $x_n \to x'$ implies $x = x'$.) A set $C \subset E$ is called *closed* if its complement is open.

Because of property (i) in the definition of a topology, for each $A \subset E$, the union of all open sets contained in A is itself an open set. We call this the *interior* of A, denoted as $\mathrm{int}(A) := \bigcup\{O : U \subset A,\ O \text{ open}\}$. Then clearly $\mathrm{int}(A)$ is the smallest

open set contained in A. Similarly, by taking complements, for each set $A \subset E$ there exists a smallest closed set containing A. We call this the *closure* of A, denoted as $\overline{A} := \bigcap\{C : C \supset A,\ C \text{ closed}\}$. A topological space is called *separable* if there exists a countable set $D \subset E$ such that D is dense in E, where we say that a set $D \subset E$ is *dense* if its closure is E, or equivalently, if every nonempty open subset of E has a nonempty intersection with D.

In particular, if d is a metric on E, and $B_\varepsilon(x) := \{y \in E : d(x,y) < \varepsilon\}$, then

$$\mathcal{O} := \big\{O \subset E : \forall x \in O\ \exists \varepsilon > 0 \text{ s.t. } B_\varepsilon(x) \subset O\big\}$$

defines a Hausdorff topology on E such that convergence $x_n \to x$ in this topology is equivalent to $d(x_n, x) \to 0$. We say that the metric d *generates* the topology $\mathcal{O}$. If for a given topology $\mathcal{O}$ there exists a metric d that generates $\mathcal{O}$, then we say that the topological space $(E, \mathcal{O})$ is *metrizable*.

Recall that a sequence x_n in a metric space (E, d) is a *Cauchy sequence* if for all $\varepsilon > 0$ there is an n such that $d(x_k, x_l) \leq \varepsilon$ for all $k, l \geq n$. A metric space is *complete* if every Cauchy sequence converges.

A *Polish space* is a separable topological space $(E, \mathcal{O})$ such that there exists a metric d on E with the property that (E, d) is complete and d generates $\mathcal{O}$. *Warning:* there may be many different metrics on E that generate the same topology. It may even happen that E is not complete in some of these metrics, and complete in others (in which case E is still Polish). Example: $\mathbb{R}$ is separable and complete in the usual metric $d(x,y) = |x - y|$, and therefore $\mathbb{R}$ is a Polish space. But $d'(x,y) := |\arctan(x) - \arctan(y)|$ is another metric that generates the same topology, while $(\mathbb{R}, d')$ is not complete. (Indeed, the completion of $\mathbb{R}$ w.r.t. the metric d' is $[-\infty, \infty]$.)

On any Polish space $(E, \mathcal{O})$ we let $\mathcal{B}(E)$ denote the Borel-σ-algebra, i.e., the smallest σ-algebra containing the open sets $\mathcal{O}$. We let $B_b(E)$ and $\mathcal{C}_b(E)$ denote the linear spaces of all bounded Borel-measurable and bounded continuous functions $f : E \to \mathbb{R}$, respectively. Then $\mathcal{C}_b(E)$ is complete in the supermumnorm $\|f\|_\infty := \sup_{x \in E} |f(x)|$, i.e., $(\mathcal{C}_b(E), \|\cdot\|_\infty)$ is a Banach space [Dud02, Theorem 2.4.9]. We let $\mathcal{M}(E)$ denote the space of all finite measures on $(E, \mathcal{B}(E))$ and write $\mathcal{M}_1(E)$ for the space of all probability measures. It is possible to equip $\mathcal{M}(E)$ with a metric d_M such that [EK86, Theorem 3.1.7]

(i) $(\mathcal{M}(E), d_H)$ is a separable complete metric space.

(ii) $d_M(\mu_n, \mu) \to 0$ if and only if $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for all $f \in \mathcal{C}_b(E)$.

The precise choice of d_M (there are several canonical ways to define such a metric) is not important to us. We denote convergence in d_M as $\mu_n \Rightarrow \mu$ and call the associated topology (which is uniquely determined by the requirements above) the *topology of weak convergence*. By property (i), the space $\mathcal{M}(E)$ equipped with the topology of weak convergence is a Polish space.

Proposition 1.1 (Weak convergence) *Let E be a Polish space and let $\mu_n, \mu \in \mathcal{M}(E)$. Then one has $\mu_n \Rightarrow \mu$ if and only if the following two conditions are satisfied.*

(i) $\limsup_{n\to\infty} \mu_n(C) \leq \mu(C) \qquad \forall C$ *closed,*

(ii) $\liminf_{n\to\infty} \mu_n(O) \geq \mu(O) \qquad \forall O$ *open.*

If the μ_n, μ are probability measures, then it suffices to check either (i) or (ii).

Before we give the proof of Proposition 1.1, we need a few preliminaries. Recall the definition of lower semi-continuity from Section 0.1. Upper semi-continuity is defined similarly: a function $f : E \to [-\infty, \infty)$ is upper semi-continuous if and only if $-f$ is lower semi-continuous. We set $\overline{\mathbb{R}} := [-\infty, \infty]$ and define

$$\begin{aligned}
\mathcal{U}(E) &:= \big\{f : E \to \overline{\mathbb{R}} \,:\, f \text{ upper semi-continuous}\big\},\\
\mathcal{U}_b(E) &:= \big\{f \in \mathcal{U}(E) : \sup_{x\in E} |f(x)| < \infty\big\},\\
\mathcal{U}_+(E) &:= \big\{f \in \mathcal{U}(E) : f \geq 0\big\},
\end{aligned}$$

and $\mathcal{U}_{b,+}(E) := \mathcal{U}_b(E) \cap \mathcal{U}_+(E)$. We define $\mathcal{L}(E), \mathcal{L}_b(E), \mathcal{L}_+(E), \mathcal{L}_{b,+}(E)$ respectively $\mathcal{C}(E), \mathcal{C}_b(E), \mathcal{C}_+(E), \mathcal{C}_{b,+}(E)$ similarly, with upper semi-continuity replaced by lower semi-continuity and resp. continuity. We will also sometimes use the notation $B(E), B_b(E), B_+(E), B_{b,+}(E)$ for the space of Borel measurable functions $f : E \to \overline{\mathbb{R}}$ and its subspaces of bounded, nonnegative, and bounded nonnegative functions, respectively.

Exercise 1.2 (Topologies of semi-continuity) Let $\mathcal{O}_{\text{up}} := \{[-\infty, a) : -\infty < a \leq \infty\} \cup \{\emptyset, \overline{\mathbb{R}}\}$. Show that $\mathcal{O}_{\text{up}}$ is a topology on $\overline{\mathbb{R}}$ (albeit a non-Hausdorff one!) and that a function $f : E \to \overline{\mathbb{R}}$ is upper semi-continuous if and only if it is continuous with respect to the topology $\mathcal{O}_{\text{up}}$. The topology $\mathcal{O}_{\text{up}}$ is known as the *Scott topology*.

The following lemma lists some elementary properties of upper and lower semi-continuous functions. We set $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$.

Lemma 1.3 *(***Upper and lower semi-continuity)**
(a) $\mathcal{C}(E) = \mathcal{U}(E) \cap \mathcal{L}(E)$.

(b) $f \in \mathcal{U}(E)$ *(resp.* $f \in \mathcal{L}(E)$*) and* $\lambda \geq 0$ *implies* $\lambda f \in \mathcal{U}(E)$ *(resp.* $\lambda f \in \mathcal{L}(E)$*).*

(c) $f, g \in \mathcal{U}(E)$ *(resp.* $f, g \in \mathcal{L}(E)$*) implies* $f + g \in \mathcal{U}(E)$ *(resp.* $f + g \in \mathcal{L}(E)$*).*

(d) $f, g \in \mathcal{U}(E)$ *(resp.* $f, g \in \mathcal{L}(E)$*) implies* $f \vee g \in \mathcal{U}(E)$ *and* $f \wedge g \in \mathcal{U}(E)$ *(resp.* $f \vee g \in \mathcal{L}(E)$ *and* $f \wedge g \in \mathcal{L}(E)$*).*

(e) $f_n \in \mathcal{U}(E)$ *and* $f_n \downarrow f$ *(resp.* $f_n \in \mathcal{L}(E)$ *and* $f_n \uparrow f$*) implies* $f \in \mathcal{U}(E)$ *(resp.* $f \in \mathcal{L}(E)$*).*

(f) *An upper (resp. lower) semi-continuous function assumes its maximum (minimum) over a compact set.*

Proof Part (a) is obvious from the fact that if $x_n \to x$, then $f(x_n) \to f(x)$ if and only if $\limsup_n f(x_n) \leq f(x)$ and $\liminf_n f(x_n) \geq f(x)$. Since f is lower semi-continuous iff $-f$ is upper semi-continuous, it suffices to prove parts (b)–(f) for upper semi-continuous functions. Parts (b) and (d) follow easily from the fact that f is upper semi-continuous if and only if $\{x : f(x) \geq a\}$ is closed for each $a \in \overline{\mathbb{R}}$, which is equivalent to $\{x : f(x) < a\}$ being open for each $a \in \overline{\mathbb{R}}$. Indeed, $f \in \mathcal{U}(E)$ implies that $\{x : \lambda f(x) < a\} = \{x : f(x) < \lambda^{-1} a\}$ is open for each $a \in \mathbb{R}$, $\lambda > 0$, hence $\lambda f \in \mathcal{U}(E)$ for each $\lambda > 0$, while obviously also $0 \cdot f \in \mathcal{U}(E)$. Likewise, $f, g \in \mathcal{U}(E)$ implies that $\{x : f(x) \vee g(x) < a\} = \{x : f(x) < a\} \cap \{x : g(x) < a\}$ is open for each $a \in \overline{\mathbb{R}}$ hence $f \vee g \in \mathcal{U}(E)$ and similarly $\{x : f(x) \wedge g(x) < a\} = \{x : f(x) < a\} \cup \{x : g(x) < a\}$ is open implying that $f \wedge g \in \mathcal{U}(E)$. Part (e) is proved in a similar way: since $\{x : f_n(x) < a\} \uparrow \{x : f(x) < a\}$, we conclude that the latter set is open for all $a \in \overline{\mathbb{R}}$ hence $f \in \mathcal{U}(E)$. Part (c) follows by observing that $\limsup_{n\to\infty}(f(x_n) + g(x_n)) \leq \limsup_{n\to\infty} f(x_n) + \limsup_{m\to\infty} g(x_m) \leq f(x) + g(x)$ for all $x_n \to x$. To prove part (f), finally let f be upper semi-continuous, K compact, and choose $a_n \uparrow \sup_{x\in K} f(x)$. Then $A_n := \{x \in K : f(x) \geq a_n\}$ is a decreasing sequence of nonempty compact sets, hence (by [Eng89, Corollary 3.1.5]) there exists some $x \in \bigcap_n A_n$ and f assumes its maximum in x. ∎

We say that an upper or lower semi-continuous function is *simple* if it assumes only finitely many values.

Lemma 1.4 (Approximation with simple functions) *For each* $f \in \mathcal{U}(E)$ *there exists simple* $f_n \in \mathcal{U}(E)$ *such that* $f_n \downarrow f$*. Analogue statements hold for*

$\mathcal{U}_b(E)$, $\mathcal{U}_+(E)$ and $\mathcal{U}_{b,+}(E)$. Likewise, lower semi-continuous functions can be approximated from below with simple lower semi-continuous functions.

Proof Let $r_- := \inf_{x\in E} f(x)$ and $r_+ := \sup_{x\in E} f(x)$. Let $\mathcal{D} \subset (r_-, r_+)$ be countable and dense and let Δ_n be finite sets such that $\Delta_n \uparrow \mathcal{D}$. Let $\Delta_n = \{a_0, \ldots, a_{m(n)}\}$ with $a_0 < \cdots < a_{m(n)}$ and set

$$f_n(x) := \begin{cases} a_0 & \text{if } f(x) < a_0, \\ a_k & \text{if } a_{k-1} \leq f(x) < a_k \qquad (k = 1, \ldots, m(n)), \\ r_+ & \text{if } a_{m(n)} \leq f(x). \end{cases}$$

Then the f_n are upper semi-continuous, simple, and $f_n \downarrow f$. If $f \in \mathcal{U}_b(E), \mathcal{U}_+(E)$ or $\mathcal{U}_{b,+}(E)$ then also the f_n are in these spaces. The same arguments applied to $-f$ yield the statements for lower semi-continuous functions. ∎

For any set $A \subset E$ and $x \in E$, we let

$$d(x, A) := \inf\{d(x, y) : y \in A\}$$

denote the distance from x to A. Recall that $\overline{A}$ denotes the closure of A.

Lemma 1.5 (Distance to a set) *For each $A \subset E$, the function $x \mapsto d(x, A)$ is continuous and satisfies $d(x, A) = 0$ if and only if $x \in \overline{A}$.*

Proof See [Eng89, Theorem 4.1.10 and Corollary 4.1.11]. ∎

Lemma 1.6 (Approximation of indicator functions) *For each closed $C \subset E$ there exist continuous $f_n : E \to [0, 1]$ such that $f_n \downarrow 1_C$. Likewise, for each open $O \subset E$ there exist continuous $f_n : E \to [0, 1]$ such that $f_n \uparrow 1_C$.*

Proof Set $f_n(x) := (1 - nd(x, C)) \vee 0$ resp. $f_n(x) := nd(x, E\backslash O) \wedge 1$. ∎

Proof of Proposition 1.1 Let $\mu_n, \mu \in \mathcal{M}(E)$ and define the 'good sets'

$$\mathcal{G}_{\text{up}} := \big\{f \in \mathcal{U}_{b,+}(E) : \limsup_{n\to\infty} \int f \mathrm{d}\mu_n \leq \int f \mathrm{d}\mu\big\},$$
$$\mathcal{G}_{\text{low}} := \big\{f \in \mathcal{L}_{b,+}(E) : \liminf_{n\to\infty} \int f \mathrm{d}\mu_n \geq \int f \mathrm{d}\mu\big\}$$

We claim that

(a) $f \in \mathcal{G}_{\text{up}}$ (resp. $f \in \mathcal{G}_{\text{low}}$), $\lambda \geq 0$ implies $\lambda f \in \mathcal{G}_{\text{up}}$ (resp. $\lambda f \in \mathcal{G}_{\text{low}}$).

(b) $f, g \in \mathcal{G}_{\text{up}}$ (resp. $f, g \in \mathcal{G}_{\text{low}}$) implies $f + g \in \mathcal{G}_{\text{up}}$ (resp. $f + g \in \mathcal{G}_{\text{low}}$).

(c) $f_n \in \mathcal{G}_{\text{up}}$ and $f_n \downarrow f$ (resp. $f_n \in \mathcal{G}_{\text{low}}$ and $f_n \uparrow f$) implies $f \in \mathcal{G}_{\text{up}}$ (resp. $f \in \mathcal{G}_{\text{low}}$).

The statements (a) and (b) are easy. To prove (c), let $f_n \in \mathcal{G}_{\text{up}}$, $f_n \downarrow f$. Then, for each k,

$$\limsup_{n\to\infty} \int f \mathrm{d}\mu_n \leq \limsup_{n\to\infty} \int f_k \mathrm{d}\mu_n \leq \int f_k \mathrm{d}\mu.$$

Since $\int f_k \mathrm{d}\mu \downarrow \int f \mathrm{d}\mu$, the claim follows. An analogue argument works for functions in $\mathcal{G}_{\text{low}}$.

We now show that $\mu_n \Rightarrow \mu$ implies the conditions (i) and (ii). Indeed, by Lemma 1.6, for each closed $C \subset E$ we can find continuous $f_k : E \to [0, 1]$ such that $f_k \downarrow 1_C$. Then $f_k \in \mathcal{G}_{\text{up}}$ by the fact that $\mu_n \Rightarrow \mu$ and therefore, by our claim (c) above, it follows that $1_C \in \mathcal{G}_{\text{up}}$, which proves condition (i). The proof of condition (ii) is similar.

Conversely, if condition (i) is satisfied, then by our claims (a) and (b) above, every simple nonnegative bounded upper semi-continuous function is in $\mathcal{G}_{\text{up}}$, hence by Lemma 1.4 and claim (c), $\mathcal{U}_{b,+}(E) \subset \mathcal{G}_{\text{up}}$. Similarly, condition (ii) implies that $\mathcal{L}_{b,+}(E) \subset \mathcal{G}_{\text{low}}$. In particular, this implies that for every $f \in \mathcal{C}_{b,+}(E) = \mathcal{U}_{b,+}(E) \cap \mathcal{L}_{b,+}(E)$, $\lim_{n\to\infty} \int f \mathrm{d}\mu_n = \int f \mathrm{d}\mu$, which by linearity implies that $\mu_n \Rightarrow \mu$.

If the μ_n, μ are probability measures, then conditions (i) and (ii) are equivalent, by taking complements. ∎

1.2 Large deviation principles

A subset K of a topological space $(E, \mathcal{O})$ is called *compact* if every open covering of K has a finite subcovering, i.e., if $\bigcup_{\gamma\in\Gamma} O_\gamma \supset K$ implies that there exist finitely many $O_{\gamma_1}, \ldots, O_{\gamma_n}$ with $\bigcup_{k=1}^n O_{\gamma_1} \supset K$. If $(E, \mathcal{O})$ is metrizable, then this is equivalent to the statement that every sequence $x_n \in K$ has a subsequence $x_{n(m)}$ that converges to a limit $x \in K$ [Eng89, Theorem 4.1.17]. If $(E, \mathcal{O})$ is Hausdorff, then each compact subset of E is closed.

Let E be a Polish space. We say that a function $f : E \to \overline{\mathbb{R}}$ has *compact level sets* if

$$\{x \in E : f(x) \leq a\} \text{ is compact for all } a \in \mathbb{R}.$$

Note that since compact sets are closed, this is (a bit) stronger than the statement that f is lower semi-continuous. We say that I is a *good rate function* if I has compact level sets, $-\infty < I(x)$ for all $x \in E$, and $I(x) < \infty$ for at least one $x \in E$. We observe that:

- A good rate function assumes its minimum on closed sets.

To see this, let C be closed. The statement is trivial if $\inf_{x\in C} I(x) = \infty$. Otherwise, we can choose $\inf_{x\in C} I(x) < a < \infty$. Then the set $K := \{x \in C : I(x) \leq a\}$ is compact and hence by Lemma 1.3 (f), there is an $y \in K$ such that $I(y) = \inf_{x\in C} I(x)$. In particular, applying this to $C = E$, we see that good rate functions and bounded from below.

Recall that $B_b(E)$ denotes the space of all bounded Borel-measurable real functions on E. If μ is a finite measure on $(E, \mathcal{B}(E))$ and $p \geq 1$ is a real constant, then we define the L^p-norm associated with μ by

$$\|f\|_{p,\mu} := \Big(\int \mathrm{d}\mu |f|^p\Big)^{1/p} \qquad (f \in B_b(E)).$$

Likewise, if I is a good rate function, then we can define a sort of 'weighted supremumnorm' by

$$\|f\|_{\infty,I} := \sup_{x\in E} e^{-I(x)}|f(x)| \qquad (f \in B_b(E)). \tag{1.1}$$

Note that $\|f\|_{\infty,I} < \infty$ by the boundedness of f and the fact that I is bounded from below. It is easy to check that $\|\cdot\|_{\infty,I}$ is a *seminorm* , i.e.,

- $\|\lambda f\|_{\infty,I} = |\lambda|\, \|f\|_{\infty,I}$,
- $\|f + g\|_{\infty,I} \leq \|f\|_{\infty,I} + \|g\|_{\infty,I}$.

If $I < \infty$ then $\|\cdot\|_{\infty,I}$ is moreover a norm, i.e.,

- $\|f\|_{\infty,I} = 0$ implies $f = 0$.

Note that what we have just called L^p-norm is in fact only a seminorm, since $\|f\|_{p,\mu} = 0$ only implies that $f = 0$ a.e. w.r.t. μ. (This is usually resolved by looking at equivalence classes of a.e. equal functions, but we won't need this here.)

(Large deviation principle) Let s_n be positive constants converging to ∞, let μ_n be finite measures on E, and let I be a good rate function on E. We say that the μ_n satisfy the *large deviation principle* (LDP) with *speed* (also called *rate*) s_n and *rate function* I if

$$\lim_{n\to\infty} \|f\|_{s_n,\mu_n} = \|f\|_{\infty,I} \qquad \big(f \in \mathcal{C}_{b,+}(E)\big). \tag{1.2}$$

While this definition may look a bit strange at this point, the next proposition looks already much more similar to things we have seen in Chapter 0.

Proposition 1.7 (Large Deviation Principle) *A sequence of finite measures μ_n satisfies the large deviation principle with speed s_n and rate function I if and only if the following two conditions are satisfied.*

(i) $\displaystyle\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C) \leq -\inf_{x\in C} I(x) \qquad \forall C$ *closed,*

(ii) $\displaystyle\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(O) \geq -\inf_{x\in O} I(x) \qquad \forall O$ *open.*

Remark 1 Recall that $\overline{A}$ and $\mathrm{int}(A)$ denote the closure and interior of a set $A \subset E$, respectively. Since for any measurable set A, one has $\mu_n(A) \leq \mu_n(\overline{A})$ and $\mu_n(A) \geq \mu_n(\mathrm{int}(A))$, conditions (i) and (ii) of Proposition 1.7 are equivalent to

(i)' $\displaystyle\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) \leq -\inf_{x\in \overline{A}} I(x),$

(ii)' $\displaystyle\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) \geq -\inf_{x\in \mathrm{int}(A)} I(x),$

for all $A \in \mathcal{B}(E)$. We say that a set $A \in \mathcal{B}(E)$ is *I-continuous* if

$$\inf_{x\in\mathrm{int}(A)} I(x) = \inf_{x\in\overline{A}} I(x)$$

It is now easy to see that if μ_n satisfy the large deviation principle with speed s_n and good rate function I, then

$$\lim_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) = -\inf_{x\in A} I(x)$$

for each I-continuous set A. For example, if I is continuous and $\overline{A} = \overline{\mathrm{int}(A)}$, then A is I-continuous. This is the reason, for example, why in our formulation of the Boltzmann-Sanov Theorem 0.7 we looked at sets that are the closure of their interior.

Remark 2 The two conditions of Proposition 1.7 are the traditional definition of a large deviation principle. Moreover, large deviation principles are often only defined for the special case that the speed s_n equals n. However, as the example of moderate deviations (Theorem 0.4) showed, it is sometimes convenient to allow more general speeds. Also parts of the abstract theory (in particular, connected to the concept of exponential tightness) are more easy to formulate if one allows general speeds. As we will see, allowing more general speeds will not cause any technical complications so this generality comes basically 'for free'.

To prepare for the proof of Proposition 1.7, we start with some preliminary lemmas.

Lemma 1.8 (Properties of the generalized supremumnorm) *Let I be a good rate function and let $\|\cdot\|_{\infty,I}$ be defined as in (1.1). Then*

(a) $\|f \vee g\|_{\infty,I} = \|f\|_{\infty,I} \vee \|g\|_{\infty,I}$ $\forall f, g \in B_{b,+}(E)$.

(b) $\|f_n\|_{\infty,I} \uparrow \|f\|_{\infty,I}$ $\forall f_n \in B_{b,+}(E)$, $f_n \uparrow f$.

(c) $\|f_n\|_{\infty,I} \downarrow \|f\|_{\infty,I}$ $\forall f_n \in \mathcal{U}_{b,+}(E)$, $f_n \downarrow f$.

Proof Property (a) follows by writing

$$\begin{aligned}\|f \vee g\|_{\infty,I} &= \sup_{x\in E} e^{-I(x)}(f(x) \vee g(x)) \\ &= \big(\sup_{x\in E} e^{-I(x)} f(x)\big) \vee \big(\sup_{y\in E} e^{-I(x)} g(y)\big) = \|f\|_{\infty,I} \vee \|g\|_{\infty,I}\end{aligned}$$

To prove (b), we start by observing that the $\|f_n\|_{\infty,I}$ form an increasing sequence and $\|f_n\|_{\infty,I} \leq \|f\|_{\infty,I}$ for each n. Moreover, for any $\varepsilon > 0$ we can find $y \in E$ such that $e^{-I(y)} f(y) \geq \sup_{x\in E} e^{-I(x)} f(x) - \varepsilon$, hence $\liminf_n \|f_n\|_{\infty,I} \geq \lim_n e^{-I(y)} f_n(y) = e^{-I(y)} f(y) \geq \|f\|_{\infty,I} - \varepsilon$. Since $\varepsilon > 0$ is arbitrary, this proves the claim.

To prove also (c), we start by observing that the $\|f_n\|_{\infty,I}$ form a decreasing sequence and $\|f_n\|_{\infty,I} \geq \|f\|_{\infty,I}$ for each n. Since the f_n are upper semi-continuous and I is lower semi-continuous, the functions $e^{-I} f_n$ are upper semi-continuous. Since the f_n are bounded and I has compact level sets, the sets $\{x : e^{-I(x)} f_n(x) \geq a\}$ are compact for each $a > 0$. In particular, for each $a > \sup_{x\in E} e^{-I(x)} f(x)$, the

sets $\{x : e^{-I(x)} f_n(x) \geq a\}$ are compact and decrease to the empty set, hence $\{x : e^{-I(x)} f_n(x) \geq a\} = \emptyset$ for n sufficiently large, which shows that $\limsup_n \|f_n\|_{\infty,I} \leq a$. ∎

Lemma 1.9 (Good sets) *Let $\mu_n \in \mathcal{M}(E)$, $s_n \to \infty$, and let I be a good rate function. Define the 'good sets'*

$$\mathcal{G}_{\rm up} := \big\{f \in \mathcal{U}_{b,+}(E) : \limsup_{n\to\infty} \|f\|_{s_n,\mu_n} \leq \|f\|_{\infty,I}\big\},$$
$$\mathcal{G}_{\rm low} := \big\{f \in \mathcal{L}_{b,+}(E) : \liminf_{n\to\infty} \|f\|_{s_n,\mu_n} \geq \|f\|_{\infty,I}\big\}.$$

Then

(a) *$f \in \mathcal{G}_{\rm up}$ (resp. $f \in \mathcal{G}_{\rm low}$), $\lambda \geq 0$ implies $\lambda f \in \mathcal{G}_{\rm up}$ (resp. $\lambda f \in \mathcal{G}_{\rm low}$).*

(b) *$f, g \in \mathcal{G}_{\rm up}$ (resp. $f, g \in \mathcal{G}_{\rm low}$) implies $f \vee g \in \mathcal{G}_{\rm up}$ (resp. $f \vee g \in \mathcal{G}_{\rm low}$).*

(c) *$f_n \in \mathcal{G}_{\rm up}$ and $f_n \downarrow f$ (resp. $f_n \in \mathcal{G}_{\rm low}$ and $f_n \uparrow f$) implies $f \in \mathcal{G}_{\rm up}$ (resp. $f \in \mathcal{G}_{\rm low}$).*

The proof of Lemma 1.9 makes use of the following elementary lemma.

Lemma 1.10 (The strongest growth wins) *For any $0 \leq a_n, b_n \leq \infty$ and $s_n \to \infty$, one has*

$$\limsup_{n\to\infty} \big(a_n^{s_n} + b_n^{s_n}\big)^{1/s_n} = \big(\limsup_{n\to\infty} a_n\big) \vee \big(\limsup_{n\to\infty} b_n\big). \tag{1.3}$$

Moreover, for any $0 \leq c_n, d_n \leq \infty$ and $s_n \to \infty$,

$$\limsup_{n\to\infty} \frac{1}{s_n} \log(c_n + d_n) = \big(\limsup_{n\to\infty} \frac{1}{s_n} \log c_n\big) \vee \big(\limsup_{n\to\infty} \frac{1}{s_n} \log d_n\big). \tag{1.4}$$

Proof To see this, set $a_\infty := \limsup_{n\to\infty} a_n$ and $b_\infty := \limsup_{n\to\infty} b_n$. Then, for each $\varepsilon > 0$, we can find an m such that $a_n \leq a_\infty + \varepsilon$ and $b_n \leq b_\infty + \varepsilon$ for all $n \geq m$. It follows that

$$\limsup_{n\to\infty} \big(a_n^{s_n} + b_n^{s_n}\big)^{1/s_n} \leq \lim_{n\to\infty} \big((a_\infty + \varepsilon)^{s_n} + (b_\infty + \varepsilon)^{s_n}\big)^{1/s_n} = (a_\infty + \varepsilon) \vee (b_\infty + \varepsilon).$$

Since $\varepsilon > 0$ is arbitrary, this shows that $\limsup_{n\to\infty} \big(a_n^{s_n} + b_n^{s_n}\big)^{1/s_n} \leq a_\infty \vee b_\infty$. Since $a_n, b_n \leq \big(a_n^{s_n} + b_n^{s_n}\big)^{1/s_n}$, the other inequality is trivial. This completes the proof of (1.3).

We claim that (1.4) is just (1.3) in another guise. Indeed, setting $a_n := c_n^{1/s_n}$ and $b_n := d_n^{1/s_n}$ we see, using (1.3), that

$$\begin{aligned} e^{\limsup_{n\to\infty} \frac{1}{s_n}\log(c_n+d_n)} &= \limsup_{n\to\infty} (a_n^{s_n} + d_n^{s_n})^{1/s_n} \\ &= \big(\limsup_{n\to\infty} a_n\big) \vee \big(\limsup_{n\to\infty} b_n\big) \\ &= e^{\big(\limsup_{n\to\infty} \frac{1}{s_n}\log(c_n)\big) \vee \big(\limsup_{n\to\infty} \frac{1}{s_n}\log(d_n)\big)}. \end{aligned}$$

∎

Proof of Lemma 1.9 Part (a) follows from the fact that for any seminorm $\|\lambda f\| = \lambda\|f\|$ ($\lambda > 0$). To prove part (b), assume that $f, g \in \mathcal{G}_{\mathrm{up}}$. Then, by (1.3),

$$\begin{aligned} &\limsup_{n\to\infty} \|f \vee g\|_{s_n,\mu_n} \\ &\quad = \limsup_{n\to\infty} \Big(\int_{\{x:f(x)\geq g(x)\}} f(x)^{s_n}\mu_n(\mathrm{d}x) + \int_{\{x:f(x)<g(x)\}} g(x)^{s_n}\mu_n(\mathrm{d}x)\Big)^{1/s_n} \\ &\quad \leq \limsup_{n\to\infty} \big(\|f\|_{s_n,\mu_n}^{s_n} + \|g\|_{s_n,\mu_n}^{s_n}\big)^{1/s_n} \leq \|f\|_{\infty,I} \vee \|g\|_{\infty,I} = \|f \vee g\|_{\infty,I}, \end{aligned} \tag{1.5}$$

proving that $f \vee g \in \mathcal{G}_{\mathrm{up}}$. Similarly, but easier, if $f, g \in \mathcal{G}_{\mathrm{low}}$, then

$$\begin{aligned} \liminf_{n\to\infty} \|f \vee g\|_{s_n,\mu_n} &\geq \big(\liminf_{n\to\infty} \|f\|_{s_n,\mu_n}\big) \vee \big(\liminf_{n\to\infty} \|g\|_{s_n,\mu_n}\big) \\ &\geq \|f\|_{\infty,I} \vee \|g\|_{\infty,I} = \|f \vee g\|_{\infty,I}, \end{aligned}$$

which proves that $f \vee g \in \mathcal{G}_{\mathrm{low}}$.

To prove part (c), finally, assume that $f_k \in \mathcal{G}_{\mathrm{up}}$ satisfy $f_k \downarrow f$. Then f is upper semi-continuous and

$$\limsup_{n\to\infty} \|f\|_{s_n,\mu_n} \leq \limsup_{n\to\infty} \|f_k\|_{s_n,\mu_n} \leq \|f_k\|_{\infty,I}$$

for each k. Since $\|f_k\|_{\infty,I} \downarrow \|f\|_{\infty,I}$, by Lemma 1.8 (c), we conclude that $f \in \mathcal{G}_{\mathrm{up}}$. The proof for $f_k \in \mathcal{G}_{\mathrm{low}}$ is similar, using Lemma 1.8 (b). ∎

Proof of Proposition 1.7 If the μ_n satisfy the large deviation principe with speed s_n and rate function I, then by Lemmas 1.6 and 1.9 (c), $1_C \in \mathcal{G}_{\mathrm{up}}$ for each closed $C \subset E$ and $1_O \in \mathcal{G}_{\mathrm{up}}$ for each open $O \subset E$, which shows that conditions (i) and (ii) are satisfied. Conversely, if conditions (i) and (ii) are satisfied, then by Lemma 1.9 (a) and (b),

$$\mathcal{G}_{\mathrm{up}} \supset \{f \in \mathcal{U}_{b,+}(E) : f \text{ simple}\} \quad \text{and} \quad \mathcal{G}_{\mathrm{low}} \supset \{f \in \mathcal{L}_{b,+}(E) : f \text{ simple}\}.$$

By Lemmas 1.4 and 1.9 (c), it follows that $\mathcal{G}_{\rm up} = \mathcal{U}_{b,+}(E)$ and $\mathcal{G}_{\rm low} = \mathcal{L}_{b,+}(E)$. In particular, this proves that

$$\lim_{n\to\infty} \|f\|_{s_n,\mu_n} = \|f\|_{\infty,I} \qquad \forall f \in \mathcal{C}_{b,+}(E),$$

which shows that the μ_n satisfy the large deviation principe with speed s_n and rate function I. ∎

Exercise 1.11 (Robustness of LDP) Let $(X_k)_{k\geq 1}$ be i.i.d. random variables with $\mathbb{P}[X_k = 0] = \mathbb{P}[X_k = 1] = \frac{1}{2}$, let $Z(\lambda) := \mathbb{E}[e^{\lambda X_1}]$ ($\lambda \in \mathbb{R}$) and let $I : \mathbb{R} \to [0,\infty]$ be defined as in (0.3). Let $\varepsilon_n \downarrow 0$ and set

$$T_n := \frac{1}{n}\sum_{k=1}^n X_k \quad \text{and} \quad T'_n := (1-\varepsilon_n)\frac{1}{n}\sum_{k=1}^n X_k.$$

In Theorem 3.5 below, we will prove that the laws $\mathbb{P}[T_n \in \cdot\,]$ satisfy the large deviation principle with speed n and rate function I. Using this fact, prove that also the laws $\mathbb{P}[T'_n \in \cdot\,]$ satisfy the large deviation principle with speed n and rate function I. Use Lemma 0.2 to conclude that

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}[T'_n \geq y] = -I(y) \quad (\tfrac{1}{2} \leq y < 1),$$

but this formula does *not* hold for $y = 1$.

1.3 Varadhan's lemma

The two conditions of Proposition 1.7 are the traditional definition of the large deviation principle, which is due to Varadhan [Var66]. Our alternative, equivalent definition in terms of convergence of L_p-norms is very similar to the road followed in Puhalskii's book [Puh01]. A very similar definition is also given in [DE97], where this is called a 'Laplace principle' instead of a large deviation principle.

From a purely abstract point of view, our definition is frequently a bit easier to work with. On the other hand, the two conditions of Proposition 1.7 are closer to the usual interpretation of large deviations in terms of exponentially small probabilities. Also, when in some practical situation one wishes to prove a large deviation principle, the two conditions of Proposition 1.7 are often a very natural way to do so. Here, condition (ii) is usually easier to check than condition (i).

Condition (ii) says that certain rare events occur wih at least a certain probability. To prove this, one needs to find one strategy by which a stochastic system can make the desired event happen, with a certain small probability. Condition (i) says that there are no other strategies that yield a higher probability for the same event, which requires one to prove something about all possible ways in which a certain event can happen.

In practically all applications, we will only be interested in the case that the measures μ_n are probability measures and the rate function satisfies $\inf_{x\in E} I(x) = 0$, but being slightly more general comes at virtually no cost.

Varadhan [Var66] was not only the first one who formulated large deviation principles in the generality that is now standard, he also first proved the lemma that is called after him, and that reads as follows.

Lemma 1.12 (Varadhan's lemma) *Let E be a Polish space and let $\mu_n \in \mathcal{M}(E)$ satisfy the large deviation principle with speed s_n and good rate function I. Let $F : E \to \mathbb{R}$ be continuous and assume that $\sup_{x\in E} F(x) < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{s_n} \log \int e^{s_n F} \mathrm{d}\mu_n = \sup_{x\in E}[F(x) - I(x)].$$

Proof Applying the exponential function to both sides of our equation, this says that

$$\lim_{n\to\infty} \Big(\int e^{s_n F} \mathrm{d}\mu_n \Big)^{1/s_n} = \sup_{x\in E} e^{F(x)-I(x)}.$$

Setting $f := e^F$, this is equivalent to

$$\lim_{n\to\infty} \|f\|_{s_n,\mu_n} = \|f\|_{\infty,I},$$

where our asumptions on F translate into $f \in \mathcal{C}_{b,+}(E)$. Thus, Varadhan's lemma is just a trivial reformulation of our definition of a large deviation principle. If we take the traditional definition of a large deviation principle as our starting point, then Varadhan's lemma corresponds to the 'if' part of Proposition 1.7. ∎

As we have just seen, Varadhan's lemma is just the statement that the two conditions of Proposition 1.7 are sufficient for (1.2). The fact that these conditions are also necessary was only proved 24 years later, by Bryc [Bry90].

We conclude this section with a little lemma that says that a sequence of measures satisfying a large deviation principle determines its rate function uniquely.

Lemma 1.13 (Uniqueness of the rate function) *Let E be a Polish space, $\mu_n \in \mathcal{M}(E)$, and let s_n be real constants converging to infinity. Assume that the μ_n satisfy the large deviation principle with speed s_n and good rate function I and also that the μ_n satisfy the large deviation principle with speed s_n and good rate function I'. Then $I = I'$.*

Proof It follows immediately from our definition of the large deviation principle that $\|f\|_{\infty,I} = \|f\|_{\infty,I'}$ for all $f \in \mathcal{C}_{b,+}(E)$. By Lemma 1.6, for each $x \in E$, we can find continuous $f_n : E \to [0,1]$ such that $f_n \downarrow 1_{\{x\}}$. By Lemma 1.8 (c), it follows that

$$e^{-I(x)} = \|1_{\{x\}}\|_{\infty,I} = \lim_{n\to\infty} \|f_n\|_{\infty,I} = \lim_{n\to\infty} \|f_n\|_{\infty,I'} = \|1_{\{x\}}\|_{\infty,I'} = e^{-I'(x)}$$

for each $x \in E$. ∎

1.4 The contraction principle

As we have seen in Propositions 1.1 and 1.7, there is a lot of similarity between weak convergence and the large deviation principle. Elaborating on this analogy, we recall that if X_n is a sequence of random variables, taking values in some Polish space E, whose laws converge weakly to the law of a random variable X, and $\psi : E \to F$ is a continuous function from E into some other Polish space, then the laws of the random variables $\psi(X_n)$ converge weakly to the law of $\psi(X)$. As we will see, an analogue statement holds for sequences of measures satisfying a large deviation principle.

Recall that if X is a random variable taking values in some measurable space $(E, \mathcal{E})$, with law $\mathbb{P}[X \in \cdot\,] = \mu$, and $\psi : E \to F$ is a measurable function from E into some other measurable space $(F, \mathcal{F})$, then the law of $\psi(X)$ is the *image measure*

$$\mu \circ \psi^{-1}(A) \quad (A \in \mathcal{F}), \quad \text{where} \quad \psi^{-1}(A) := \{x \in E : \psi(x) \in A\}$$

is the *inverse image* (or *pre-image*) of A under ψ.

The next result shows that if X_n are random variables whose laws satisfy a large deviation principle, and ψ is a continuous function, then also the laws of the $\psi(X_n)$ satify a large deviation principle. This fact is known a the *contraction principle*.

Note that we have already seen this principle at work when we derived Proposition 0.5 from Theorem 0.7. As is clear from this example, it is in practice not always easy to explicitly calculate the 'image' of a rate function under a continuous map, as defined formally in (1.6) below.

Proposition 1.14 (Contraction principle) *Let E, F be Polish spaces and let $\psi : E \to F$ be continuous. Let μ_n be finite measures on E satisfying a large deviation principle with speed s_n and good rate function I. Then the image measures $\mu \circ \psi^{-1}$ satisfying the large deviation principle with speed s_n and good rate function J given by*

$$J(y) := \inf_{x \in \psi^{-1}(\{y\})} I(x) \qquad (y \in F), \tag{1.6}$$

where $\inf_{x \in \emptyset} I(x) := \infty$.

Proof Recall that a function ψ from one topological space E into another topological space F is continuous if and only if the inverse image under ψ of any open set is open, or equivalently, the inverse image of any closed set is closed (see, e.g., [Eng89, Proposition 1.4.1] or [Kel75, Theorem 3.1]). As a result, condition (i) of Proposition 1.7 implies that

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n \circ \psi^{-1}(C) &\leq - \inf_{x \in \psi^{-1}(C)} I(x) \\ = - \inf_{y \in C} \inf_{x \in \psi^{-1}(\{y\})} I(x) &= - \inf_{y \in C} J(y), \end{aligned} \tag{1.7}$$

where we have used that $\psi^{-1}(C) = \bigcup_{y \in C} \psi^{-1}(y\})$. Condition (ii) of Proposition 1.7 carries over in the same way. We are left with the task of showing that J is a good rate function. Indeed, for each $a \in \mathbb{R}$, we have that

$$\begin{aligned} \{y \in F : J(y) \leq a\} &= \big\{y \in F : \inf_{x \in \psi^{-1}(\{y\})} I(x) \leq a\big\} \\ &= \big\{y \in F : \exists x \in E \text{ s.t. } \psi(x) = y,\ I(x) \leq a\big\} \\ &= \{\psi(x) : x \in E,\ I(x) \leq a\} = \psi(\{x : I(x) \leq a\}), \end{aligned}$$

where in the second equality we have used that I assumes its minimum on the closed set $\psi^{-1}(\{y\})$. Our calculation shows that the level set $\{y \in F : J(y) \leq a\}$ is the image under ψ of the level set $\{x : I(x) \leq a\}$. Since the continuous image of a compact set is compact[Eng89, Theorem 3.1.10],[1] this proves that J has compact level sets. Finally, we observe (compare (1.7)) that $\inf_{y \in F} J(y) = \inf_{x \in \psi^{-1}(F)} I(x) = \inf_{x \in E} I(x) < \infty$, proving that J is a good rate function. ∎

[1]This is a well-known fact that can be found in any book on general topology. It is easy to show by counterexample that the continuous image of a *closed* set needs in general not be closed!

1.5 Exponential tilts

It is not hard to see that if μ_n are measures satisfying a large deviation principle, then we can transform these measures by weighting them with an exponential density, in such a way that the new measures also satisfy a large deviation principle. Recall that if μ is a measure and f is a nonnegative measurable function, then setting

$$f\mu(A) := \int_A f \mathrm{d}\mu$$

defines a new measure $f\mu$ which is μ weighted with the density f.

Lemma 1.15 (Exponential weighting) *Let E be a Polish space and let $\mu_n \in \mathcal{M}(E)$ satisfy the large deviation principle with speed s_n and good rate function I. Let $F : E \to \overline{\mathbb{R}}$ be continuous and assume that $-\infty < \sup_{x\in E} F(x) < \infty$. Then the measures*

$$\tilde{\mu}_n := e^{s_n F}\mu_n$$

satisfy the large deviation principle with speed s_n and good rate function $\tilde{I} := I - F$.

Proof Note that $e^F \in \mathcal{C}_{b,+}(E)$. Therefore, for any $f \in \mathcal{C}_{b,+}(E)$,

$$\begin{aligned}\|f\|_{s_n,\tilde{\mu}_n} &= \int f^{s_n} e^{s_n F} \mathrm{d}\mu_n = \|fe^F\|_{s_n,\mu_n} \\ &\underset{n\to\infty}{\longrightarrow} \|fe^F\|_{\infty,I} = \sup_{x\in E} f(x)e^{F(x)}e^{-I(x)} = \|f\|_{\infty,\tilde{I}}.\end{aligned}$$

Since F is continuous, $I - F$ is lower semi-continuous. Since F is bounded from above, any level set of $I - F$ is contained in some level set of I, and therefore compact. Since F is not identically $-\infty$, finally, $\inf_{x\in I}(I(x) - F(x)) < \infty$, proving that $I - F$ is a good rate function. ∎

Lemma 1.15 is not so useful yet, since in practice we are usually interested in probability measures, while exponential weighting may spoil the normalization. Likewise, we are usually interested in rate functions that are properly 'normalized'. Let us say that a function I is a *normalized rate function* if I is a good rate function and $\inf_{x\in E} I(x) = 0$. Note that if μ_n are probability measures satisfying a large deviation principle with speed s_n and rate function I, then I must be normalized, since E is both open and closed, and therefore by conditions (i) and (ii) of Proposition 1.7

$$-\inf_{x\in E} I(x) = \lim_{n\to\infty} \frac{1}{s_n} \log \mu_n(E) = 0.$$

Lemma 1.16 (Exponential tilting) *Let E be a Polish space and let μ_n be probability measures on E satisfy the large deviation principle with speed s_n and normalized rate function I. Let $F : E \to \overline{\mathbb{R}}$ be continuous and assume that $-\infty < \sup_{x\in E} F(x) < \infty$. Then the measures*

$$\tilde{\mu}_n := \frac{1}{\int e^{s_n F} \mathrm{d}\mu_n} e^{s_n F} \mu_n$$

satisfy the large deviation principle with speed s_n and normalized rate function $\tilde{I}(x) := I(x) - F(x) - \inf_{y\in E}(I(y) - F(y))$.

Proof Since $e^F \in \mathcal{C}_{b,+}(E)$, much in the same way as in the proof of the previous lemma, we see that

$$\begin{aligned}
\|f\|_{s_n,\tilde{\mu}_n} &= \Big(\frac{1}{\int e^{s_n F} \mathrm{d}\mu_n} \int f^{s_n} e^{s_n F} \mathrm{d}\mu_n\Big)^{1/s_n} = \frac{\|fe^F\|_{s_n,\mu_n}}{\|e^F\|_{s_n,\mu_n}} \\
&\underset{n\to\infty}{\longrightarrow} \frac{\|fe^F\|_{\infty,I}}{\|e^F\|_{\infty,I}} = \frac{\sup_{x\in E} f(x)e^{F(x)}e^{-I(x)}}{\sup_{x\in E} e^{F(x)}e^{-I(x)}} \\
&= e^{-\inf_{y\in E}(I(y)-F(y))} \sup_{x\in E} f(x)e^{-(I(x)-F(x))} = \|f\|_{\infty,\tilde{I}}.
\end{aligned}$$

The fact that $\tilde{I}$ is a good rate function follows from the same arguments as in the proof of the previous lemma, and $\tilde{I}$ is obviously normalized. ∎

1.6 Robustness

Often, when one wishes to prove that the laws $\mathbb{P}[X_n \in \cdot\,]$ of some random variables X_n satisfy a large deviation principle with a given speed and rate function, it is convenient to replace the random variables X_n by some other random variables Y_n that are 'sufficiently close', so that the large deviation principle for the laws $\mathbb{P}[Y_n \in \cdot\,]$ implies the LDP for $\mathbb{P}[X_n \in \cdot\,]$. The next result (which we copy from [DE97, Thm 1.3.3]) gives sufficient conditions for this to be allowed.

Proposition 1.17 (Superexponential approximation) *Let $(X_n)_{n\geq 1}$, $(Y_n)_{n\geq 1}$ be random variables taking values in a Polish space E and assume that the laws $\mathbb{P}[Y_n \in \cdot\,]$ satisfy a large deviation principle with speed s_n and rate function I. Let d be any metric generating the topology on E, and assume that*

$$\lim_{n\to\infty} \frac{1}{s_n} \log \mathbb{P}\big[d(X_n, Y_n) \geq \varepsilon\big] = -\infty \qquad (\varepsilon > 0). \tag{1.8}$$

Then the laws $\mathbb{P}[X_n \in \cdot\,]$ satisfy the large deviation principle with speed s_n and rate function I.

Remark If (1.8) holds, then we say that the random variables X_n and Y_n are *exponentially close*. Note that condition (1.8) is in particular satisfied if for each $\varepsilon > 0$ there is an N such that $d(X_n, Y_n) < \varepsilon$ a.s. for all $n \geq N$. We can even allow for $d(X_n, Y_n) \geq \varepsilon$ with a small probability, but in this case these probabilities must tend to zero faster than any exponential.

Proof of Proposition 1.17 Let $C \subset E$ be closed and let $C_\varepsilon := \{x \in E : d(x, C) \leq \varepsilon\}$. Then

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{s_n} \log \mathbb{P}[X_n \in C] \\
&\quad \leq \limsup_{n\to\infty} \frac{1}{s_n} \log \big(\mathbb{P}[Y_n \in C_\varepsilon,\ d(X_n, Y_n) \leq \varepsilon] + \mathbb{P}[d(X_n, Y_n) > \varepsilon]\big) \\
&\quad \leq \limsup_{n\to\infty} \frac{1}{s_n} \log \mathbb{P}[Y_n \in C_\varepsilon] = -\inf_{x\in C_\varepsilon} I(x) \underset{\varepsilon\downarrow 0}{\longrightarrow} -\inf_{x\in C} I(x),
\end{aligned}$$

where we have used (1.4) and in the last step we have applied (the logarithmic version of) Lemma 1.8 (c). Similarly, if $O \subset E$ is open and $O_\varepsilon := \{x \in E : d(x, E\backslash O) > \varepsilon\}$, then

$$\liminf_{n\to\infty} \frac{1}{s_n} \log \mathbb{P}[X_n \in O] \geq \liminf_{n\to\infty} \frac{1}{s_n} \log \mathbb{P}[Y_n \in O_\varepsilon,\ d(X_n, Y_n) \leq \varepsilon].$$

The large deviations lower bound is trivial if $\inf_{x\in O} I(x) = \infty$, so without loss of generality we may assume that $\inf_{x\in O} I(x) < \infty$. Since $\inf_{x\in O_\varepsilon} I(x) \downarrow \inf_{x\in O} I(x)$, it follows that for ε sufficiently small, also $\inf_{x\in O_\varepsilon} I(x) < \infty$. By the fact that the Y_n satisfy the large deviation lower bound and by (1.8),

$$\begin{aligned}
&\mathbb{P}[Y_n \in O_\varepsilon,\ d(X_n, Y_n) \leq \varepsilon] \geq \mathbb{P}[Y_n \in O_\varepsilon] - \mathbb{P}[d(X_n, Y_n) > \varepsilon] \\
&\quad \geq e^{-s_n \inf_{x\in O_\varepsilon} I(x) + o(s_n)} - e^{-s_n/o(s_n)}
\end{aligned}$$

as $n \to \infty$, where $o(s_n)$ is the usual small 'o' notation, i.e., $o(s_n)$ denotes any term such that $o(s_n)/s_n \to 0$. It follows that

$$\liminf_{n\to\infty} \frac{1}{s_n} \log \mathbb{P}[Y_n \in O_\varepsilon,\ d(X_n, Y_n) \leq \varepsilon] \geq -\inf_{x\in O_\varepsilon} I(x) \underset{\varepsilon\downarrow 0}{\longrightarrow} -\inf_{x\in O} I(x),$$

which proves the the large deviation lower bound for the X_n. ∎

Proposition 1.17 shows that large deviation principles are 'robust', in a certain sense, with repect to small perturbations. The next result is of a similar nature:

we will prove that weighting measures with densities does not affect a large deviation principle, as long as these densities do not grow exponentially fast. This complements the case of exponentialy growing densities which has been treated in Section 1.5.

Lemma 1.18 (Subexponential weighting) *Let E be a Polish space and let $\mu_n \in \mathcal{M}(E)$ satisfy the large deviation principle with speed s_n and good rate function I. Let $F_n : E \to \mathbb{R}$ be measurable and assume that $\lim_{n\to\infty} \|F_n\|_\infty = 0$, where $\|F_n\|_\infty := \sup_{x\in E} |F_n(x)|$. Then the measures*

$$\tilde{\mu}_n := e^{s_n F_n} \mu_n$$

satisfy the large deviation principle with speed s_n and rate function I.

Proof We check the large deviations upper and lower bound from Proposition 1.7. For any closed set $C \subset E$, by the fact that the μ_n satisfy the large deviation principle, we have

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{s_n} \log \tilde{\mu}_n(C) &= \limsup_{n\to\infty} \frac{1}{s_n} \log \int_C \mu_n(\mathrm{d}x) e^{s_n F_n(x)} \\ &\leq \limsup_{n\to\infty} \frac{1}{s_n} \log \big(e^{s_n \|F_n\|} \mu_n(C)\big) = \limsup_{n\to\infty} \big(\|F_n\| + \frac{1}{s_n} \log \mu_n(C)\big),\end{aligned}$$

which equals $-\inf_{x\in C} I(x)$. Similarly, for any open $O \subset E$, we have

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{s_n} \log \tilde{\mu}_n(O) &= \liminf_{n\to\infty} \frac{1}{s_n} \log \int_O \mu_n(\mathrm{d}x) e^{s_n F_n(x)} \\ &\geq \liminf_{n\to\infty} \frac{1}{s_n} \log \big(e^{-s_n \|F_n\|} \mu_n(O)\big) = \liminf_{n\to\infty} \big(-\|F_n\| + \frac{1}{s_n} \log \mu_n(O)\big),\end{aligned}$$

which yields $-\inf_{x\in O} I(x)$, as required. ∎

1.7 Tightness

In Sections 1.1 and 1.2, we have stressed the similarity between weak convergence of measures and large deviation principles. In the remainder of this chapter, we will pursue this idea further. In the present section, we recall the concept of tightness and Prohorov's theorem. In particular, we will see that any tight sequence of probability measures on a Polish space has a weakly convergent subsequence. In

the next sections (to be precise, in Theorem 1.25), we will prove an analogue of this result, which says that every exponentially tight sequence of probability measures on a Polish space has a subsequence that satisfies a large deviation principle.

A set A is called *relatively compact* if its closure $\overline{A}$ is compact. The next result is known as Prohorov's theorem (see, e.g., [Ste87, Theorems III.3.3 and III.3.4] or [Bil99, Theorems 5.1 and 5.2]).

Proposition 1.19 (Prohorov) *Let E be a Polish space and let $\mathcal{M}_1(E)$ be the space of probability measures on $(E, \mathcal{B}(E))$, equipped with the topology of weak convergence. Then a subset $\mathcal{C} \subset \mathcal{M}_1(E)$ is relatively compact if and only if $\mathcal{C}$ is* tight*, i.e.,*

$$\forall \varepsilon > 0 \ \exists K \subset E \text{ compact, s.t. } \sup_{\mu \in \mathcal{C}} \mu(E \backslash K) \leq \varepsilon.$$

Note that since sets consisting of a single point are always compact, Proposition 1.19 implies that every probability measure (and therefore also every finite measure) on a Polish space E has the property that for all $\varepsilon > 0$ there exists a compact K such that $\mu(E \backslash K) \leq \varepsilon$. This result, that is sometimes known as *Ulam's theorem*, is in itself already nontrivial, since Polish spaces need in general not be locally compact.

By definition, a set of functions $\mathcal{D} \subset \mathcal{C}_b(E)$ is called *distribution determining* if for any $\mu, \nu \in \mathcal{M}_1(E)$,

$$\int f \mathrm{d}\mu = \int f \mathrm{d}\nu \quad \forall f \in \mathcal{D} \quad \text{implies} \quad \mu = \nu.$$

We say that a sequence of probability measures $(\mu_n)_{n \geq 1}$ is *tight* if the set $\{\mu_n : n \geq 1\}$ is tight, i.e., $\forall \varepsilon > 0$ there exists a compact K such that $\sup_n \mu_n(E \backslash K) \leq \varepsilon$. By Prohorov's theorem, each tight sequence of probability measures has a convergent subsequence. This fact is often applied as in the following lemma.

Lemma 1.20 (Tight sequences) *Let E be a Polish space and let μ_n, μ be probability measures on E. Assume that $\mathcal{D} \subset \mathcal{C}_b(E)$ is distribution determining. Then one has $\mu_n \Rightarrow \mu$ if and only if the following two conditions are satisfied:*

(i) *The sequence $(\mu_n)_{n \geq 1}$ is tight.*

(ii) *$\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for all $f \in \mathcal{D}$.*

The proof of Lemma 1.20 uses a simple fact from general topology. Recall that $(x'_n)_{n\in\mathbb{N}}$ is a subsequence of $(x_n)_{n\in\mathbb{N}}$ if there exist $n(m) \to \infty$ such that $x'_m = x_{n(m)}$ $(m \in \mathbb{N})$.

Lemma 1.21 (Convergence along subsequences) *Let E be a topological space and let $x_n, x \in E$. Assume that each subsequence (x'_n) of (x_n) contains a further subsequence (x''_n) such that $x''_n \to x$. Then $x_n \to x$.*

Proof Assume that $x_n \not\to x$. Then there exists an open set $O \ni x$ such that $x_n \notin O$ for infinitely many n, hence there exists a subsequence (x'_n) such that $x'_n \notin O$ for all n. But then no subsequence (x''_n) of (x'_n) can converge to x, contradicting our assumption. ∎

Proof of Lemma 1.20 In any metrizable space, if $(x_n)_{n\geq 1}$ is a convergent sequence, then $\{x_n : n \geq 1\}$ is relatively compact. Thus, by Prohorov's theorem, conditions (i) and (ii) are clearly necessary.

To prove the sufficiency of conditions (i) and (ii) we apply Lemma 1.21. By (i) and Prohorov's theorem, each subsequence (μ'_n) of (μ_n) contains a further subsequence (μ''_n) that converges weakly to some limit μ''. By (ii) $\int f \mathrm{d}\mu'' = \int f \mathrm{d}\mu$ for all $f \in \mathcal{D}$ so $\mu'' = \mu$ and hence by Lemma 1.21 wd conclude that the original sequence (μ_n) converges weakly to μ. ∎

1.8 LDP's on compact spaces

Our aim is to prove an analogue of Lemma 1.20 for large deviation principles. To prepare for this, in the present section, we will study large deviation principles on compact spaces. The results in this section will also shed some light on some elements of the theory that have up to now not been very well motivated, such as why rate functions are lower semi-continuous.

It is well-known that a compact metrizable space is separable, and complete in any metric that generates the topology. In particular, all compact metrizabe spaces are Polish. Note that if E is a compact metrizable space, then $\mathcal{C}(E) = \mathcal{C}_b(E)$, i.e., continuous functions are automatically bounded. We equip $\mathcal{C}(E)$ with the supremumnorm $\|\cdot\|_\infty$, under which it is a separable Banach space.[2] Below, $|f|$ denotes the absolute value of a function, i.e., the function $x \mapsto |f(x)|$.

[2]The separability of $\mathcal{C}(E)$ is an easy consequence of the Stone-Weierstrass theorem [Dud02,

Proposition 1.22 (Generalized supremumnorms) *Let E be a compact metrizable space and let $\Lambda : \mathcal{C}(E) \to [0,\infty)$ be a function such that*

(i) *Λ is a seminorm.*

(ii) *$\Lambda(f) = \Lambda(|f|)$ for all $f \in \mathcal{C}(E)$.*

(iii) *$\Lambda(f) \leq \Lambda(g)$ for all $f, g \in \mathcal{C}_+(E)$, $f \leq g$.*

(iv) *$\Lambda(f \vee g) = \Lambda(f) \vee \Lambda(g)$ for all $f, g \in \mathcal{C}_+(E)$.*

Then

(a) *$\Lambda : \mathcal{C}(E) \to [0,\infty)$ is continuous w.r.t. the supremumnorm.*

Moreover, there exists a function $I : E \to (-\infty,\infty]$ such that

(b) *$\Lambda(f_n) \downarrow e^{-I(x)}$ for any $f_n \in \mathcal{C}_+(E)$ s.t. $f_n \downarrow 1_{\{x\}}$.*

(c) *I is lower semi-continuous.*

(d) *$\Lambda(f) = \sup_{x\in E} e^{-I(x)}|f(x)| \qquad \big(f \in \mathcal{C}(E)\big)$.*

Proof To prove part (a), we observe that by (ii), (iii) and (i)

$$\Lambda(f) = \Lambda(|f|) \leq \Lambda(\|f\|_\infty \cdot 1) = \|f\|_\infty \Lambda(1),$$

where $1 \in \mathcal{C}(E)$ denotes the function that is identically one. Using again that Λ is a seminorm, we see that

$$\big|\Lambda(f) - \Lambda(g)\big| \leq \Lambda(f-g) \leq \Lambda(1)\|f-g\|_\infty.$$

This shows that Λ is continuous w.r.t. the supremumnorm.

Next, define $I : E \to (-\infty,\infty]$ (or equivalently $e^{-I} : E \to [0,\infty)$) by

$$e^{-I(x)} := \inf\{\Lambda(f) : f \in \mathcal{C}_+(E),\ f(x) = 1\} \qquad (x \in E).$$

Thm 2.4.11]. Let $\mathcal{D} \subset E$ be dense and let $\mathcal{A} := \{\phi_{n,x} : x \in \mathcal{D},\ n \geq 1\}$, where $\phi_{\delta,x}(y) := 0 \vee (1 - nd(x,y))$. Let $\mathcal{B}$ be the set containing the function that is identically 1 and all functions of the form $f_1 \cdots f_m$ with $m \geq 1$ and $f_1, \ldots, f_m \in \mathcal{A}$. Let $\mathcal{C}$ be the linear span of $\mathcal{B}$ and let $\mathcal{C}'$ be the set of functions of the form $a_1 f_1 + \cdots + a_m f_m$ with $m \geq 1$, $a_1, \ldots, a_m \in \mathbb{Q}$ and $f_1, \ldots, f_m \in \mathcal{B}$. Then $\mathcal{C}$ is an algebra that separates points, hence by the Stone-Weierstrass theorem, $\mathcal{C}$ is dense in $\mathcal{C}(E)$. Since $\mathcal{C}'$ is dense in $\mathcal{C}'$ and $\mathcal{C}'$ is countable, it follows that $\mathcal{C}(E)$ is separable.

We claim that this function satisfies the properties (b)–(d). Indeed, if $f_n \in \mathcal{C}_+(E)$ satisfy $f_n \downarrow 1_{\{x\}}$ for some $x \in E$, then the $\Lambda(f_n)$ decrease to a limit by the monotonicity of Λ. Since

$$\Lambda(f_n) \geq \Lambda(f_n/f_n(x)) \geq \inf\{\Lambda(f) : f \in \mathcal{C}_+(E),\ f(x) = 1\} = e^{-I(x)}$$

we see that this limit is larger or equal than $e^{-I(x)}$. To prove the other inequality, we note that by the definition of I, for each $\varepsilon > 0$ we can choose $f \in \mathcal{C}_+(E)$ with $f(x) = 1$ and $\Lambda(f) \leq e^{-I(x)} + \varepsilon$. We claim that there exists an n such that $f_n < (1+\varepsilon)f$. Indeed, this follows from the fact that the the sets $C_n := \{y \in E : f_n(y) \geq (1+\varepsilon)f(y)\}$ are compact sets decreasing to the empty set, hence $C_n = \emptyset$ for some n [Eng89, Corollary 3.1.5]. As a result, we obtain that $\Lambda(f_n) \leq (1+\varepsilon)\Lambda(f) \leq (1+\varepsilon)(e^{-I(x)} + \varepsilon)$. Since $\varepsilon > 0$ is arbitrary, this completes the proof of property (b).

To prove part (c), consider the functions

$$\phi_{\delta,y}(x) := 0 \vee (1 - d(y,x)/\delta) \qquad (x, y \in E,\ \delta > 0).$$

Observe that $\phi_{\delta,y}(y) = 1$ and $\phi_{\delta,y} = 0$ on $B_\delta(y)^{\mathrm{c}}$, and recall from Lemma 1.5 that $\phi_{\delta,y} : E \to [0,1]$ is continuous. Since

$$\|\phi_{\delta,y} - \phi_{\delta,z}\|_\infty \leq \delta^{-1} \sup_{x \in E} |d(x,y) - d(x,z)| \leq \delta^{-1} d(y,z),$$

we see that the map $x \mapsto \phi_{\delta,x}$ is continuous w.r.t. the supremumnorm. By part (a), it follows that for each $\delta > 0$, the functions

$$x \mapsto \Lambda(\phi_{\delta,x})$$

are continuous. Since by part (b) these functions decrease to e^{-I} as $\delta \downarrow 0$, we conclude that e^{-I} is upper semi-continuous or equivalently I is lower semi-continuous.

To prove part (d), by assumption (ii), it suffices to consider the case that $f \in \mathcal{C}_+(E)$. We start by observing that

$$e^{-I(x)} \leq \Lambda(f) \qquad \forall x \in E,\ f \in \mathcal{C}_+(E),\ f(x) = 1,$$

hence, more generally, for any $x \in E$ and $f \in \mathcal{C}_+(E)$ such that $f(x) > 0$,

$$e^{-I(x)} \leq \Lambda(f/f(x)) = \Lambda(f)/f(x),$$

which implies that

$$e^{-I(x)} f(x) \leq \Lambda(f) \qquad \forall x \in E,\ f \in \mathcal{C}_+(E),$$

and therefore

$$\Lambda(f) \geq \sup_{x\in E} e^{-I(x)} f(x) \qquad \big(f \in \mathcal{C}_+(E)\big).$$

To prove the other inequality, we claim that for each $f \in \mathcal{C}_+(E)$ and $\delta > 0$ we can find some $x \in E$ and $g \in \mathcal{C}_+(E)$ supported on $B_{2\delta}(x)$ such that $f \geq g$ and $\Lambda(f) = \Lambda(g)$. To see this, consider the functions

$$\psi_{\delta,y}(x) := 0 \vee (1 - d(B_\delta(y), x)/\delta) \qquad (x, y \in E,\ \delta > 0).$$

Note that $\psi_{\delta,y} : E \to [0, 1]$ is continuous and equals one on $B_\delta(y)$ and zero on $B_{2\delta}(y)^{\mathrm{c}}$. Since E is compact, for each $\delta > 0$ we can find a finite set $\Delta \subset E$ such that $\bigcup_{x\in\Delta} B_\delta(x) = E$. By property (iv), it follows that

$$\Lambda(f) = \Lambda\big(\bigvee_{x\in\Delta} \psi_{\delta,x} f\big) = \bigvee_{x\in\Delta} \Lambda(\psi_{\delta,x} f).$$

In particular, we may choose some x such that $\Lambda(f) = \Lambda(\psi_{\delta,x} f)$. Continuing this process, we can find $x_k \in E$ and $f_k \in \mathcal{C}_+(E)$ supported on $B_{1/k}(x_k)$ such that $f \geq f_1 \geq f_2$ and $\Lambda(f) = \Lambda(f_1) = \Lambda(f_2) = \cdots$. It is not hard to see that the f_n decrease to zero except possibly in one point x, i.e.,

$$f_n \downarrow c1_{\{x\}}$$

for some $0 \leq c \leq f(x)$ and $x \in E$. By part (b), it follows that $\Lambda(f) = \Lambda(f_n) \downarrow ce^{-I(x)} \leq f(x)e^{-I(x)}$. This completes the proof of part (d). ∎

Recall the definition of a normalized rate function from page 32. The following proposition prepares for Theorem 1.25 below.

Proposition 1.23 (LDP along a subsequence) *Let E be a compact metrizable space, let μ_n be probability measures on E and let s_n be positive constants converging to infinity. Then there exists $n(m) \to \infty$ and a normalized rate function I such that the $\mu_{n(m)}$ satisfy the large deviation principle with speed $s_{n(m)}$ and rate function I.*

Proof Since $\mathcal{C}(E)$, the space of continuous real functions on E, equipped with the supremumnorm, is a separable Banach space, we can choose a countable dense subset $\mathcal{D} = \{f_k : k \geq 1\} \subset \mathcal{C}(E)$. Using the fact that the μ_n are probability measures, we see that

$$\|f\|_{s_n,\mu_n} = \Big(\int |f|^{s_n} \mathrm{d}\mu_n\Big)^{1/s_n} \leq \big(\|f\|_\infty^{s_n}\big)^{1/s_n} = \|f\|_\infty \qquad \big(f \in \mathcal{C}(\overline{E})\big).$$

By Tychonoff's theorem, the product space

$$X := \bigtimes_{k=1}^{\infty} \big[0, \|f_k\|_\infty\big],$$

equipped with the product topology is compact. Therefore, we can find $n(m) \to \infty$ such that

$$\big(\|f\|_{s_{n(m)},\mu_{n(m)}}\big)_{k\geq 1}$$

converges as $m \to \infty$ to some limit in X. In other words, this says that we can find a subsequence such that

$$\lim_{m\to\infty} \|f\|_{s_{n(m)},\mu_{n(m)}} =: \Lambda(f)$$

exists for each $f \in \mathcal{D}$. We claim that this implies that for the same subsequence, this limit exists in fact for all $f \in \mathcal{C}(E)$. To prove this, we observe that for each $f, g \in \mathcal{C}(E)$,

$$\big|\|f\|_{s_n,\mu_n} - \|g\|_{s_n,\mu_n}\big| \leq \|f-g\|_{s_n,\mu_n} \leq \|f-g\|_\infty.$$

Letting $n(m) \to \infty$ we see that also

$$|\Lambda(f) - \Lambda(g)| \leq \|f-g\|_\infty \tag{1.9}$$

for all $f, g \in \mathcal{D}$. Since a uniformly continuous function from one metric space into another can uniquely be extended to a continuous function from the completion of one space to the completion of the other, we see from (1.9) that Λ can be uniquely extended to a function $\Lambda : \mathcal{C}(E) \to [0,\infty)$ such that (1.9) holds for all $f, g \in \mathcal{C}(E)$. Moreover, if $f \in \mathcal{C}(E)$ is arbitrary and $f_i \in \mathcal{D}$ satisfy $\|f - f_i\|_\infty \to 0$, then

$$\begin{aligned}
&\big|\|f\|_{s_{n(m)},\mu_{n(m)}} - \Lambda(f)\big| \\
&\quad \leq \big|\|f\|_{s_{n(m)},\mu_{n(m)}} - \|f_i\|_{s_{n(m)},\mu_{n(m)}}\big| + \big|\|f_i\|_{s_{n(m)},\mu_{n(m)}} - \Lambda(f_i)\big| + \big|\Lambda(f_i) - \Lambda(f)\big| \\
&\qquad \leq \big|\|f_i\|_{s_{n(m)},\mu_{n(m)}} - \Lambda(f_i)\big| + 2\|f - f_i\|_\infty,
\end{aligned}$$

hence

$$\limsup_{m\to\infty} \big|\|f\|_{s_{n(m)},\mu_{n(m)}} - \Lambda(f)\big| \leq 2\|f - f_i\|_\infty$$

for each i, which proves that $\|f\|_{s_{n(m)},\mu_{n(m)}} \to \Lambda(f)$.

Our next aim is to show that the function $\Lambda : \mathcal{C}(E) \to [0,\infty)$ satisfies properties (i)–(iv) of Proposition 1.22. Properties (i)–(iii) are satisfied by the norms $\|\cdot\|_{s_{n(m)},\mu_{n(m)}}$ for each m, so by taking the limit $m \to \infty$ we see that also Λ has

these properties. To prove also property (iv), we use an argument similar to the one used in the proof of Lemma 1.9 (b). Arguing as in (1.5), we obtain

$$\begin{aligned}\Lambda(f\vee g) &= \lim_{m\to\infty}\|f\vee g\|_{s_{n(m)},\mu_{n(m)}} \leq \limsup_{m\to\infty}\big(\|f\|^{s_{n(m)}}_{s_{n(m)},\mu_{n(m)}} + \|g\|^{s_{n(m)}}_{s_{n(m)},\mu_{n(m)}}\big)^{1/s_{n(m)}}\\ &= \big(\limsup_{m\to\infty}\|f\|_{s_{n(m)},\mu_{n(m)}}\big)\vee\big(\limsup_{m\to\infty}\|g\|_{s_{n(m)},\mu_{n(m)}}\big) = \Lambda(f)\vee\Lambda(g),\end{aligned}$$

where we have used (1.3). Since $f,g\leq f\vee g$, it follows from property (iii) that moreover $\Lambda(f)\vee\Lambda(g)\leq\Lambda(f\vee g)$, completing the proof of property (iv).

By Proposition 1.22, it follows that there exists a lower semi-continuous function $I:E\to(-\infty,\infty]$ such that

$$\Lambda(f) = \sup_{x\in E} e^{-I(x)}|f(x)| \qquad \big(f\in\mathcal{C}(E)\big).$$

Since E is compact, I has compact level sets, i.e., I is a good rate function, hence the $\mu_{n(m)}$ satisfy the large deviation principle with speed $s_{n(m)}$ and rate function I. Since the $\mu_{n(m)}$ are probability measures, it follows that I is normalized. ∎

1.9 Exponential tightness

We wish to generalize Proposition 1.23 to spaces that are not compact. To do this, we need a condition whose role is similar to that of tightness in the theory of weak convergence.

Let μ_n be a sequence of finite measures on a Polish space E and let s_n be positive contants, converging to infinity. We say that the μ_n are *exponentially tight* with speed s_n if

$$\forall M\in\mathbb{R}\ \exists K\subset E \text{ compact, s.t. } \limsup_{n\to\infty}\frac{1}{s_n}\log\mu_n(E\backslash K)\leq -M.$$

Letting $A^{\mathrm{c}} := E\backslash A$ denote the complement of a set $A\subset E$, it is easy to check that exponential tightness is equivalent to the statement that

$$\forall\varepsilon>0\ \exists K\subset E \text{ compact, s.t. } \limsup_{n\to\infty}\|1_{K^{\mathrm{c}}}\|_{s_n,\mu_n}\leq\varepsilon.$$

The next lemma says that exponential tightness is a necessary condition for a large deviation principle.

Lemma 1.24 (LDP implies exponential tightness) *Let E be a Polish space and let μ_n be finite measures on E satisfying a large deviation principle with speed s_n and good rate function I. Then the μ_n are exponentially tight with speed s_n.*

Proof This proof of this statement is more tricky than might be expected at first sight. We follow [DZ93, Excercise 4.1.10]. If the space E is locally compact, then an easier proof is possible, see [DZ93, 1.2.19].

Let d be a metric generating the topology on E such that (E,d) is complete, and let $B_r(x)$ denote the open ball (w.r.t. this metric) of radius r around x. Since E is separable, we can choose a dense sequence $(x_k)_{k\geq 1}$ in E. Then, for every $\delta > 0$, the open sets $O_{\delta,m} := \bigcup_{k=1}^m B_\delta(x_k)$ increase to E. By Lemma 1.8 (c), $\|1_{O^{\rm c}_{\delta,m}}\|_{\infty,I} \downarrow 0$. Thus, for each $\varepsilon, \delta > 0$ we can choose an $m \geq 1$ such that

$$\limsup_{n\to\infty} \|1_{O^{\rm c}_{\delta,m}}\|_{s_n,\mu_n} \leq \|1_{O^{\rm c}_{\delta,m}}\|_{\infty,I} \leq \varepsilon.$$

In particular, for any $\varepsilon > 0$, we can choose $(m_k)_{k\geq 1}$ such that

$$\limsup_{n\to\infty} \|1_{O^{\rm c}_{1/k,m_k}}\|_{s_n,\mu_n} \leq 2^{-k}\varepsilon \qquad (k \geq 1).$$

It follows that

$$\begin{aligned}\limsup_{n\to\infty} \|1_{\bigcup_{k=1}^\infty O^{\rm c}_{1/k,m_k}}\|_{s_n,\mu_n} &\leq \limsup_{n\to\infty} \sum_{k=1}^\infty \|1_{O^{\rm c}_{1/k,m_k}}\|_{s_n,\mu_n} \\ &\leq \sum_{k=1}^\infty \limsup_{n\to\infty} \|1_{O^{\rm c}_{1/k,m_k}}\|_{s_n,\mu_n} \leq \sum_{k=1}^\infty 2^{-k}\varepsilon = \varepsilon.\end{aligned}$$

Here

$$\bigcup_{k=1}^\infty O^{\rm c}_{1/k,m_k} = \Big(\bigcap_{k=1}^\infty O_{1/k,m_k}\Big)^{\rm c} = \Big(\bigcap_{k=1}^\infty \bigcup_{l=1}^{m_k} B_{1/k}(x_l)\Big)^{\rm c}.$$

Let K be the closure of $\bigcap_{k=1}^\infty O_{1/k,m_k}$. We claim that K is compact. Recall that a subset A of a metric space (E,d) is *totally bounded* if for every $\delta > 0$ there exist a finite set $\Delta \subset A$ such that $A \subset \bigcup_{x\in\Delta} B_\delta(x)$. It is well-known [Dud02, Thm 2.3.1] that a subset A of a metric space (E,d) is compact if and only if it is complete and totally bounded. In particular, if (E,d) is complete, then A is compact if and only if A is closed and totally bounded. In light of this, it suffices to show that K is totally bounded. But this is obvious from the fact that $K \subset \bigcup_{l=1}^{m_k} B_{2/k}(x_l)$ for each $k \geq 1$. Since

$$\limsup_{n\to\infty} \|1_{K^{\rm c}}\|_{s_n,\mu_n} \leq \limsup_{n\to\infty} \|1_{(\bigcap_{k=1}^\infty O_{1/k,m_k})^{\rm c}}\|_{s_n,\mu_n} \leq \varepsilon$$

and $\varepsilon > 0$ is arbitrary, this proves the exponential tightness of the μ_n. ∎

The following theorem generalizes Proposition 1.23 to non-compact spaces. This result is due to O'Brian and Verwaat [OV91] and Puhalskii [Puk91]; see also the treatment in Dupuis and Ellis [DE97, Theorem 1.3.7].

Theorem 1.25 (Exponential tightness implies LDP along a subsequence) *Let E be a Polish space, let μ_n be probability measures on E and let s_n be positive constants converging to infinity. Assume that the μ_n are exponentially tight with speed s_n. Then there exist $n(m) \to \infty$ and a normalized rate function I such that the $\mu_{n(m)}$ satisfy the large deviation principle with speed $s_{n(m)}$ and good rate function I.*

We will derive Theorem 1.25 from Proposition 1.23 using compactification techniques. For this, we need to recall some general facts about compactifications of metrizable spaces.

If $(E, \mathcal{O})$ is a topological space (with $\mathcal{O}$ the collection of open subsets of E) and $E' \subset E$ is any subset of E, then E' is also naturally equipped with a topology given by the collection of open subsets $\mathcal{O}' := \{O \cap E' : O \in \mathcal{O}\}$. This topology is called the *induced* topology from E. If $x_n, x \in E'$, then $x_n \to x$ in the induced topology on E' if and only if $x_n \to x$ in E.

If $(E, \mathcal{O})$ is a topological space, then a *compactification* of E is a compact topological space $\overline{E}$ such that E is a dense subset of $\overline{E}$ and the topology on E is the induced topology from $\overline{E}$. If $\overline{E}$ is metrizable, then we say that $\overline{E}$ is a *metrizable compactification* of E. It turns out that each separable metrizable space E has a metrizable compactification [Cho69, Theorem 6.3].

A topological space E is called *locally compact* if for every $x \in E$ there exists an open set O and compact set C such that $x \in O \subset C$. We cite the following proposition from [Eng89, Thms 3.3.8 and 3.3.9].

Proposition 1.26 (Compactification of locally compact spaces) *Let E be a metrizable topological space. Then the following statements are equivalent.*

(i) *E is locally compact and separable.*

(ii) *There exists a metrizable compactification $\overline{E}$ of E such that E is an open subset of $\overline{E}$.*

(iii) *For each metrizable compactification $\overline{E}$ of E, E is an open subset of $\overline{E}$.*

A subset $A \subset E$ of a topological space E is called a *G_δ-set* if A is a countable intersection of open sets (i.e., there exist $O_i \in \mathcal{O}$ such that $A = \bigcap_{i=1}^{\infty} O_i$. The following result can be found in [Bou58, §6 No. 1, Theorem. 1]. See also [Oxt80, Thms 12.1 and 12.3].

Proposition 1.27 (Compactification of Polish spaces) *Let E be a metrizable topological space. Then the following statements are equivalent.*

(i) *E is Polish.*

(ii) *There exists a metrizable compactification $\overline{E}$ of E such that E is a G_δ-subset of $\overline{E}$.*

(iii) *For each metrizable compactification $\overline{E}$ of E, E is a G_δ-subset of $\overline{E}$.*

Moreover, a subset $F \subset E$ of a Polish space E is Polish in the induced topology if and only if F is a G_δ-subset of E.

Lemma 1.28 (Restriction principle) *Let E be a Polish space and let $F \subset E$ be a G_δ-subset of E, equipped with the induced topology. Let $(\mu_n)_{n\geq 1}$ be finite measures on E such that $\mu_n(E\backslash F) = 0$ for all $n \geq 1$, let s_n be positive constants converging to infinity and let I be a good rate function on E such that $I(x) = \infty$ for all $x \in E\backslash F$. Let $\mu_n|_F$ and $I|_F$ denote the restrictions of μ_n and I, respectively, to F. Then $I|_F$ is a good rate function on F and the following statements are equivalent.*

(i) *The μ_n satisfy the large deviation principle with speed s_n and rate function I.*

(ii) *The $\mu_n|_F$ satisfy the large deviation principle with speed s_n and rate function $I|_F$.*

Proof Since the level sets of I are compact in E and contained in F, they are also compact in F, hence $I|_F$ is a good rate function. To complete the proof, by Proposition 1.7, it suffices to show that the large deviations upper and lower bounds for the μ_n and $\mu_n|_F$ are equivalent. A subset of F is open (resp. closed) in the induced topology if and only if it is of the form $O \cap F$ (resp. $C \cap F$) with O an open subset of E (resp. C a closed subset of E). The equivalence of the upper bounds now follows from the observation that for each closed $C \subset E$,

$$\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n\big|_F(C \cap F) = \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C)$$

and

$$\inf_{x\in C} I(x) = \inf_{x\in C\cap F} I\big|_F(x).$$

In the same way, we see that the large deviations lower bounds for the μ_n and $\mu_n|_F$ are equivalent. ∎

Exercise 1.29 (Weak convergence and the induced topology) Let E be a Polish space and let $\overline{E}$ be a metrizable compactification of E. Let d be a metric generating the topology on $\overline{E}$, and denote the restriction of this metric to E also by d. Let $\mathcal{C}_u(E)$ denote the class of functions $f : E \to \mathbb{R}$ that are uniformly continuous w.r.t. the metric d, i.e.,

$$\forall\varepsilon > 0\ \exists\delta > 0 \text{ s.t. } d(x,y) \le \delta \quad \text{implies} \quad |f(x) - f(y)| \le \varepsilon.$$

Let $(\mu_n)_{n\ge1}$ and μ be probability measures on E. Show that the following statements are equivalent:

(i) $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for all $f \in \mathcal{C}_b(E)$,

(ii) $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for all $f \in \mathcal{C}_u(E)$,

(iii) $\mu_n \Rightarrow \mu$ where $\Rightarrow$ denotes weak convergence of probability measures on E,

(iv) $\mu_n \Rightarrow \mu$ where $\Rightarrow$ denotes weak convergence of probability measures on $\overline{E}$.

Hint: Identify $\mathcal{C}_u(E) \cong \mathcal{C}(\overline{E})$ and apply Proposition 1.1.

We note that compactifications are usually not unique, i.e., it is possible to construct many different compactifications of one and the same space E. If E is locally compact (but not compact), however, then we may take $\overline{E}$ such that $\overline{E}\backslash E$ consists one a single point (usually denoted by ∞). This *one-point compactification* is (up to homeomorphisms) unique. For example, the one-point compactification of $[0,\infty)$ is $[0,\infty]$ and the one-point compactification of $\mathbb{R}$ looks like a circle. Another useful compactification of $\mathbb{R}$ is of course $\overline{\mathbb{R}} := [-\infty,\infty]$. To see an example of a compactification of a Polish space that is not locally compact, consider the space $E := \mathcal{M}_1(\mathbb{R})$ of probability measures on $\mathbb{R}$, equipped with the topology of weak convergence. A natural compactification of this space is the space $\overline{E} := \mathcal{M}_1(\overline{\mathbb{R}})$ of probability measures on $\overline{\mathbb{R}}$. Note that $\mathcal{M}_1(\mathbb{R})$ is not an open subset[3] of $\mathcal{M}_1(\overline{\mathbb{R}})$,

[3]Indeed $(1-n^{-1})\delta_0 + n^{-1}\delta_\infty \in \mathcal{M}_1(\overline{\mathbb{R}})\backslash\mathcal{M}_1(\mathbb{R})$ converge to $\delta_0 \in \mathcal{M}_1(\mathbb{R})$ which show that the complement of $\mathcal{M}_1(\mathbb{R})$ is not closed.

which by Proposition 1.26 proves that $\mathcal{M}_1(\mathbb{R})$ is not locally compact. On the other hand, since by Excercise 1.29, $\mathcal{M}_1(\mathbb{R})$ is Polish in the induced topology, we can conclude by Proposition 1.27 that $\mathcal{M}_1(\mathbb{R})$ must be a G_δ-subset $\mathcal{M}_1(\overline{\mathbb{R}})$. (Note that in particular, this is a very quick way of proving that $\mathcal{M}_1(\mathbb{R})$ is a measurable subset of $\mathcal{M}_1(\overline{\mathbb{R}})$.)

Note that in all these examples, though the *topology* on E coincides with the (induced) topology from $\overline{E}$, the *metrics* on E and $\overline{E}$ may be different. Indeed, if d is a metric generating the topology on $\overline{E}$, then E will never be complete in this metric (unless E is compact).

Proof of Theorem 1.25 Let $\overline{E}$ be a metrizable compactification of E. By Proposition 1.23, there exists $n(m) \to \infty$ and a normalized rate function $I : \overline{E} \to [0, \infty]$ such that the $\mu_{n(m)}$ (viewed as probability measures on $\overline{E}$) satisfy the large deviation principle with speed $s_{n(m)}$ and rate function I.

We claim that for each $a < \infty$, the level set $L_a := \{x \in \overline{E} : I(x) \leq a\}$ is a compact subset of E (in the induced topology). To see this, choose $a < b < \infty$. By exponential tightness, there exists a compact $K \subset E$ such that

$$\limsup_{m\to\infty} \frac{1}{s_{n(m)}} \log \mu_{n(m)}(K^{\mathrm{c}}) \leq -b. \tag{1.10}$$

Note that since the identity map from E into $\overline{E}$ is continuous, and the continuous image of a compact set is compact, K is also a compact subset of $\overline{E}$. We claim that $L_a \subset K$. Assume the converse. Then we can find some $x \in L_a \backslash K$ and open subset O of $\overline{E}$ such that $x \in O$ and $O \cap K = \emptyset$. Since the $\mu_{n(m)}$ satisfy the LDP on $\overline{E}$, by Proposition 1.7 (ii),

$$\liminf_{m\to\infty} \frac{1}{s_{n(m)}} \log \mu_{n(m)}(O) \geq -\inf_{x\in O} I(x) \geq -a,$$

contradicting (1.10). This shows that $L_a \subset K$. Since L_a is a closed subset of $\overline{E}$, it follows that L_a is a compact subset of E (in the induced topology). In particular, our arguments show that $I(x) = \infty$ for all $x \in \overline{E} \backslash E$. The statement now follows from the restriction principle (Lemma 1.28) and the fact that the $\mu_{n(m)}$ viewed as probability measures on $\overline{E}$ satisfy the large deviation principle with speed $s_{n(m)}$ and rate function I. ∎

In the next section, we will look at applications of Theorem 1.25. As an appetizer, we conclude the present section by proving two simple lemmas. The argument used in the proof of Lemma 1.21 also applies to large deviation principles. In the

following lemma, instead of saying that μ_n satisfies the large deviation principle with speed s_n and rate function I, we say more briefly that (μ_n, s_n) satisfies the large deviation principle with rate function I.

Lemma 1.30 (Large deviation principles along subsequences) *Let E be a Polish space, let μ_n be probability measures on E, let s_n be positive constants tending to infinity, and let I be a good rate function on E. Assume that each subsequence (μ'_n, s'_n) of (μ_n, s_n) contains a further subsequence (μ''_n, s''_n) that satisfies the large deviation principle with rate function I. Then (μ_n, s_n) satisfies the large deviation principle with rate function I.*

Proof Assume that (μ_n, s_n) does not satisfy the large deviation principle with rate function I. Then there exists a function $f \in \mathcal{C}_{b,+}(E)$ and an $\varepsilon > 0$ such that $\big|\|f\|_{s_n,\mu_n} - \|f\|_{\infty,I}\big| \geq \varepsilon$ for infinitely many n, hence there exists a subsequence (μ'_n, s'_n) such that $\big|\|f\|_{s'_n,\mu'_n} - \|f\|_{\infty,I}\big| \geq \varepsilon$ for all n. But then no subsequence (μ''_n, s''_n) of (μ'_n, s'_n) can satisfy the large deviation principle with rate function I, contradicting our assumption. ∎

The following lemma generalizes Lemmas 1.12 and 1.15 to unbounded functions.

Lemma 1.31 (Varadhan's lemma for unbounded functions) *Let E be a Polish space and let $\mu_n \in \mathcal{M}(E)$ satisfy the large deviation principle with speed s_n and good rate function I. Let $F : E \to [-\infty, \infty)$ be continuous and assume that the weighted measures $\nu_n(\mathrm{d}x) := e^{s_n F(x)}\mu_n(\mathrm{d}x)$ are exponentially tight. Then*

$$\lim_{n\to\infty} \frac{1}{s_n} \log \int e^{s_n F} \mathrm{d}\mu_n = \sup_{x\in E} \big[F(x) - I(x)\big]. \tag{1.11}$$

Moreover, the weighted measures ν_n satisfy the large deviation principle with speed s_n and good rate function $I - F$.

Proof We start by proving the final claim of the lemma. By Lemma 1.30, it suffices to prove that $I - F$ is a good rate function and that each subsequence (ν'_n, s'_n) of (ν_n, s_n) contains a further subsequence (ν''_n, s''_n) that satisfies the large deviation principle with rate function $I - F$. By our exponential tightness assumption and Theorem 1.25, (ν'_n, s'_n) contains a further subsequence (ν''_n, s''_n) that satisfies the large deviation principle for some good rate function J. It therefore suffices to show that $J = I - F$.

Let $G : E \to [-\infty, \infty)$ be continuous and assume that both G and $G + F$ are bounded from above. Then Varadhan's lemma tells us that

$$\begin{aligned}\sup_{x\in E}\big[G(x) + F(x) - I(x)\big] &= \lim_{n\to\infty} \frac{1}{s''_n} \int_E e^{s''_n(G(x)+F(x))}\mu''_n(\mathrm{d}x)\\ &= \lim_{n\to\infty} \frac{1}{s''_n} \int_E e^{s''_n G(x)}\nu''_n(\mathrm{d}x) = \sup_{x\in E}\big[G(x) - J(x)\big].\end{aligned}$$

In other words, setting $g := e^G$, $f := e^F$ this says that if $g \in \mathcal{C}_{b,+}(E)$ has the property that also $fg \in \mathcal{C}_{b,+}(E)$, then $\|fg\|_{\infty,I} = \|g\|_{\infty,J}$.

We claim that for each $x \in E$, we can find $g_n \in \mathcal{C}_{b,+}(E)$ such that $fg_n \in \mathcal{C}_{b,+}(E)$ for each n and $g_n \downarrow 1_{\{x\}}$. To prove this, we first use Lemma 1.6 to construct $h_n \in \mathcal{C}_{b,+}(E)$ with $h_n \downarrow 1_{\{x\}}$. Setting

$$g_n(y) := \frac{f(x) \vee 1}{f(y) \vee 1} h_n(y) \qquad (y \in E)$$

then does the job, since the inequality $(f \vee 1)g_n \leq (f(x) \vee 1)h_n$ shows that both fg_n and g_n are bounded.

By our earlier claim, Lemma 1.8 (c) now implies that

$$e^{F(x)-I(x)} = \|f(x)1_{\{x\}}\|_{\infty,I} = \lim_{n\to\infty} \|fg_n\|_{\infty,I} = \lim_{n\to\infty} \|g_n\|_{\infty,J} = \|1_{\{x\}}\|_{\infty,J} = e^{-J(x)}$$

for each $x \in E$, which proves that $J = I - F$. This completes the proof that the weighted measures ν_n satisfy the large deviation principle with speed s_n and good rate function $I - F$. Applying Varadhan's lemma to the function that is constantly zero and the measures ν_n then implies (1.11). ∎

1.10 Applications of exponential tightness

In this section, we look at some applications of Theorem 1.25. By definition, if I is a normalized good rate function, then we say that a set of functions $\mathcal{D} \subset \mathcal{C}_b(E)$ *determines* I if for any normalized good rate function J,

$$\|f\|_{\infty,I} = \|f\|_{\infty,J} \quad \forall f \in \mathcal{D} \quad \text{implies} \quad I = J.$$

We say that $\mathcal{D}$ is *rate function determining* if $\mathcal{D}$ determines any normalized good rate function I. By combining Lemma 1.24 and Theorem 1.25, we obtain the following analogue of Lemma 1.20. Note that by Lemma 1.24, the conditions (i) and (ii) below are clearly necessary for the measures μ_n to satisfy a large deviation principle.

Proposition 1.32 (Conditions for LDP) *Let E be a Polish space, let μ_n be probability measures on E, and let s_n be positive constants converging to infinity. Assume that $\mathcal{D} \subset \mathcal{C}_b(E)$ is rate function determining and that:*

(i) *The sequence $(\mu_n)_{n\geq 1}$ is exponentially tight with speed s_n.*

(ii) *The limit $\Lambda(f) = \lim_{n\to\infty} \|f\|_{s_n,\mu_n}$ exists for all $f \in \mathcal{D}$.*

Then there exists a good rate function I on E which is uniquely characterized by the requirement that $\Lambda(f) = \|f\|_{\infty,I}$ for all $f \in \mathcal{D}$, and the μ_n satisfy the large deviation principle with speed s_n and rate function I.

Proof By exponential tightness and Theorem 1.25, there exist $n(m) \to \infty$ and a normalized rate function I such that the $\mu_{n(m)}$ satisfy the large deviation principle with speed $s_{n(m)}$ and good rate function I. It follows that

$$\Lambda(f) = \lim_{m\to\infty} \|f\|_{s_{n(m)},\mu_{n(m)}} = \|f\|_{\infty,I} \qquad (f \in \mathcal{D}),$$

which characterizes I uniquely by the fact that $\mathcal{D}$ is rate function determining. By the same argument, each subsequence (μ'_n, s'_n) of (μ_n, s_n) contains a further subsequence (μ''_n, s''_n) such that the μ''_n satisfy the large deviation principle with speed s''_n and rate function I. By Lemma 1.30, this implies that the μ_n satisfy the large deviation principle with speed s_n and rate function I. ∎

A somewhat weaker version of Proposition 1.32 where $\mathcal{D}$ is replaced by $\mathcal{C}_{b,+}$ is known as Bryc's theorem [Bry90], which can also be found in [DZ93, Theorem 4.4.2].

In view of Proposition 1.32, we are interested in finding sufficient conditions for a set $\mathcal{D} \subset \mathcal{C}_{b,+}$ to be rate function determining. The following simple observation is useful.

Lemma 1.33 (Sufficient conditions to be rate function determining)
(a) *Let E be a Polish space, $\mathcal{D} \subset \mathcal{C}_{b,+}(E)$, and assume that for each $x \in E$ there exist $f_k \in \mathcal{D}$ such that $f_k \downarrow 1_{\{x\}}$. Then $\mathcal{D}$ is rate function determining.*
(b) *Let E be a compact metrizable space, let $\mathcal{C}(E)$ be the Banach space of all continuous real functions on E, equipped with the supremumnorm, and let $\mathcal{D} \subset \mathcal{C}(E)$ be dense. Then $\mathcal{D}$ is rate function determining.*

Proof If $f_k \downarrow 1_{\{x\}}$, then, by Lemma 1.8, $\|f_k\|_{\infty,I} \downarrow \|1_{\{x\}}\|_{\infty,I} = e^{-I(x)}$, proving part (a). Part (b) follows from the fact that the map $f \mapsto \|f\|_{\infty,I}$ is continuous w.r.t. the supremumnorm, as proved in Proposition 1.22. ∎

Proposition 1.32 shows that in the presence of exponential tightness, it is possible to prove large deviation principles by showing that the limit $\lim_{n\to\infty} \|f\|_{s_n,\mu_m}$ exists for sufficiently many continuous functions f. Often, it is more convenient to prove that the large deviations upper and lower bounds from Proposition 1.7 hold for suffiently many closed and open sets.

Let $\mathcal{A}$ be a collection of measurable subsets of some Polish space E. We say that $\mathcal{A}$ is *rate function determining* if for any pair I, J of normalized good rate functions on E, the condition

$$\inf_{x\in\overline{A}} I(x) \leq \inf_{x\in\mathrm{int}(A)} J(x) \quad \forall A \in \mathcal{A} \tag{1.12}$$

implies that $I \leq J$. A set $\mathcal{O}' \subset \mathcal{O}$ is a *basis for the topology* if every $O \in \mathcal{O}$ can be written as a (possibly uncountable) union of sets in $\mathcal{O}'$. Equivalently, this says that for each $x \in E$ and open set $O \ni x$, there exists some $O' \in \mathcal{O}'$ such that $x \in O' \subset O$. For example, in any metric space, the open balls form a basis for the topology.

Lemma 1.34 (Rate function determining sets) *Let $\mathcal{A}$ be a collection of measurable subsets of a Polish space E. Assume that $\{\mathrm{int}(A) : A \in \mathcal{A}\}$ is a basis for the topology. Then $\mathcal{A}$ is rate function determining.*

Proof Choose $\varepsilon_k \downarrow 0$. Since $\{\mathrm{int}(A) : A \in \mathcal{A}\}$ is a basis for the topology, for each $z \in E$ and k there exists some $A_k \in \mathcal{A}$ such that $z \in \mathrm{int}(A_k) \subset B_{\varepsilon_k}(z)$. Since I is a good rate function, it assumes its minimum over $\overline{A}_k$, so (1.12) implies that there exist $z_k \in \overline{A}_k$ such that $I(z_k) \leq \inf_{x\in\mathrm{int}(A_k)} J(x) \leq J(z)$. Since $z_k \to z$, the lower semi-continuity of I implies that $I(z) \leq \liminf_{k\to\infty} I(z_k) \leq J(z)$. ∎

Theorem 1.35 (Conditions for LDP) *Let E be a Polish space, let μ_n be probability measures on E, let s_n be positive constants converging to infinity, let I be a normalized good rate function on E, and let $\mathcal{A}_{\mathrm{up}}, \mathcal{A}_{\mathrm{low}}$ be collections of measurable subsets of E that are rate function determining. Then the μ_n satisfy the large deviation principle with speed s_n and rate function I if and only if the following three conditions are satisfied.*

(i) $\displaystyle\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) \leq -\inf_{x\in\overline{A}} I(x) \quad \forall A \in \mathcal{A}_{\mathrm{up}},$

(ii) $\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) \geq - \inf_{x\in \mathrm{int}(A)} I(x) \quad \forall A \in \mathcal{A}_{\mathrm{low}},$

(iii) *the μ_n are exponentially tight.*

Proof The necessity of the conditions (i)–(iii) follows from Remark 1 below Proposition 1.7 and Lemma 1.24. To prove sufficiency, we use Lemma 1.30. By Theorem 1.25, exponential tightness implies that each subsequence (μ_n', s_n') of (μ_n, s_n) contains a further subsequence (μ_n'', s_n'') of such that the μ_n'' satisfy a large deviations principle with speed s_n'' and some good rate function J. By Lemma 1.30, if we can show that for each such subsequence, $J = I$, then it follows that the μ_n satisfy the large deviations principle with speed s_n and rate function I.

In view of this, it suffices to show that if the μ_n satisfy a large deviations principle with speed s_n and some good rate function J and conditions (i) and (ii) are satisfied, then $J = I$. Indeed, condition (i) and the large deviation principle for J imply that for any $A \in \mathcal{A}_{\mathrm{up}}$,

$$- \inf_{x\in \mathrm{int}(A)} J(x) \leq \liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(\mathrm{int}(A)) \leq \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) \leq - \inf_{x\in \overline{A}} I(x),$$

which by the assumption that $\mathcal{A}_{\mathrm{up}}$ is rate function determining implies that $I \leq J$. Similarly, using (ii) instead of (i), we find that for any $A \in \mathcal{A}_{\mathrm{low}}$,

$$- \inf_{x\in \mathrm{int}(A)} I(x) \leq \liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) \leq \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(\overline{A}) \leq - \inf_{x\in \overline{A}} J(x),$$

which by the assumption that $\mathcal{A}_{\mathrm{low}}$ is rate function determining implies that $J \leq I$. ∎

Remark In Theorem 1.35, instead of assuming that $\mathcal{A}_{\mathrm{low}}$ is rate function determining, it suffices to assume that

$$\forall \varepsilon > 0 \text{ and } z \in E \text{ s.t. } I(z) < \infty,\ \exists A \in \mathcal{A}_{\mathrm{low}} \text{ s.t. } z \in A \subset B_\varepsilon(z). \tag{1.13}$$

Indeed, the proof of Lemma 1.34 shows that if (1.12) holds with I and J interchanged, and we moreover have (1.13), then $J(z) \leq I(z)$ for all $z \in E$ such that $I(z) < \infty$. Trivially, this also holds if $I(z) = \infty$, and the proof proceeds as before.∎

The next lemma shows that in Theorem 1.35, instead of assuming that $\mathcal{A}_{\mathrm{up}}$ is rate function determining, we can also take for $\mathcal{A}_{\mathrm{up}}$ the set of all compact subsets of E. If E is locally compact, then $\{\mathrm{int}(K) : K \text{ compact}\}$ is a basis for the topology, so

in view of Lemma 1.34 this does not add anything new. However, if E is not locally compact, then $\{\mathrm{int}(K) : K \text{ compact}\}$ is never a basis for the topology. In fact, there exist Polish spaces in which every compact set has empty interior. Clearly, in such spaces, the compact sets are not rate function determining and hence the lemma below does add something new.

Lemma 1.36 (Upper bound for compact sets) *Let E be a Polish space, let μ_n be finite measures on E, let s_n be positive constants converging to infinity, and let I be a good rate function on E. Assume that*

(i) *The sequence $(\mu_n)_{n\geq 1}$ is exponentially tight with speed s_n.*

(ii) $\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(K) \leq -\inf_{x\in K} I(x) \qquad \forall K$ *compact.*

Then

$$\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C) \leq -\inf_{x\in C} I(x) \qquad \forall C \text{ closed.}$$

Remark If $I : E \to (-\infty, \infty]$ is lower semi-continuous and not identically ∞, but not necessarily has compact level sets, and if μ_n are measures and $s_n \to \infty$ constants such that

(i) $\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(K) \leq -\inf_{x\in K} I(x) \qquad \forall K$ compact.

(ii) $\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(O) \leq -\inf_{x\in O} I(x) \qquad \forall O$ open,

then one says that the μ_n satisfy the *weak large deviation principle* with speed s_n and rate function I. Thus, a weak large deviation principle is basically a large deviation principle without exponential tightness. The theory of weak large deviation principles is much less elegant than for large deviation principles. For example, the contraction principle (Proposition 1.14 below) may fail for measures satisfying a weak large deviation principle.

Proof of Lemma 1.36 By exponential tightness, for each $M < \infty$ we can find a compact $K \subset E$ such that

$$\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(E\backslash K) \leq -M.$$

By (1.4), it follows that, for any closed $C \subset E$,

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C) &= \limsup_{n\to\infty} \frac{1}{s_n} \log \big(\mu_n(C\cap K) + \mu_n(C\backslash K)\big) \\ &= \Big(\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C\cap K)\Big) \vee \Big(\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C\backslash K))\Big) \\ &\leq -\big(M \wedge \inf_{x\in C\cap K} I(x)\big) \leq -\big(M \wedge \inf_{x\in C} I(x)\big) \underset{M\to\infty}{\longrightarrow} -\inf_{x\in C} I(x).\end{aligned}$$

∎

Let E and F be sets and let $(f_\gamma)_{\gamma\in\Gamma}$ be a collection of functions $f : E \to F$. By definition, we say that $(f_\gamma)_{\gamma\in\Gamma}$ *separates points* if for each $x, y \in E$ with $x \neq y$, there exists a $\gamma \in \Gamma$ such that $f_\gamma(x) \neq f_\gamma(y)$. The following theorem is a sort of 'inverse' of the contraction principle, in the sense that a large deviation principle for sufficiently many image measures implies a large deviation principle for the original measures. For weak convergence, the analogous statement is that if we have a sequence $X^{(n)}$ of discrete-time processes $(X_i^{(n)})_{i\in\mathbb{N}}$, then weak convergence of the finite dimensional distributions implies weak convergence in law of the processes.

Theorem 1.37 (Projective limit) *Let E and F be Polish spaces, let μ_n be probability measures on E, and let s_n be positive constants converging to infinity. Let $(\psi_i)_{i\in\mathbb{N}_+}$ be continuous functions $\psi_i : E \to F$. For each $m \geq 1$, let $\vec{\psi}_m : E \to F^m$ be defined as $\vec{\psi}_m(x) = \big(\psi_1(x), \ldots, \psi_m(x)\big)$ $(x \in E)$. Assume that $(\psi_i)_{i\in\mathbb{N}_+}$ separates points and that:*

(i) *The sequence $(\mu_n)_{n\geq 1}$ is exponentially tight with speed s_n.*

(ii) *For each finite $m \geq 1$, there exists a good rate function I_m on F^m, equipped with the product topology, such that the measures $\mu_n \circ \vec{\psi}_m^{-1}$ satisfy the large deviation principle with speed s_n and rate function I_m.*

Then there exists a good rate function I on E which is uniquely characterized by the requirement that

$$I_m(y) = \inf_{x:\, \vec{\psi}_m(x)=y} I(x) \qquad (m \geq 1,\ y \in F^m).$$

Moreover, the measures μ_n satisfy the large deviation principle with speed s_n and rate function I.

Proof Our assumptions imply that for each $f \in \mathcal{C}_{b,+}(F^m)$,

$$\|f \circ \vec{\psi}_m\|_{s_n,\mu_n} = \|f\|_{s_n,\mu_n \circ \vec{\psi}_m^{-1}} \underset{n\to\infty}{\longrightarrow} \|f\|_{\infty,I_m}.$$

We claim that the set

$$\mathcal{D} := \big\{f \circ \vec{\psi}_m : m \geq 1,\ f \in \mathcal{C}_{b,+}(F^m)\big\}$$

is rate function determining. To see this, fix $z \in E$ and define $f_{i,k} \in \mathcal{D}$ by

$$f_{i,k}(x) := \big(1 - kd\big(\psi_i(x), \psi_i(z)\big)\big) \vee 0 \qquad (i,k \geq 1,\ y \in E),$$

where d is any metric generating the topology on F. We claim that

$$\mathcal{D} \ni \bigwedge_{i=1}^m f_{i,m} \downarrow 1_{\{z\}} \quad \text{as } m \uparrow \infty.$$

Indeed, since the $(\psi_i)_{i\in\mathbb{N}_+}$ separate points, for each $x \neq z$ there is an $i \geq 1$ such that $\psi_i(x) \neq \psi_i(z)$ and hence $f_{i,m}(y) = 0$ for m large enough. By Lemma 1.33 (a), it follows that $\mathcal{D}$ is rate function determining.

Proposition 1.32 now implies that there exists a good rate function I on E such that the μ_n satisfy the large deviation principle with speed s_n and rate function I. Moreover, I is uniquely characterized by the requirement that

$$\|f \circ \vec{\psi}_m\|_{\infty,I} = \|f\|_{\infty,I_m} \qquad \big(m \geq 1,\ f \in \mathcal{C}_{b,+}(F^m)\big). \tag{1.14}$$

Set

$$I'_m(y) := \inf_{x:\ \vec{\psi}_m(x)=y} I(x) \qquad (y \in F^m),$$

which by the contraction principle (Proposition 1.14) is a good rate function on F^m. Since

$$\begin{aligned}\|f \circ \vec{\psi}_m\|_{\infty,I} &= \sup_{x\in E} e^{-I(x)} f\big(\vec{\psi}_m(x)\big)\\ &= \sup_{y\in F^m} e^{-\inf_{x:\ \vec{\psi}_m(x)=y} I(x)} f(y) = \|f\|_{\infty,I'_m},\end{aligned}$$

formula (1.14) implies that $\|f\|_{\infty,I'_m} = \|f\|_{\infty,I_m}$ for all $f \in \mathcal{C}_{b,+}(F^m)$, which is in turn implies that $I_m = I'_m$. ∎

The following lemma gives a more explicit expression for the rate function I from Theorem 1.37 in terms of the rate functions $\vec{\psi}_m$.

Lemma 1.38 (Formula for high-level rate function) *In the set-up of Theorem 1.37,*

$$I_m(\vec{\psi}_m(x)) \uparrow I(x) \quad \text{as } m \uparrow \infty.$$

Proof We observe that

$$I_m(\vec{\psi}_m(x)) = \inf_{x' \in E:\, \vec{\psi}_m(y) = \vec{\psi}_m(x)} I(x').$$

The sets $C_m := \{x' \in E : \vec{\psi}_m(y) = \vec{\psi}_m(x)\}$ are closed and decrease to $\{x\}$ as $m \uparrow \infty$ by the fact that the ψ_i separate points. Therefore, by Lemma 1.8 (c), $\inf_{x' \in C_m} I(x') \uparrow I(x)$ as $\uparrow \infty$. ∎

Chapter 2

Convex analysis

2.1 Convex sets

In the rather long Chapter 1, we developed the abstract theory of large deviation principles. It may seem that we are we are now finally ready to turn to applications and prove some concrete large deviation principles like Cramér's theorem, the moderate deviations theorem, or the Boltzmann-Sanov theorem. This impression is not correct. Large deviation theory, as it turns out, is built on two abstract pillars. One of these, the abstract theory of large deviation principles, we have just seen in the previous chapter. The second one, that we are still missing, is convex analysis, and in particular the theory of the Legendre transform. In view of this, we ask the reader to be patient and bear with us during the present chapter which, while less voluminous than the previous one, still contains a sizable bit of theory. Luckily, the theory of convex functions is quite interesting in itself and also has many applications outside the theory of large deviations.

By definition, a set $C \subset \mathbb{R}^d$ is *convex* if $(1-p)x + py \in C$ for all $x, y \in C$ and $p \in [0,1]$. The *convex hull* $C(A)$ of a set $A \subset \mathbb{R}^d$ is the smallest convex set that contains it, which is given by

$$C(A) = \Big\{ \sum_{k=1}^{n} p_k x_k : x_1, \ldots, x_n \in A,\ p_1, \ldots, p_n \geq 0,\ \sum_{k=1}^{n} p_k = 1 \Big\}.$$

In particular, A is convex if and only if $C(A) = A$. The *closed convex hull* $\overline{C}(A)$ of A is the closure of $C(A)$. A set $C \subset \mathbb{R}^d$ is a *convex cone* if $p_1 x + p_2 y \in C$ for all $x, y \in C$ and $p_1, p_2 \geq 0$. A set $A \subset \mathbb{R}^d$ is *affine* if $(1-p)x + py \in C$ for all

$x, y \in C$ and $p \in \mathbb{R}$. The *affine hull* of a set $A \subset \mathbb{R}^d$ is the set

$$\Big\{ \sum_{k=1}^{n} p_k x_k : x_1, \ldots, x_n \in A,\ p_1, \ldots, p_n,\ \sum_{k=1}^{n} p_k = 1 \Big\},$$

where this time we do not require that the real constants $p_1, \ldots, p_n$ are nonnegative. Each affine set $A \subset \mathbb{R}^d$ is of the form $A = \{x + y : y \in V\}$ where V is a linear subspace of $\mathbb{R}^d$. In particular, affine sets are always closed.

Recall that the *interior* $\mathrm{int}(A)$ of a set A is the largest open set contained in A. The *relative interior* of a closed convex set $C \subset \mathbb{R}^d$ is the interior of C when viewed as a subset of its affine hull. Each nonempty convex set $C \subset \mathbb{R}^d$ has a nonempty relative interior[1] and each closed convex set $C \subset \mathbb{R}^d$ is the closure of its relative interior.

We denote a vector in $\mathbb{R}^d$ as $x = \big(x(1), \ldots, x(d)\big)$ and let

$$\langle x, y \rangle := \sum_{i=1} x(i) y(j) \qquad (x, y \in \mathbb{R}^d)$$

denote the usual inner product. Each $x^* \in \mathbb{R}^d \backslash \{0\}$ and $c^* \in \mathbb{R}$ define two closed *half-spaces* by

$$H^{\leq}_{x^*,c^*} := \big\{ x \in \mathbb{R}^d : \langle x^*, x \rangle \leq c^* \big\},$$
$$H^{\geq}_{x^*,c^*} := \big\{ x \in \mathbb{R}^d : \langle x^*, x \rangle \geq c^* \big\}.$$

We let $H_{x^*,c^*} := H^{\leq}_{x^*,c^*} \cap H^{\geq}_{x^*,c^*}$ denote the $(d-1)$-dimensional hyperplane that separates the half-spaces $H^{\leq}_{x^*,c^*}$ and $H^{\geq}_{x^*,c^*}$. One can prove that the closed convex hull of a set A is equal to the intersection of all closed half-spaces that contain it:

$$\overline{C}(A) = \bigcap \big\{ H^{\leq}_{x^*,c^*} : x^* \in \mathbb{R}^d \backslash \{0\},\ c^* \in \mathbb{R},\ A \subset H^{\leq}_{x^*,c^*} \big\}. \tag{2.1}$$

A formal proof may easily be deduced from [Roc70, Theorem 11.5] or [Dud02, Thm 6.2.9]. The basic ingredient in the proof of (2.1) is the following separation theorem, which we cite from [Roc70, Theorem 11.3]. See Figure 2.1 for an illustration.

Theorem 2.1 (Separating hyperplane *Let $C_1, C_2 \subset \mathbb{R}^d$ be convex sets with relative interiors* $\mathrm{ri}(C_i)$ $(i = 1, 2)$. *Then the following statements are equivalent.*

(i) $\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2) = \emptyset$

(ii) *There exists a $x^* \in \mathbb{R}^d \backslash \{0\}$ and $c^* \in \mathbb{R}$ such that $C_1 \in H^{\leq}_{x^*,c^*}$ and $C_2 \in H^{\leq}_{x^*,c^*}$.*

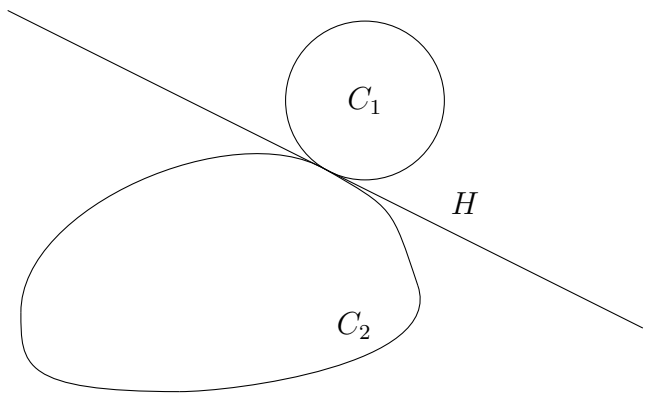

Figure 2.1: A hyperplane H separating the convex sets C_1 and C_2.

The following lemma is a simple consequence of Theorem 2.1.

Lemma 2.2 (Supporting hyperplane) *Let $C \subset \mathbb{R}^d$ be a closed convex set. Assume that the interior* $\mathrm{int}(C)$ *is nonempty and let $x \in C \backslash \mathrm{int}(C)$ be a point on the boundary of C. Then there exists an $x^* \in \mathbb{R}^d \backslash \{0\}$ and $c^* \in \mathbb{R}$ such that*

$$C \subset H^{\leq}_{x^*,c^*} \quad \textit{and} \quad x \in H^{\geq}_{x^*,c^*}. \tag{2.2}$$

Proof Apply Theorem 2.1 to the convex sets C and $\{x\}$, using the fact that the relative interior of C is $\mathrm{int}(C)$ and the relative interior of $\{x\}$ is $\{x\}$, and these are disjoint. ∎

If (2.2) holds, then we say that H_{x^*} is a *supporting hyperplane* at x.

2.2 Convex functions

For any function $f : \mathbb{R}^d \to (-\infty, \infty]$, we call

$$\mathcal{D}_f := \big\{x \in \mathbb{R}^d : f(x) < \infty\big\} \quad \text{and} \quad \mathcal{U}_f := \mathrm{int}(\mathcal{D}_f).$$

the *domain* of f and the interior of the domain, respectively, and we call

$$\mathcal{E}(f) := \big\{(x, c) : x \in \mathcal{D}_f,\ c \in \mathbb{R},\ f(x) \leq c\big\}$$

[1]This is true even when C consists of s single point x. In this case, the relative interior of C is $\{x\}$, which is both open and closed as a subset of the affine hull of C, which is also $\{x\}$.

the *epigraph* of f.

Recall that a function $f : \mathbb{R}^d \to (-\infty, \infty]$ is *convex* if $f(px_1 + (1-p)x_2) \leq pf(x_1) + (1-p)f(x_2)$ for all $0 \leq p \leq 1$ and $x_1, x_2 \in \mathbb{R}^d$. We say that a function f is *strictly convex* on an convex set U if $f(px + (1-p)y) < pf(x) + (1-p)f(y)$ for all $0 < p < 1$ and $x, y \in U$ with $x \neq y$. We let $\mathrm{Conv}(\mathbb{R}^d)$ denote the space of functions $f : \mathbb{R}^d \to (-\infty, \infty]$ such that:

(i) f is convex,

(ii) f is not identically ∞,

(iii) f is lower semi-continuous.

In view of the following two exercises, a function $f : \mathbb{R}^d \to (-\infty, \infty]$ satisfies $f \in \mathrm{Conv}(\mathbb{R}^d)$ if and only if the epigraph $\mathcal{E}(f)$ is a nonempty, closed, and convex subset of $\mathbb{R}^{d+1}$.

Exercise 2.3 (Epigraph of a lower semi-continuous function) Show that a function $f : \mathbb{R}^d \to (-\infty, \infty]$ is lower semi-continuous if and only if its epigraph $\mathcal{E}(f)$ is a closed subset of $\mathbb{R}^{d+1}$.

Exercise 2.4 (Epigraph of a convex function) Show that a function $f : \mathbb{R}^d \to (-\infty, \infty]$ is convex if and only if its epigraph $\mathcal{E}(f)$ is a convex subset of $\mathbb{R}^{d+1}$.

We note that if f is convex, then $\mathcal{D}_f$ is a convex subset of $\mathbb{R}^d$. For a proof of the following well-known fact we refer to [Roc70, Thm 10.2].

Lemma 2.5 (Continuity of convex functions) *If $f \in \mathrm{Conv}(\mathbb{R}^d)$, then its restriction to $\mathcal{D}_f$ is a continuous function.*

Assume that $f \in \mathrm{Conv}(\mathbb{R}^d)$ and that $\mathcal{U}_f \neq \emptyset$. We will be interested in the supporting hyperplanes of the epigraph $\mathcal{E}_f$. For each $x \in \mathcal{D}_f$, we let $Hf(x)$ denote the set of all $(x^*, a^*) \in \mathbb{R}^{d+1}$ such that

$$\langle x^*, y - x \rangle + a^*\big(z - f(x)\big) \leq 0 \quad \forall y \in \mathcal{D}_f \text{ and } z \geq f(y). \tag{2.3}$$

Note that this implies $a^* \leq 0$, since otherwise (2.3) is violated for z large enough. As we will see shortly, $Hf(x)$ roughly corresponds to the set of all supporting hyperplanes for $\mathcal{E}(f)$ at $\big(x, f(x)\big)$. We also let

$$Hf := \big\{\big(x, (x^*, a^*)\big) : x \in \mathcal{D}_f,\ (x^*, a^*) \in Hf(x)\big\} \tag{2.4}$$

denote the space of all pairs $\big(x,(x^*,a^*)\big)$ such that $x \in \mathcal{D}_f$ and $(x^*,a^*) \in Hf(x)$. For $x \in \mathcal{D}_f$, we moreover set

$$H'f(x) := \big\{(x^*,a^*) \in Hf : (x^*,a^*) \neq 0\big\},$$
$$H''f(x) := \big\{(x^*,a^*) \in Hf : a^* < 0\big\},$$

and we define $H'f$ and $H''f$ as in (2.4) but with $Hf(x)$ replaced by $H'f(x)$ or $H''f(x)$, respectively. Hyperplanes $H_{(x^*,a^*),c^*}$ with $a^* = 0$ are called *vertical*, for obvious reasons.

Proposition 2.6 (Supporting hyperplanes) *Assume that $f \in \mathrm{Conv}(\mathbb{R}^d)$ and that $\mathcal{U}_f \neq \emptyset$. Then,*

(a) *$Hf(x)$ is a closed convex cone in $\mathbb{R}^{d+1}$,*

(b) *$H'f(x) \neq \emptyset$ for each $x \in \mathcal{D}_f$,*

(c) *$H''f(x) \neq \emptyset$ for each $x \in \mathcal{U}_f$,*

(d) *$H''f$ is a closed subset of $\mathbb{R}^d \times \mathbb{R}^d \times (\mathbb{R}\backslash\{0\})$.*

(e) *$H''f$ is a connected subset of $\mathbb{R}^d \times \mathbb{R}^d \times (\mathbb{R}\backslash\{0\})$.*

Proof Part (a) is immediate from (2.3). Since $\mathcal{U}_f \neq \emptyset$, the interior of $\mathcal{E}(f)$ is nonempty and for each $x \in \mathcal{D}_f$, the point $\big(x, f(x)\big)$ lies on the boundary of $\mathcal{E}(f)$. We can therefore apply Lemma 2.2 to conclude that for each $x \in \mathcal{D}_f$, there exist $(x^*,a^*) \in \mathbb{R}^{d+1}\backslash\{0\}$ and $c^* \in \mathbb{R}$ such that

$$\mathcal{E}(f) \subset H^{\leq}_{(x^*,a^*),c^*} \quad \text{and} \quad x \in H^{\geq}_{(x^*,a^*),c^*}.$$

In other words, this says that

$$\langle x^*, y\rangle + a^* z \leq c^* \quad \big(y \in \mathcal{D}_f,\ z \geq f(y)\big) \quad \text{and} \quad \langle x^*, x\rangle + a^* f(x) \geq c^*.$$

Since this implies that $\langle x^*, x\rangle + a^* f(x) = c^*$, we can simplify this to (2.3). This proves part (b).

To prove part (c), we use part (b) and observe that by (2.3), $(x^*, 0) \in H'f(x)$ implies

$$\langle x^*, y - x\rangle \leq 0 \quad \forall y \in \mathcal{D}_f,$$

so setting $c^* := \langle x^*, x\rangle$, we see that $\mathcal{D}_f \subset H^{\leq}_{x^*,c^*}$ and $x \in H^{\geq}_{x^*,c^*}$, which is only possible if x lies on the boundary of $\mathcal{D}_f$.

To prove part (d), assume that $\big(x_n,(x_n^*,a_n^*)\big)\in H''f$ converge to a limit $\big(x,(x^*,a^*)\big)$ in $\mathbb{R}^d\times\mathbb{R}^d\times(\mathbb{R}\backslash\{0\})$. Then, taking the limit in (2.3), we see that $\big(x_n,(x_n^*,a_n^*)\big)\in Hf$. Since $a_n^*<0$, formula (2.3) moreover implies that $f(x)<\infty$ and hence $x\in\mathcal{D}_f$, so we see that $\big(x_n,(x_n^*,a_n^*)\big)\in H''f$.

It remains to prove part (e). We recall that a closed set A is *connected* if it cannot be written as the union $A=A_1\cup A_2$ of two disjoint nonempty closed sets A_1,A_2. Since $Hf(x)$ is convex by part (a), we see that $H''f(x)$ is convex too and therefore connected. If $H''f$ is not connected, then $H''f=A_1\cup A_2$ where A_1,A_2 are disjoint nonempty closed subsets of $\mathcal{D}_f\times\mathbb{R}^d\times(\mathbb{R}\backslash\{0\})$. Since the sets $H''f(x)$ are connected, for each $x\in\mathcal{D}_f$, the set $\{x\}\times H''f(x)$ must be either entirely contained in A_1, or in A_2. It follows that setting $B_i:=\{x\in\mathcal{D}_f:\{x\}\times H''f(x)\subset A_i\}$ $(i=1,2)$ defines disjoint nonempty closed subsets B_1,B_2 of $\mathcal{D}_f$ whose union is $\mathcal{D}_f$. But $\mathcal{D}_f$ is convex and hence connected, so we arrive at a contradiction. ∎

2.3 The generalized gradient

A function $f:\mathbb{R}^d\to\mathbb{R}$ is *affine* if

$$f\big((1-p)x+py\big)=(1-p)f(x)+pf(y)\qquad(x,y\in\mathbb{R}^d,\ p\in\mathbb{R}).$$

Each affine function is the sum of a linear function and a constant, and can therefore be written in the form

$$f(x)=\langle x^*,x\rangle-c^*\qquad(x\in\mathbb{R}^d)$$

for some $x^*\in\mathbb{R}^d$ and $c^*\in\mathbb{R}$. We say that an affine function $y\mapsto x^*y-c^*$ is *supporting* at a point $x\in\mathcal{D}_f$ if

$$\langle x^*,x\rangle-c^*=f(x)\quad\text{and}\quad x^*y-c^*\leq f(y)\quad(y\in\mathbb{R}^d).$$

We call x^* the *slope* of the supporting affine function $y\mapsto x^*y-c^*$. For any $f\in\mathrm{Conv}(\mathbb{R}^d)$ and $x\in\mathcal{D}_f$, we write

$$\begin{aligned}Df(x)&:=\big\{x^*\in\mathbb{R}^d:f(x)+\langle x^*,y-x\rangle\leq f(y)\ \forall y\in\mathbb{R}^d\big\},\\ Df&:=\big\{(x,x^*):x\in\mathcal{D}_f,\ x^*\in Df(x)\big\}.\end{aligned}\tag{2.5}$$

Df is the collection of all slopes of supporting affine functions at x. The *gradient* of a continuously differentiable function f is defined as

$$\partial f(x):=\big(\tfrac{\partial}{\partial x(1)}f(x),\ldots,\tfrac{\partial}{\partial x(d)}f(x)\big).\tag{2.6}$$

We note that for any $y \in \mathbb{R}^d$,

$$\langle y, \partial f(x) \rangle = \lim_{\varepsilon \to 0} \varepsilon^{-1} \big(f(x + \varepsilon y) - f(x) \big)$$

is the *directional derivative* of f at x in the direction y. If $f \in \mathrm{Conv}(\mathbb{R}^d)$ is differentiable in $x \in \mathcal{D}_f$, then there is a unique supporting affine function at x, whose slope is given by the gradient of f, so in this case $Df(x) = \big\{ \partial f(x) \big\}$. Thus, we can view $Df(x)$ as a possibly multi-valued generalization of the gradient of f.

As the reader may already have guessed, there is a one-to-one correpondence between the set of all supporting affine functions of f and the set of all supporting hyperplanes that are not vertical. We will use this to derive the following proposition from Proposition 2.6.

Proposition 2.7 (Generalized gradient) *Assume that $f \in \mathrm{Conv}(\mathbb{R}^d)$ and that $\mathcal{U}_f \neq \emptyset$. Then:*

(a) *$Df(x) \neq \emptyset$ for all $x \in \mathcal{U}_f$,*

(b) *$Df(x)$ is a closed convex set for all $x \in \mathcal{D}_f$,*

(c) *Df is a closed subset of $\mathbb{R}^{2d}$,*

(d) *Df is a connected subset of $\mathbb{R}^{2d}$.*

Proof Let $x \in \mathcal{D}_f$. If $(x^*, a^*) \in H''f(x)$ and $r > 0$, then $(rx^*, ra^*) \in H''f(x)$. Therefore, Proposition 2.6 (c) implies that for each $x \in \mathcal{U}_f$, there exists an $x^* \in \mathbb{R}^d$ such that $(x^*, -1) \in H''f(x)$. Then (2.3) tels us that

$$\langle x^*, y - x \rangle - \big(f(y) - f(x) \big) \leq 0 \quad \forall y \in \mathcal{D}_f,$$

which shows that $x^* \in Df(x)$. In view of this, part (a) follows from Proposition 2.6 (c). Part (b) is immediate from the definition of $Df(x)$ in (2.5). Parts (c) and (d) follow from Proposition 2.6 (d) and (e) and our earlier observation that each $(x^*, a^*) \in H''f(x)$ can be normalized so that $a^* = -1$. ∎

We have already argued that we can view $Df(x)$ as a generalization of the gradient. The following lemma makes this observation more precise.

Lemma 2.8 (Uniqueness of the slope) *Let $f \in \mathrm{Conv}(\mathbb{R}^d)$. Then the following conditions are equivalent:*

(i) *f is continuously differentiable on $\mathcal{U}_f$,*

(ii) *for each $x \in \mathcal{U}_f$, the set $Df(x)$ consists of a single element.*

Moreover, under these conditions, $Df(x) = \{\partial f(x)\}$ $(x \in \mathcal{U}_f)$, where ∂f is the gradient of f, defined in (2.6)

Proof If f is differentiable at x, then there is a unique supporting affine function at x, so the implication (i)⇒(ii) is trivial. For the converse, we refer to [Roc70, Thm 25.1]. To make this implication at least a bit plausible, we observe that (ii) implies that there exists a function $g : \mathcal{U}_f \to \mathbb{R}^d$ such that

$$Df(x) = \{g(x)\} \qquad (x \in \mathcal{U}_f),$$

Now Proposition 2.7 (b) says that the graph of g is a closed subset of $\mathcal{U}_f \times \mathbb{R}^d$, so the closed graph theorem implies that g is continuous. The technical part of the proof is showing that g is indeed the gradient of f. ∎

2.4 The convex hull of a function

The *convex hull* $\overline{f}$ of a function $f : \mathbb{R}^d \to (-\infty, \infty]$ is the pointwise supremum of all affine functions that lie below f, i.e.,

$$\overline{f}(x) := \sup\big\{\langle x^*, x\rangle - c^* : x^* \in \mathbb{R}^d,\ c^* \in \mathbb{R},\ \langle x^*, y\rangle - c^* \le f(y)\ \forall y \in \mathbb{R}\big\}.$$

It can be shown that $\overline{f}$ is the largest lower semi-continuous convex function such that $\overline{f} \le f$. We cite the following lemma from [Roc70, Thm 12.1].

Lemma 2.9 (Convex hull of a function) *Assume that $f : \mathbb{R}^d \to (-\infty, \infty]$ is not identically ∞. Then $\overline{f} \in \mathrm{Conv}(\mathbb{R}^d)$ and $\overline{f} \le f$. Moreover, if $g \in \mathrm{Conv}(\mathbb{R}^d)$ satisfies $g \le f$, then $g \le \overline{f}$. In particular, $f \in \mathrm{Conv}(\mathbb{R}^d)$ if and only if $f = \overline{f}$.*

Sometimes, to know a function, it suffices to know only its convex hull. Recall the definition of strict convexity from Section 2.2.

Lemma 2.10 (Function determined by its convex hull) *Assume that $f : \mathbb{R}^d \to (-\infty, \infty]$ is lower semi-continuous and assume that its convex hull $\overline{f}$ is strictly convex on $\mathcal{U}_{\overline{f}}$ and that $\mathcal{U}_{\overline{f}} \neq \emptyset$. Then $f = \overline{f}$.*

Proof Let us say that $x \in \mathcal{D}_f$ is an *exposed point* of a function $h \in \mathrm{Conv}(\mathbb{R}^d)$ if there exists a supporting affine function $y \mapsto h(x) + \langle x^*, y - x\rangle$ at x such that

$$h(x) + \langle x^*, y - x\rangle < h(y) \quad \forall y \in \mathbb{R}^d \backslash \{x\}. \tag{2.7}$$

We claim that $f(x) = \overline{f}(x)$ for each exposed point x of $\overline{f}$. To see this, let x^* be as in (2.7) with $\overline{f}$ in place of h and let ε_n be positive constants converging to zero. For each n, there must be an y_n such that $\overline{f}(x) + \varepsilon_n + x^*(y_n - x) > f(y_n)$ since otherwise, the affine function $y \mapsto \overline{f}(x) + \varepsilon_n + \langle x^*, y - x\rangle$ would lie below f contradicting the maximality of $\overline{f}$. Since $\overline{f} \leq f$ it follows that

$$\overline{f}(y_n) \leq f(y_n) < \overline{f}(x) + \varepsilon_n + x^*(y_n - x).$$

It is not hard to see that the closed convex sets

$$C_n := \left\{y \in \mathbb{R}^d : f(y) < \overline{f}(x) + \varepsilon_n + \langle x^*, y - x\rangle\right\}$$

are in fact compact. Since the sets C_n decrease to $\{x\}$, we see that $y_n \to x$ and hence, by the lower semi-continuity of f, it follows that

$$f(x) \leq \liminf_{n\to\infty} f(y_n) \leq \liminf_{n\to\infty} \left[\overline{f}(x) + \varepsilon_n + x^*(y_n - x)\right] = \overline{f}(x).$$

Since $\overline{f} \leq f$, the other inequality is trivial and we conclude that $f(x) = \overline{f}(x)$ as claimed.

If $\overline{f}$ is strictly convex on $\mathcal{U}_{\overline{f}}$, then each point in $\mathcal{U}_{\overline{f}}$ is exposed. By what we have just proved, it follows that $\overline{f} = f$ on $\mathcal{U}_{\overline{f}}$. Since each convex set is the closure of its relative interior and since $\mathcal{U}_{\overline{f}} \neq \emptyset$, for each $x \in \mathcal{D}_{\overline{f}} \backslash \mathcal{U}_{\overline{f}}$, we can choose $\mathcal{U}_{\overline{f}} \ni x_n \to x$. Since f is lower semi-continuous and $\overline{f}$ is continuous on $\mathcal{D}_{\overline{f}}$, it follows that

$$f(x) \leq \liminf_{n\to\infty} f(x_n) \leq \lim_{n\to\infty} \overline{f}(x_n) = \overline{f}(x).$$

This proves that $f(x) \leq \overline{f}(x)$ for all $x \in \mathcal{D}_{\overline{f}}$. Trivially also $f(x) \leq \infty = \overline{f}(x)$ for $x \notin \mathcal{D}_{\overline{f}}$ and $\overline{f} \leq f$ on $\mathbb{R}^d$ since $\overline{f}$ is the convex hull of f, so we conclude that $f = \overline{f}$. ∎

2.5 The Legendre transform

The *Legendre transform*[2] of a function $f : \mathbb{R}^d \to (-\infty, \infty]$ is defined as

$$f^*(x^*) := \sup_{x\in\mathbb{R}^d} \left[\langle x^*, x\rangle - f(x)\right] \qquad (x^* \in \mathbb{R}^d).$$

[2]Sometimes also called *Legendre-Fenchel transform* or *Fenchel-Legendre transform*, to honor Fenchel who first studied the transformation for non-smooth functions.

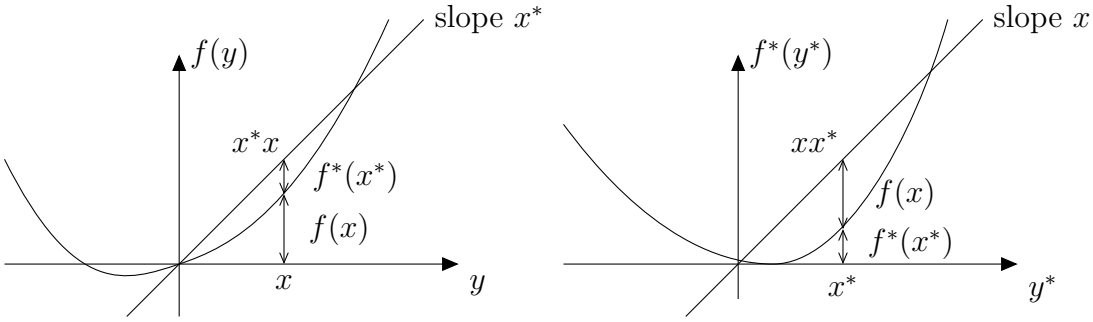

Figure 2.2: The Legendre transform.

This definition is demonstrated in Figure 2.2.

Exercise 2.11 For $a \in \mathbb{R}^d$, let l_a denote the linear function $l_a(x) := \langle a, x\rangle$, and for any function $f : \mathbb{R}^d \to [-\infty, \infty]$, define $T_a f(x) := f(x-a)$ $(x \in \mathbb{R}^d)$. Show that:

(a) $f \leq g \;\Rightarrow\; f^* \geq g^*$.

(b) $(f+c)^* = f^* - c$.

(c) $(f + l_a)^* = T_a f^*$

(d) $(T_a f)^* = f^* + l_a$.

Exercise 2.12 Let $a > 0$. Show that the Legendre transform of the function $f(x) = \frac{1}{2}ax^2$ $(x \in \mathbb{R})$ is given by $f^*(y) = \frac{1}{2a}y^2$ $(y \in \mathbb{R})$.

The following lemma implies that the Legendre transform maps $\mathrm{Conv}(\mathbb{R}^d)$ into itself and that $(f^*)^* = f$ for each $f \in \mathrm{Conv}(\mathbb{R}^d)$.

Lemma 2.13 (Legendre transform) *Assume that $f : \mathbb{R}^d \to (-\infty, \infty]$ is not identically ∞. Then $f^* \in \mathrm{Conv}(\mathbb{R}^d)$. One has*

$$\text{(i) } f^*(x^*) = \sup_{(x,c)\in\mathcal{E}(f)} \big[\langle x^*, x\rangle - c\big], \qquad \text{(ii) } \overline{f}(x) = \sup_{(x^*,c^*)\in\mathcal{E}(f^*)} \big[\langle x, x^*\rangle - c^*\big], \tag{2.8}$$

and

$$\begin{aligned} \text{(i)}\quad \mathcal{E}(\overline{f}) &= \big\{(x,c) : \langle x^*, x\rangle - c \leq f^*(x^*)\ \forall x^* \in \mathbb{R}\big\}, \\ \text{(ii)}\quad \mathcal{E}(f^*) &= \big\{(x^*,c^*) : \langle x, x^*\rangle - c^* \leq f(x)\ \forall x \in \mathbb{R}\big\}. \end{aligned} \tag{2.9}$$

Moreover, $f^* = (\overline{f})^*$ and $f^{**} = \overline{f}$.

Proof Since f is not identically ∞, the function f^* takes values in $(-\infty, \infty]$. Since the supremum of a collection of convex functions is convex and the supremum of a collection of lower semi-continuous functions is lower semi-continuous, we see that f^*, being the supremum of a collection of affine functions, is convex and lower semi-continuous. This proves that $f^* \in \mathrm{Conv}(\mathbb{R}^d)$.

Since $\langle x^*, x\rangle - f(x) \geq \langle x^*, x\rangle - c$ for each $(x, c) \in \mathcal{E}(f)$, it is clear that

$$f^*(x^*) := \sup_{x\in\mathbb{R}^d} \big[\langle x^*, x\rangle - f(x)\big] = \sup_{(x,c)\in\mathcal{E}(f)} \big[\langle x^*, x\rangle - c\big],$$

which proves (2.8) (i). We next observe that

$$\begin{aligned}\mathcal{E}(f^*) &= \big\{(x^*, c^*) : c^* \geq \sup_{x\in\mathbb{R}^d} \big[\langle x^*, x\rangle - f(x)\big]\big\}\\ &= \big\{(x^*, c^*) : \langle x^*, x\rangle - c^* \leq f(x)\ \forall x \in \mathbb{R}^d\big\},\end{aligned}$$

which proves (2.9) (ii). This in turn implies

$$\overline{f}(x) = \sup_{(x^*,c^*)\in\mathcal{E}(f^*)} \big[\langle x^*, x\rangle - c^*\big],$$

which proves (2.8) (ii). We postpone the proof of (2.9) (i) and first prove the remaining statements.

Since $\langle x^*, x\rangle - c^* \leq f(x)\ \forall x \in \mathbb{R}^d$ if and only if $\langle x^*, x\rangle - c^* \leq \overline{f}(x)\ \forall x \in \mathbb{R}^d$, formula (2.9) (ii) shows that $\mathcal{E}(f^*) = \mathcal{E}((\overline{f})^*)$ and hence $f^* = (\overline{f})^*$.

If $f \in \mathrm{Conv}(\mathbb{R}^d)$ or equivalently $f = \overline{f}$, then (2.8) shows that f is defined in terms of f^* by exactly the same formula that defines f^* in terms of f, which proves that $f^{**} = f$. More generally, if $f : \mathbb{R}^d \to (-\infty, \infty]$ is not identically ∞, then we can apply what we have just proved to $\overline{f}$ to conclude that $f^{**} = ((\overline{f})^*)^* = \overline{f}$. Formula (2.9) (i) now follows by applying (2.9) (ii) to f^*. ∎

The following lemma says that the generalized gradient of f^* is the inverse of the gradient of f. The relation $\langle x^*, x\rangle = f(x) + f^*(x^*)$ is demonstrated in Figure 2.2. See also Figure 2.3 for an illustration of the Legendre transform of a non-smooth function.

Lemma 2.14 (Slope of the Legendre transform) *For any $f \in \mathrm{Conv}(\mathbb{R}^d)$ and $x, x^* \in \mathbb{R}^d$, one has*

$$\langle x^*, x\rangle \leq f(x) + f^*(x^*) \tag{2.10}$$

Moreover,

$$(x, x^*) \in Df \quad \Leftrightarrow \quad \langle x^*, x\rangle = f(x) + f^*(x^*) \quad \Leftrightarrow \quad (x^*, x) \in Df^*.$$

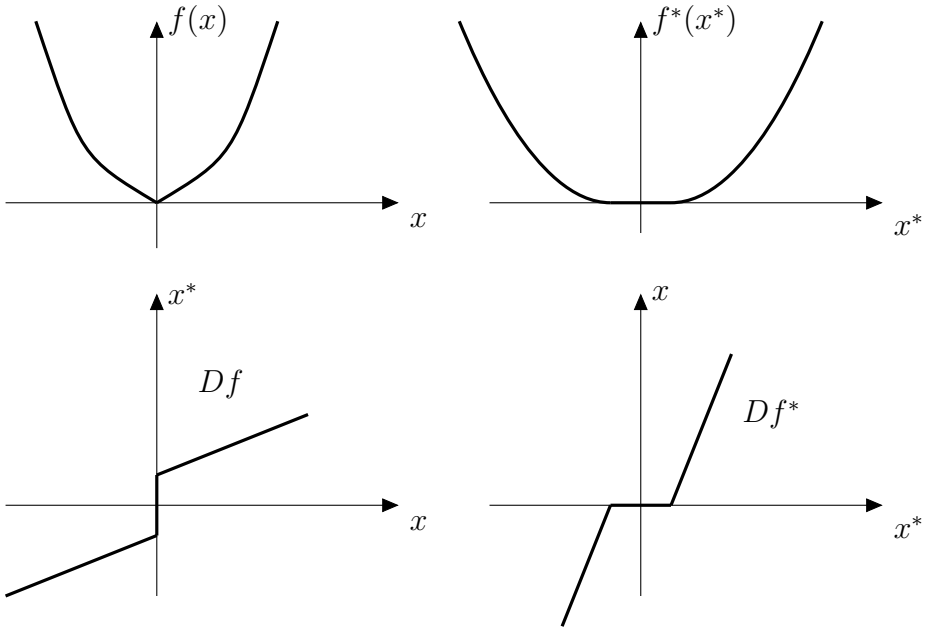

Figure 2.3: Legendre transform of a non-smooth function.

Proof The inequality (2.10) follows immediately from the definition $f^*(x^*) = \sup_{x\in\mathbb{R}^d}\big[\langle x^*, x\rangle - f(x)\big]$. Assume that $(x, x^*) \in Df$. Then there exists a $c^* \in \mathbb{R}$ such that $\langle x^*, x\rangle - c^* = f(x)$ and $x^*y - c^* \leq f(y)$ for all $y \in \mathbb{R}^d$. By (2.9) (ii), this implies that $(x^*, c^*) \in \mathcal{E}(f^*)$. On the other hand, since $\langle x^*, x\rangle - c^* = f(x)$, for each $\varepsilon > 0$ it is not true that $x^*y - c^* + \varepsilon \leq f(y)$ for all $y \in \mathbb{R}^d$, which again by (2.9) (ii) implies that $(x^*, c^* - \varepsilon) \notin \mathcal{E}(f^*)$ for all $\varepsilon > 0$ and hence $c^* = f^*(x^*)$ and $\langle x^*, x\rangle = f(x) + f^*(x^*)$.

Assume, conversely, that $\langle x^*, x\rangle = f(x) + f^*(x^*)$. Trivially $\big(x^*, f^*(x^*)\big) \in \mathcal{E}(f^*)$ so (2.9) (ii) implies that $x^*y - f^*(x^*) \leq f(y)$ for all $y \in \mathbb{R}^d$. Since moreover $\langle x^*, x\rangle - f^*(x^*) = f(x)$, this proves that the affine function $x \mapsto \langle x^*, x\rangle - f^*(x^*)$ is supporting at x and hence $(x, x^*) \in Df$.

This proves that $(x, x^*) \in Df$ if and only if $\langle x^*, x\rangle = f(x) + f^*(x^*)$. By symmetry, reversing the roles of x and x^* and of f and f^*, this is in turn equivalent to $(x^*, x) \in Df^*$. ∎

2.6 Extensions of convex functions

For any $g \in \mathrm{Conv}(\mathbb{R}^d)$ and closed convex set $D \subset \mathbb{R}^d$, setting $f(x) := g(x)$ for $x \in D$ and $:= \infty$ otherwise defines a function $f \in \mathrm{Conv}(\mathbb{R}^d)$. In such a situation, we that f is the *restriction* of g to D and that g *extends* f. We say that f is a restriction of g if there exists a closed convex D such that f is the restriction of g to D.

Lemma 2.15 (Restriction of a convex function) *Let $f, g \in \mathrm{Conv}(\mathbb{R}^d)$ and assume that $\mathcal{U}_f$ is nonempty. Then f is a restriction of g if and only if $f(x) = g(x)$ for all $x \in \mathcal{U}_f$.*

Proof The condition is clearly necessary. To prove sufficiency, we observe that since $\mathcal{U}_f$ is nonempty, $\mathcal{D}_f \subset \overline{\mathcal{U}}_f$. Let $x \in \overline{\mathcal{U}}_f \backslash \mathcal{U}_f$ and $\mathcal{U}_f \ni x_n \to x$. If $f(x) < \infty$, then the lower semi-continuity of g and the fact that f is continuous on $\mathcal{D}_f$ imply that

$$g(x) \leq \lim_{n\to\infty} f(x_n) = f(x),$$

and this inequality also trivially holds if $f(x) = \infty$. The inequality $g(x) \leq f(x)$ cannot be strict since this would contradict the fact that g is continuous on $\mathcal{D}_g$, so we conclude that $f(x) = g(x)$ for all $x \in \overline{\mathcal{U}}_f$. Setting $D := \overline{\mathcal{U}}_f$, it follows that $f(x) = g(x)$ for all $x \in D$ and $= \infty$ otherwise, i.e., f is a restriction of g. ∎

Let $f \in \mathrm{Conv}(\mathbb{R}^d)$ and assume that $\mathcal{U}_f \neq \emptyset$. By definition, we say that a f is *on maximal domain* if it satisfies the equivalent conditions (i) and (ii) of the following lemma.

Lemma 2.16 (Convex functions on maximal domain) *Let $f \in \mathrm{Conv}(\mathbb{R}^d)$ and assume that $\mathcal{U}_f \neq \emptyset$. Let $D_\circ f := \big\{(x, x^*) \in Df : x \in \mathcal{U}_f\big\}$. Then of the following conditions, (i) and (ii) are equivalent and imply (iii).*

(i) *$D_\circ f$ is a closed subset of $\mathbb{R}^{2d}$,*

(ii) *$D_\circ f = Df$,*

(iii) *$f = g$ for all $g \in \mathrm{Conv}(\mathbb{R}^d)$ that extend f.*

Proof The implication (ii)⇒(i) follows from the fact that by Proposition 2.7 (c), Df is a closed subset of $\mathbb{R}^{2d}$. Assume, conversely, that $D_\circ f$ is a closed subset of $\mathbb{R}^{2d}$.

Then $D_\circ f$ is a closed subset of Df. On the other hand, since $D_\circ f = Df \cap (\mathcal{U}_f \times \mathbb{R}^d)$, it is also open as a subset of Df. By Proposition 2.7 (d), Df is connected so aso using the fact that $D_\circ f \neq \emptyset$ by Proposition 2.7 (a) we see that $D_\circ f = Df$. This proves the implication (i)⇒(ii).

To complete the proof, we need to show that (ii)⇒(iii). Assume that f satisfies (ii) and that there exists a $g \in \mathrm{Conv}(\mathbb{R}^d)$ with $g \neq f$ that extends f. Then $\mathcal{U}_g \backslash D$ is nonempty and hence there exists an $x \in \mathcal{U}_g$ that lies on the boundary of D. By Proposition 2.7 (a), there exists a supporting affine function for g at x. Since this is also a supporting affine function for f at x, we conclude that there exists an $(x, x^*) \in Df$ for which $x \notin \mathcal{U}_f$, contradicting (ii). ∎

Remark I conjecture that the conditions (i)–(iii) are in fact all equivalent, but I have not found a reference (though the claim, if it is true, is probably known). For any $f \in \mathrm{Conv}(\mathbb{R}^d)$ one can define

$$h(y) := \sup_{(x,x^*) \in D_\circ f} \big[f(x) + \langle x^*, y - x \rangle \big] \qquad (y \in \mathbb{R}^d).$$

Since h is the supremum of all affine functions that support f in some point $x \in \mathcal{U}_f$, we see that h is lower semi-continuous and convex and that $h(x) = f(x)$ for all $x \in \mathcal{U}_f$, which by Lemma 2.15 implies that h extends f. In fact, it is easy to see that each extension g of f must satisfy $h \leq g \leq f$, so (iii) is equivalent to the statement that $f = h$. To prove the conjecture, one would have to show that if (i) and (ii) fail, then $h \neq f$. Since this is not completely straightforward we will not need this in what follows, we will skip this.

2.7 Well-behaved convex functions

Recall the definition of strict convexity from Section 2.2 and the definition of "maximal domain" just above Lemma 2.16. For each $n \in \mathbb{N}_+ \cup \{\infty\}$, we let $\mathrm{Conv}_n(\mathbb{R}^d)$ denote the class of functions $f \in \mathrm{Conv}(\mathbb{R}^d)$ such that

(i) $\mathcal{U}_f \neq \emptyset$,

(ii) f is n times continuously differentiable on $\mathcal{U}_f$,

(iii) f is strictly convex on $\mathcal{U}_f$,

(iv) f is on maximal domain.

For a twice continuously differentiable function f, we let $\partial^2 f(x)$ defined as

$$\partial^2_{ij} f(x) := \tfrac{\partial^2}{\partial x(i)\partial x(j)} f(x)$$

denote the matrix of its second derivatives. Then

$$\begin{aligned}\tfrac{\partial^2}{\partial \varepsilon^2} f(x+\varepsilon y)\big|_{\varepsilon=0} &= \sum_{i=1}^d y(i) \tfrac{\partial}{\partial x(i)} \Big(\sum_{j=1}^d y(j) \tfrac{\partial}{\partial x(j)} f(x) \Big) \\ &= \sum_{i,j=1}^d y(i) \partial^2_{ij} f(x) y(j) = \langle y, \partial^2 f(x) y \rangle.\end{aligned}$$

For the spaces $\mathrm{Conv}_n(\mathbb{R}^d)$ with $n \geq 2$, an equivalent formulation of condition (iii) is

(iii)' $\langle y, \partial^2 f(x) y \rangle > 0$ for all $x \in \mathcal{U}_f$ and $y \in \mathbb{R}^d \backslash \{0\}$,

which says that the symmetric matrix $\partial^2 f(x)$ is strictly positive definite for all $x \in \mathcal{U}_f$. The condition that f is on maximal domain can alternatively be formulated as

(iv)' $\big|\partial f(x_n)\big| \underset{n\to\infty}{\longrightarrow} \infty$ whenever $\mathcal{U}_f \ni x_n \underset{n\to\infty}{\longrightarrow} x \in \mathbb{R}^d \backslash \mathcal{U}_f$.

Indeed, if (iv)' holds, then $\{(x, x^*) \in Df : x \in \mathcal{U}_f\}$ is a closed subset of Df and hence, by Proposition 2.7 (c), also of $\mathbb{R}^{2d}$. On the other hand, if (iv)' fails, then by going to a subsequence we can find $\mathcal{U}_f \ni x_n \to x \in \mathbb{R}^d \backslash \mathcal{U}_f$ such that $\partial f(x_n)$ converges to a finite limit $x^* \in \mathbb{R}^d$. Since Df is closed by Proposition 2.7 (c), it follows that $(x, x^*) \in Df$ which contradicts condition (ii) of Lemma 2.16.

The following proposition links the spaces $\mathrm{Conv}_n(\mathbb{R}^d)$ to the Legendre transform. There does not seem to exist established terminology for these spaces but we will sometimes call elements of $\mathrm{Conv}_1(\mathbb{R}^d)$ *well-behaved*. Recall that a homeomorphism is a bijection ϕ such that both ϕ and ϕ^{-1} are continuous functions.

Proposition 2.17 (Well-behaved functions) *Let $n \geq 1$ and $f \in \mathrm{Conv}_n(\mathbb{R}^d)$. Then:*

(a) $f^* \in \mathrm{Conv}_n(\mathbb{R}^d)$.

(b) $\partial f : \mathcal{U}_f \to \mathcal{U}_{f^*}$ *is a homeomorphism, and* $(\partial f)^{-1} = \partial f^*$.

(c) *For each* $x^* \in \mathcal{U}_{f^*}$*, the function* $y \mapsto \langle x^*, y\rangle - f(y)$ *assumes its maximum in the unique point* $x = \partial f^*(x^*)$.

Proof We start by proving (b). Using Lemma 2.16 and the assumptions that f is on maximal domain and continuously differentiable, we see that

$$Df = \big\{(x, x^*) \in Df : x \in \mathcal{U}_f\big\} = \big\{\big(x, \partial f(x)\big) \in Df : x \in \mathcal{U}_f\big\}.$$

Since f is strictly convex, the function ∂f is one-to-one. Since Df is closed by Proposition 2.7 (c), the closed graph theorem tells us that ∂f is a homeomorphism from $\mathcal{U}_f$ to some open set $O \subset \mathbb{R}^d$. By Proposition 2.7 (a) and Lemma 2.14, $\mathcal{U}_{f^*} \subset O \subset \mathcal{D}_{f^*}$ and hence $O = \mathcal{U}_{f^*}$ since O is open. Lemma 2.14 now tells us that

$$Df^* = \{(x^*, (\partial f)^{-1}(x^*)) : x^* \in \mathcal{U}_{f^*}\},$$

so Lemma 2.8 allows us to conclude that f^* is continuously differentiable and $(\partial f)^{-1} = \partial f^*$, concluding the proof of part (b).

We next prove (a). Since $\partial f : \mathcal{U}_f \to \mathcal{U}_{f^*}$ is a bijection, clearly $\mathcal{U}_f \neq \emptyset$ implies $\mathcal{U}_{f^*} \neq \emptyset$. Since ∂f is strictly increasing and $n-1$ times continuously differentiable, its inverse $(\partial f)^{-1} = \partial f^*$ has the same properties and hence f^* is strictly convex and n times continuously differentiable on $\mathcal{U}_{f^*}$. By Lemma 2.14, the fact that f is continuously differentiable and on maximal domain, and part (b), we see that

$$\begin{aligned} Df^* &= \big\{(x^*, x) : (x, x^*) \in Df\big\} \\ &= \big\{\big(\partial f(x), x\big) : x \in \mathcal{U}_f\big\} = \big\{\big(x^*, \partial f^*(x^*)\big) : x^* \in \mathcal{U}_{f^*}\big\}, \end{aligned}$$

which proves that f^* is on maximal domain.

It remains to prove (c). Since $\partial f : \mathcal{U}_f \to \mathcal{U}_{f^*}$ is a bijection, $\mathcal{U}_{f^*}$ is the range of values that $\partial f(x)$ can take. Since $\mathcal{U}_{f^*}$ is an open convex set, it follows that the strictly concave function $y \mapsto \langle x^*, y\rangle - f(y)$ is increasing for small values of y, decreasing for large values of y, and assumes its maximum in the point x that is uniquely characterized by the condition that $x^* = \partial f(x)$. Since ∂f^* is the inverse of ∂f, this proves (c). ∎

Chapter 3

Sums of i.i.d. random variables

3.1 Cramér's rate function

After two chapters of abstract real analysis we are now slowly returning to probability theory and preparing to prove our first large deviation principles. We start by studying the rate function in Cramér's theorem (Theorem 0.1). Because we will need this in what follows, we will generalize somewhat and immediately state our results in the multi-dimensional case.

For any probability measure μ on $\mathbb{R}^d$ which has at least finite first, respectively second moments, we let

$$\begin{aligned}\langle\mu\rangle(i) &:= \int \mu(\mathrm{d}x)x(i),\\ \mathrm{Cov}_{ij}(\mu) &:= \int \mu(\mathrm{d}x)x(i)x(j) - \Big(\int \mu(\mathrm{d}x)x(i)\Big)\Big(\int \mu(\mathrm{d}x)x(j)\Big)\end{aligned}$$

$(i,j=1,\ldots,d)$ denote the *mean* and *covariance matrix* of μ. The class $\mathrm{Conv}_\infty(\mathbb{R}^d)$ mentioned below is defined in Section 2.7.

Lemma 3.1 (Smoothness of logarithmic moment generating function) *Let μ be a probability measure on $\mathbb{R}^d$ and let Z be given by*

$$Z(\lambda) := \int e^{\langle\lambda,\,x\rangle}\mu(\mathrm{d}x) \qquad (\lambda\in\mathbb{R}^d). \tag{3.1}$$

Assume that $Z(\lambda)<\infty$ for all $\lambda\in\mathbb{R}^d$ and for $\lambda\in\mathbb{R}$, let μ_λ denote the tilted law

$$\mu_\lambda(\mathrm{d}x) := \frac{1}{Z(\lambda)}e^{\langle\lambda,\,x\rangle}\mu(\mathrm{d}x) \qquad (\lambda\in\mathbb{R}^d). \tag{3.2}$$

Then $\lambda \mapsto \log Z(\lambda)$ *is infinitely differentiable and*

$$\left.\begin{array}{ll} \text{(i)} & \frac{\partial}{\partial\lambda(i)} \log Z(\lambda) = \langle \mu_\lambda \rangle(i), \\ \text{(ii)} & \frac{\partial^2}{\partial\lambda(i)\partial\lambda(j)} \log Z(\lambda) = \mathrm{Cov}_{ij}(\mu_\lambda) \end{array}\right\} \quad (\lambda \in \mathbb{R}^d,\ i,j = 1,\dots,d).$$

In particular, if $\langle y, \mathrm{Cov}(\mu) y\rangle > 0$ *for all* $y \in \mathbb{R}^d \backslash \{0\}$*, then* $\log Z \in \mathrm{Conv}_\infty(\mathbb{R}^d)$.

Proof We only give the proof in the one-dimensional case. The proof in the multidimensional case is basically the same but notationally more complicated. We claim that $\lambda \mapsto Z(\lambda)$ is infinitely differentiable and

$$\left(\tfrac{\partial}{\partial\lambda}\right)^n Z(\lambda) = \int x^n e^{\lambda x} \mu(\mathrm{d}x).$$

To justify this, we must show that the interchanging of differentiation and integral is allowed. By symmetry, it suffices to prove this for $\lambda \geq 0$. We observe that

$$\tfrac{\partial}{\partial\lambda} \int x^n e^{\lambda x} \mu(\mathrm{d}x) = \lim_{\varepsilon\to 0} \int x^n \varepsilon^{-1} (e^{(\lambda+\varepsilon)x} - e^{\lambda x}) \mu(\mathrm{d}x),$$

where

$$|x|^n \varepsilon^{-1} \left| e^{(\lambda+\varepsilon)x} - e^{\lambda x} \right| = |x|^n \left| \varepsilon^{-1} \int_\lambda^{\lambda+\varepsilon} x e^{\kappa x} \mathrm{d}\kappa \right| \leq |x|^{n+1} e^{(\lambda+1)x} \qquad (x \in \mathbb{R},\ \varepsilon \leq 1).$$

It follows from the existence of all exponential moments that this function is integrable, hence we may use dominated convergence to interchange the limit and integral.

It follows that

$$\begin{aligned} \text{(i)} \quad & \tfrac{\partial}{\partial\lambda} \log Z(\lambda) = \tfrac{\partial}{\partial\lambda} \log \int e^{\lambda x} \mu(\mathrm{d}x) = \frac{\int x e^{\lambda x} \mu(\mathrm{d}x)}{\int e^{\lambda x} \mu(\mathrm{d}x)} = \langle \mu_\lambda \rangle, \\ \text{(ii)} \quad & \tfrac{\partial^2}{\partial\lambda^2} \log Z(\lambda) = \frac{Z(\lambda) \int x^2 e^{\lambda x} \mu(\mathrm{d}x) - (\int x e^{\lambda x} \mu(\mathrm{d}x))^2}{Z(\lambda)^2} \\ & \qquad = \int x^2 \mu_\lambda(\mathrm{d}x) - \Big(\int x \mu_\lambda(\mathrm{d}x) \Big)^2 = \mathrm{Var}(\mu_\lambda). \end{aligned} \tag{3.3}$$

Since $\log Z(\lambda)$ is finite for all $\lambda \in \mathbb{R}$ it satisfies conditions (i) and (iv) of the definition of Conv_∞, and since it is infinitely differentiable it satisfies condition (ii) too. If $\mathrm{Var}(\mu) > 0$, then $\mathrm{Var}(\mu_\lambda) > 0$ for all λ and hence $\frac{\partial^2}{\partial\lambda^2} \log Z(\lambda) > 0$, showing that $\log Z$ is strictly convex. In the multidimensional case, the condition $\mathrm{Var}(\mu) > 0$ must be replaced by strict positive definiteness of the covariance matrix. ∎

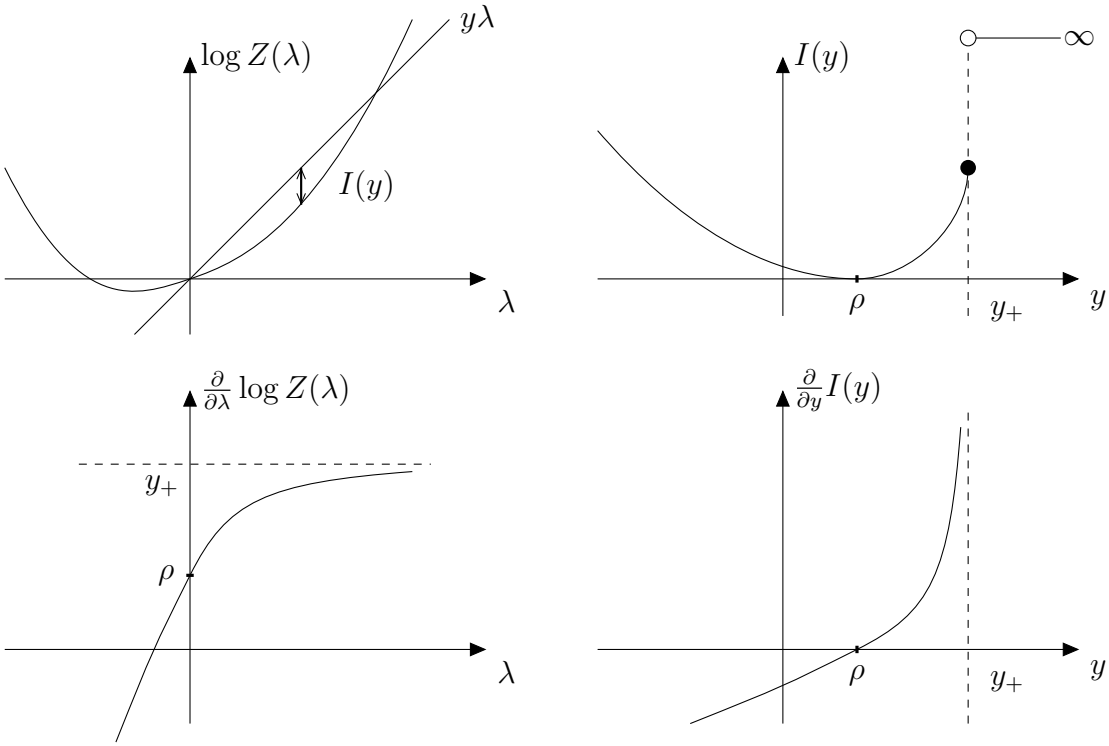

Figure 3.1: Definition of the rate function in Cramér's theorem. The functions below are derivatives of the functions above, and inverses of each other.

Exercise 3.2 (Maximal and minimal mean of tilted law) Let μ be a probability law on $\mathbb{R}$ such that $\int e^{\lambda x}\mu(\mathrm{d}x) < \infty$ for all $\lambda \in \mathbb{R}$ and let μ_λ be defined as in Lemma 3.1. Show that

$$\lim_{\lambda\to-\infty}\langle\mu_\lambda\rangle = y_- \quad \text{and} \quad \lim_{\lambda\to+\infty}\langle\mu_\lambda\rangle = y_+,$$

where $y_- := \inf(\text{support}(\mu))$, $y_+ := \sup(\text{support}(\mu))$.

We next turn our attention to the Legendre transform I of $\log Z$, which plays the role of the rate function in Cramér's theorem. The following lemma lists some properties of the function I that is the Legendre transform of $\log Z$. See Figure 3.1 for an illustration.

Lemma 3.3 (Properties of the rate function) *Let μ be a probability measure on $\mathbb{R}^d$. Assume that the moment generating function Z defined in (3.1) is finite*

for all $\lambda \in \mathbb{R}^d$ *and that*

$$\langle y, \mathrm{Cov}(\mu) y\rangle > 0 \qquad (0 \neq y \in \mathbb{R}^d).$$

For $\lambda \in \mathbb{R}^d$*, define* μ_λ *as in (3.2) and let* $\langle \mu\rangle$ *resp.* $\langle \mu_\lambda\rangle$ *be the mean of* μ *and* μ_λ*. Let* $I : \mathbb{R}^d \to (-\infty, \infty]$ *be the Legendre transform of* $\log Z$*. Then:*

(i) $I \in \mathrm{Conv}_\infty(\mathbb{R}^d)$.

(ii) $I\big(\langle \mu\rangle\big) = 0$ *and* $I(y) > 0$ *for all* $y \neq \langle \mu\rangle$.

(iii) I *is a good rate function.*

(iv) $\mathcal{U}_I = \{\langle \mu_\lambda\rangle : \lambda \in \mathbb{R}^d\}$.

(v) $\overline{\mathcal{U}}_I$ *is the closed convex hull of* $\mathrm{support}(\mu)$.

(vi) *For each* $y_\circ \in \mathcal{U}_I$*, the function* $\mathbb{R}^d \ni \lambda \mapsto \langle y_\circ, \lambda\rangle - \log Z(\lambda)$ *assumes its maximum in a unique point* $\lambda_\circ \in \mathbb{R}^d$*, which is uniquely characterized by the requirement that* $\langle \mu_{\lambda_\circ}\rangle = y_\circ$.

Proof The fact that $I \in \mathrm{Conv}_\infty(\mathbb{R}^d)$ follows from Proposition 2.17 (a) and the fact that $\log Z \in \mathrm{Conv}(\mathbb{R}^d)$, which follows from Lemma 3.1 and the assumption that the matrix $\mathrm{Cov}(\mu)$ is strictly positive.

It is immediate from the definition of $Z(\lambda)$ that $Z(0) = 1$ and hence $\log Z(0) = 0$. Since I is the Legendre transform of $\log Z$, Lemma 2.13 tells us that $\log Z$ is the Legendre transform of I. In particular, this shows that

$$0 = \log Z(0) = \sup_{y\in\mathbb{R}} \big[\langle 0, y\rangle - I(y)\big] = -\inf_{y\in\mathbb{R}} I(y),$$

proving that $I \geq 0$. By Lemma 3.1, $\partial \log Z(0) = \langle \mu\rangle =: \rho$, which means that $\lambda \mapsto \langle \rho, \lambda\rangle$ is a supporting affine function to $\log Z$ at the point $\lambda = 0$ and hence

$$I(\rho) = \sup_{\lambda\in\mathbb{R}} \big[\langle \rho, \lambda\rangle - \log Z(\lambda)\big] = 0.$$

Since $I \in \mathrm{Conv}_\infty(\mathbb{R})$, it is strictly convex, so $I(y) > 0$ for all $y \neq \rho$, proving part (ii).

Since $I \in \mathrm{Conv}_\infty(\mathbb{R}^d)$, it is lower semi-continuous, while part (ii) and the convexity of I imply that the level sets of I are bounded, hence I is a good rate function.

Property (iv) is immediate from Proposition 2.17 (b) and Lemma 3.1. Proposition 2.17 (c) moreover tells us that for each $y_\circ \in \mathcal{U}_I$, the function $\mathbb{R}^d \ni \lambda \mapsto \langle y_\circ, \lambda\rangle - \log Z(\lambda)$ assumes its maximum in a unique point $\lambda_\circ \in \mathbb{R}^d$, which is given by $\lambda_\circ = \partial I(y_\circ)$. By Proposition 2.17 (b), the function $\lambda \mapsto \partial \log Z(\lambda)$ is the inverse of $y \mapsto \partial I(y)$, so the condition $\lambda_\circ = \partial I(y_\circ)$ is equivalent to $\partial \log Z(\lambda_\circ) = y_\circ$. By Lemma 3.1, this says that $\langle \mu_{\lambda_\circ}\rangle = y_\circ$, proving (vi).

It remains to prove (v). Since $\text{support}(\mu_\lambda) = \text{support}(\mu)$ for all $\lambda \in \mathbb{R}^d$, it is easy to see that if H is an open half-space such that $H \cap \text{support}(\mu) = \emptyset$, then $\langle \mu_\lambda\rangle \notin H$. Since by (2.1), the complement of $\overline{\mathrm{C}}(\text{support}(\mu))$ is the union of all open half-spaces that do not intersect $\text{support}(\mu)$, this proves the inclusion $\mathcal{U}_I \subset \overline{\mathrm{C}}(\text{support}(\mu))$.

On the other hand, if $H = \{y \in \mathbb{R}^d : \langle \lambda, y\rangle > c\}$ is an open half-space such that $H \cap \text{support}(\mu) \neq \emptyset$, then, in the same way as in Exercise 3.2, one can check that there exists some $r > 0$ large enough such that $\langle \mu_{r\lambda}\rangle \in H$. This proves that $\overline{\mathrm{C}}(\mathcal{U}_I) \supset \overline{\mathrm{C}}(\text{support}(\mu))$. Since I is convex, so is $\mathcal{U}_I$, and therefore the closed convex hull of $\mathcal{U}_I$ is just the closure of $\mathcal{U}_I$. Thus, we have $\overline{\mathcal{U}}_I \supset \overline{\mathrm{C}}(\text{support}(\mu))$, completing our proof. ∎

We also provide the proof of Lemma 0.2 from the introduction, which gives a bit more detail than Lemma 3.3 in the one-dimensional case.

Proof of Lemma 0.2 Properties (i), (ii), (vi) follow from Lemma 3.3 (i), properties (iii) and (iv) follow from Lemma 3.3 (ii), and property (v) follows from Lemma 3.3 (vi).

Property (vii) follows from the fact that, by Proposition 2.17 (b), $\partial I : (y_-, y_+) \to \mathbb{R}$ is a bijection. The fact that $I'' > 0$ on $\mathcal{U}_I$ follows from the fact that $I \in \text{Conv}$. We recall that if f is smooth and strictly increasing and $f(x) = y$, then $\frac{\partial}{\partial x} f(x) = 1/(\frac{\partial}{\partial y} f^{-1}(y))$. Therefore, Proposition 2.17 (b), the fact that $\partial \log Z(0) = \rho$, and Lemma 3.1 imply that $\partial^2 I(\rho) = 1/(\partial^2 \log Z(0)) = 1/\sigma^2$, proving part (viii).

To prove part (ix), finally, by symmetry it suffices to prove the statement for y_+. If $y_+ < \infty$, then

$$\begin{aligned} e^{-I(y_+)} &= \inf_{\lambda\in\mathbb{R}} \big[e^{\log Z(\lambda) - y_+\lambda}\big] = \inf_{\lambda\in\mathbb{R}} e^{-y_+\lambda} Z(\lambda) \\ &= \inf_{\lambda\in\mathbb{R}} e^{-y_+\lambda} \int e^{\lambda y} \mu(\mathrm{d}y) = \inf_{\lambda\in\mathbb{R}} \int e^{\lambda(y\,-\,y_+)} \mu(\mathrm{d}y) \\ &= \lim_{\lambda\to\infty} \int e^{\lambda(y\,-\,y_+)} \mu(\mathrm{d}y) = \mu(\{y_+\}), \end{aligned}$$

which completes our proof. ∎

3.2 Cramér's theorem

We are finally ready to leave the abstract theory and prove our first "real" large deviations result, which is Cramér's theorem. We could (as Cramér did in 1938) have proved this result by more elementary means and indeed the proof below does by far not use all the abstract theory that we have developed so far, but we will soon move on to the Gärtner-Ellis theorem, for which we will be able to give a very short and elegant proof that however essentially depends on exponential tightness and fairly subtle results from convex analysis like Lemmas 2.16 and 2.10.

Proof of Theorem 0.1 By symmetry, it suffices to prove (0.2) (i). In view of the fact that $1_{[0,\infty)}(z) \leq e^z$, we have, for each $y \in \mathbb{R}$ and $\lambda \geq 0$,

$$\begin{aligned}
&\mathbb{P}\big[\frac{1}{n}\sum_{k=1}^n X_k \geq y\big] = \mathbb{P}\big[\frac{1}{n}\sum_{k=1}^n (X_k - y) \geq 0\big] = \mathbb{P}\big[\lambda\sum_{k=1}^n (X_k - y) \geq 0\big] \\
&\leq \mathbb{E}\big[e^{\lambda\sum_{k=1}^n (X_k - y)}\big] = \prod_{k=1}^n \mathbb{E}\big[e^{\lambda(X_k - y)}\big] = e^{-n\lambda y}\mathbb{E}\big[e^{\lambda X_1}\big]^n \\
&= e^{(\log Z(\lambda) - \lambda y)n}.
\end{aligned}$$

If $y > \rho$, then, by Lemma 3.1, $\frac{\partial}{\partial\lambda}[\log Z(\lambda) - \lambda y]|_{\lambda=0} = \rho - y < 0$, so, by the convexity of the function $\lambda \mapsto [\log Z(\lambda) - \lambda y]$,

$$\inf_{\lambda\geq 0}[\log Z(\lambda) - \lambda y] = \inf_{\lambda\in\mathbb{R}}[\log Z(\lambda) - \lambda y] =: -I(y).$$

Together with our previous formula, this shows that

$$\mathbb{P}\big[\frac{1}{n}\sum_{k=1}^n X_k \geq y\big] \leq e^{-nI(y)} \qquad (y > \rho),$$

and hence, in particular,

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbb{P}\big[T_n \geq y\big] \leq -I(y) \qquad (y > \rho).$$

To estimate the limit inferior from below, we distinguish three cases. If $y > y_+$, then $\mathbb{P}[T_n \geq y] = 0$ for all $n \geq 1$ while $I(y) = \infty$ by Lemma 0.2 (v), so (0.2) (i) is trivially fulfilled. If $y = y_+$, then $\mathbb{P}[T_n \geq y] = \mathbb{P}[X_1 = y_+]^n$ while $I(y_+) = -\log \mathbb{P}[X_1 = y_+]$ by Lemma 0.2 (ix), hence again (0.2) (i) holds.

If $y < y_+$, finally, then by Proposition 2.17 (b) and (c), $I(y) = \sup_{\lambda\in\mathbb{R}}[y\lambda - \log Z(\lambda)] = y\lambda_\circ - \log Z(\lambda_\circ)$, where $\lambda_\circ = (\partial \log Z)^{-1}(y)$. In other words, recalling

Lemma 3.1, this says that $\lambda_\circ$ is uniquely characterized by the requirement that

$$\langle \mu_{\lambda_\circ} \rangle = \partial \log Z(\lambda_\circ) = y.$$

We observe that if $(\hat{X}_k)_{k\geq 1}$ are i.i.d. random variables with common law $\mu_{\lambda_\circ}$, and $\hat{T}_n := \frac{1}{n}\sum_{k=1}^n \hat{X}_k$, then $\lim_{n\to\infty} \mathbb{P}[\hat{T}_n \geq y] = \frac{1}{2}$ by the central limit theorem and therefore $\lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}[\hat{T}_n \geq y] = 0$. The idea of the proof is to replace the law μ of the $(X_k)_{k\geq 1}$ by $\mu_{\lambda_\circ}$ at an exponential cost of size $I(y)$. More precisely, we estimate

$$\begin{aligned}
\mathbb{P}\big[T_n \geq y\big] &= \mathbb{P}\big[\sum_{k=1}^n (X_k - y) \geq 0\big] = \int \mu(\mathrm{d}x_1)\cdots\int \mu(\mathrm{d}x_n) 1_{\{\sum_{k=1}^n (x_k - y) \geq 0\}} \\
&= Z(\lambda_\circ)^n \int e^{-\lambda_\circ x_1}\mu_{\lambda_\circ}(\mathrm{d}x_1)\cdots\int e^{-\lambda_\circ x_n}\mu_{\lambda_\circ}(\mathrm{d}x_n) 1_{\{\sum_{k=1}^n (x_k - y) \geq 0\}} \\
&= Z(\lambda_\circ)^n e^{-n\lambda_\circ y}\int \mu_{\lambda_\circ}(\mathrm{d}x_1)\cdots\int \mu_{\lambda_\circ}(\mathrm{d}x_n) \\
&\qquad\qquad \times e^{-\lambda_\circ \sum_{k=1}^n (x_k - y)} 1_{\{\sum_{k=1}^n (x_k - y) \geq 0\}} \\
&= e^{-nI(y)}\mathbb{E}\big[e^{-n\lambda_\circ(\hat{T}_n - y)} 1_{\{\hat{T}_n - y \geq 0\}}\big].
\end{aligned} \tag{3.4}$$

By the central limit theorem,

$$\mathbb{P}\big[y \leq \hat{T}_n \leq y + \sigma n^{-1/2}\big] \underset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}}\int_0^1 e^{-z^2/2}\mathrm{d}z =: \theta > 0.$$

Since

$$\mathbb{E}\big[e^{-n\lambda_\circ(\hat{T}_n - y)} 1_{\{\hat{T}_n - y \geq 0\}}\big] \geq \mathbb{P}\big[y \leq \hat{T}_n \leq y + \sigma n^{-1/2}\big] e^{-\sqrt{n}\sigma\lambda_\circ},$$

this implies that

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{E}\big[e^{-n\lambda_\circ(\hat{T}_n - y)} 1_{\{\hat{T}_n - y \geq 0\}}\big] \\
&\quad \geq \liminf_{n\to\infty} \frac{1}{n}\log\big(\theta e^{-\sqrt{n}\sigma\lambda_\circ}\big) = -\liminf_{n\to\infty}\frac{1}{n}\big(\log\theta + \sqrt{n}\sigma\lambda_\circ\big) = 0.
\end{aligned}$$

Inserting this into (3.4) we find that

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big[T_n \geq y\big] \geq -I(y) \qquad (y > \rho).$$

∎

Remark Our proof of Cramér's theorem actually shows that for any $\rho < y < y_+$,

$$e^{-nI(y)} - O(\sqrt{n}) \leq \mathbb{P}[T_n \geq y] \leq e^{-nI(y)} \qquad \text{as } n \to \infty.$$

Here the term of order $\sqrt{n}$ in the lower bound comes from the central limit theorem. A simpler method to obtain a more crude lower bound is to use the weak law of large numbers instead. For each $\lambda_* > \lambda_\circ$, the calculation in (3.4) shows that

$$\mathbb{P}[T_n \geq y] = e^{-n[\lambda_* y - \log Z(\lambda_*)]}\mathbb{E}\big[e^{-n\lambda_*(\hat{T}_n - y)}1_{\{\hat{T}_n - y \geq 0\}}\big],$$

where $\hat{T}_n$ now denotes the mean of n i.i.d. random variables with common law μ_{λ_*}, instead of $\mu_{\lambda_\circ}$. Let $\varepsilon := \langle \mu_{\lambda_*} \rangle - \langle \mu_{\lambda_\circ} \rangle = \langle \mu_{\lambda_*} \rangle - y$. By the weak law of large numbers

$$\mathbb{P}\big[y \leq \hat{T}_n \leq y + 2\varepsilon\big] \underset{n\to\infty}{\longrightarrow} 1.$$

Inserting this into our previous formula yields

$$\mathbb{P}[T_n \geq y] \geq e^{-n[\lambda_* y - \log Z(\lambda_*)]}\, e^{-n2\varepsilon\lambda_*},$$

and hence

$$\liminf_{n\to\infty} \mathbb{P}[T_n \geq y] \geq \lambda_* y - \log Z(\lambda_*) - 2\varepsilon\lambda_*.$$

Since $\varepsilon \downarrow 0$ as $\lambda_* \downarrow \lambda_\circ$, taking the limit, we obtain that

$$\liminf_{n\to\infty} \mathbb{P}[T_n \geq y] \geq \lambda_\circ y - \log Z(\lambda_\circ) = I(y).$$

∎

Remark Using Theorem 0.1, it is not hard to show that indeed, the laws $\mathbb{P}[T_n \in \cdot\,]$ satisfy a large deviation principle with speed n and good rate function I. We will postpone this until we treat the multi-dimensional case in Theorem 3.5. Theorem 0.1 is in fact a bit stronger than the large deviation principle. Indeed, if $y_+ < \infty$ and $\mu(\{y_+\}) > 0$, then the large deviation principle tells us that

$$\limsup_{n\to\infty} \mu_n([y_+, \infty)) \leq -\inf_{y\in[y_+,\infty)} I(y) = -I(y_+),$$

but, as we have seen in Excercise 1.11, the complementary statement for the limit inferior does not follow from the large deviation principle since $[y_+, \infty)$ is not an open set.

Remark Let $\mathcal{U}_Z$ be the interior of the interval $\{\lambda \in \mathbb{R} : Z(\lambda) < \infty\}$. Theorem 0.1 remains true if the assumption that $\mathcal{U}_Z = \mathbb{R}$ is replaced by the weaker condition that $0 \in \mathcal{U}_Z$, see [DZ98, Section 2.2.1]. When we treat the multi-dimensional case in Theorem 3.5 below, we will prove Cramér's theorem under the assumptions that $0 \in \mathcal{U}_Z$ and $\log Z \in \mathrm{Conv}_1(\mathbb{R})$. This is more general than Theorem 0.1 but does not cover all cases where $0 \in \mathcal{U}_Z$.

Remark For $\rho < y < y_+$, it can be shown that for fixed $m \geq 1$,

$$\mathbb{P}\big[X_1 \in \mathrm{d}x_1, \ldots, X_m \in \mathrm{d}x_m \,\big|\, \tfrac{1}{n}\sum_{k=1}^n X_k \geq y\big] \underset{n\to\infty}{\Longrightarrow} \mu_{\lambda_\circ}(\mathrm{d}x_1)\cdots\mu_{\lambda_\circ}(\mathrm{d}x_m),$$

where μ_λ denotes a tilted law as in Lemma 3.1 and $\lambda_\circ$ is defined by the requirement that $\langle\mu_{\lambda_\circ}\rangle = y$. This means that conditioned on the rare event $\frac{1}{n}\sum_{k=1}^n X_k \geq y$, in the limit $n \to \infty$, the random variables $X_1, \ldots, X_n$ are approximately distributed as if they are i.i.d. with common law $\mu_{\lambda_\circ}$.

3.3 The Gärtner-Ellis theorem

In Theorem 3.5 below, we will see that Cramér's theorem generalizes to the multi-dimensional case in a more or less straightforward manner. We will also prove Theorem 0.4 about moderate deviations from the introduction. As preparation for both proofs, we will prove a theorem that is considerably more general and can also often be applied to dependent random variables. Below is a version of the Gärtner-Ellis theorem. The class $\mathrm{Conv}_1(\mathbb{R}^d)$ mentioned below is defined in Section 2.7.

Theorem 3.4 (Gärtner-Ellis) *Let μ_n be probability measures on $\mathbb{R}^2$ and let s_n be positive constants such that $s_n \to \infty$. Assume that for each $\lambda \in \mathbb{R}^d$, the limit*

$$\Gamma(\lambda) := \lim_{n\to\infty} \frac{1}{s_n} \log \int e^{s_n\langle\lambda, x\rangle}\mu_n(\mathrm{d}x) \tag{3.5}$$

exists in $[0, \infty]$ and that $\Gamma \in \mathrm{Conv}_1(\mathbb{R}^d)$ and $0 \in \mathcal{U}_\Gamma$. Then the measures μ_n satisfy the large deviation principle with speed s_n and good rate function I given by

$$I(x) := \sup_{\lambda\in\mathbb{R}^d} \big[\langle\lambda, x\rangle - \Gamma(\lambda)\big] \qquad (x \in \mathbb{R}^d).$$

Proof We start by proving exponential tightness. For each $0 \neq \lambda \in \mathbb{R}^d$ and $c > 0$, let $H_{\lambda,c}$ denote the half-space

$$H_{\lambda,c} := \big\{x \in \mathbb{R}^d : \langle \lambda, x\rangle > c\big\}.$$

Then we can estimate

$$\begin{aligned}\frac{1}{s_n}\log \mu_n(H_{\lambda,c}) &\leq \frac{1}{s_n}\log \int e^{s_n(\langle \lambda, x\rangle - c)}\mu_n(\mathrm{d}x)\\ &= \frac{1}{s_n}\log \int e^{s_n\langle \lambda, x\rangle}\mu_n(\mathrm{d}x) - c \underset{n\to\infty}{\longrightarrow} \Gamma(\lambda) - c.\end{aligned}$$

Since $0 \in \mathcal{U}_\Gamma$, we can choose vectors $\lambda_1, \ldots, \lambda_n \in \mathcal{U}_\Gamma$ such that

$$K_{\lambda_1,\ldots,\lambda_n,c} := \mathbb{R}^d \setminus \bigcup_{k=1}^n H_{\lambda_k,c}$$

is compact for each $c > 0$. The minimum number of vectors we need is $n = d + 1$ but it is simpler to choose two vectors, one positive and one negative, in each basis direction, so that $K_{\lambda_1,\ldots,\lambda_n,c}$ has the shape of a hyperrectangle and $n = 2d$. Applying our previous estimate with c large enough, using also Lemma 1.10, then yields exponential tightness.

We claim that in fact, for each $\lambda \in \mathcal{U}_\Gamma$, the measures

$$\mu_n^\lambda(\mathrm{d}x) := e^{s_n\langle \lambda, x\rangle}\mu_n(\mathrm{d}x)$$

are exponentially tight. Indeed, for these measures, the limit

$$\Gamma_\lambda(\lambda') := \lim_{n\to\infty}\frac{1}{s_n}\log \int e^{s_n\langle \lambda', x\rangle}\mu_n^\lambda(\mathrm{d}x) = \Gamma(\lambda + \lambda')$$

exists in $[0, \infty]$ for all $\lambda' \in \mathbb{R}^d$ and Γ_λ satisfies the same properties as Γ, so the claim follows from our previous argument.

We now prove the large deviation principle. We aim to apply Lemma 1.30. By Theorem 1.25, exponential tightness implies that each subsequence (μ_n', s_n') of (μ_n, s_n) contains a further subsequence (μ_n'', s_n'') of such that the μ_n'' satisfy a large deviations principle with speed s_n'' and some good rate function J. By Lemma 1.30, to complete the proof, it suffices to prove that $J = I$.

Using the exponential tightness of the μ_n^λ and Varadhan's lemma for unbounded functions (Lemma 1.31), we see that for each $\lambda \in \mathcal{U}_\Gamma$

$$\Gamma(\lambda) = \lim_{n\to\infty}\frac{1}{s_n''}\log \int e^{s_n''\langle \lambda, x\rangle}\mu_n''(\mathrm{d}x) = \sup_{x\in\mathbb{R}^d}\big[\langle \lambda, x\rangle - J(x)\big]. \tag{3.6}$$

Let $g(\lambda)$ be defined by the right-hand side of (3.6). Then Lemma 2.15 tells us that Γ is a restriction of γ. Since by assumption $\Gamma \in \mathrm{Conv}_1(\mathbb{R}^d)$, the function Γ is on maximal domain and hence by Lemma 2.16 we conclude that (3.6) holds for all $\lambda \in \mathbb{R}^d$.

Taking the Legendre transform on both sides of (3.6), applying Lemma 2.13, we see that $I = \overline{J}$, where $\overline{J}$ denotes the convex hull of J. Since $\Gamma \in \mathrm{Conv}_1(\mathbb{R}^d)$, Proposition 2.17 (a) tells us that $I \in \mathrm{Conv}_1(\mathbb{R}^d)$. In particular, I is strictly convex on $\mathcal{U}_I$ and $\mathcal{U}_I \neq \emptyset$, so we can apply Lemma 2.10 to conclude that $I = J$. ∎

As an application of the Gärtner-Ellis theorem, we can give a quick proof of a multi-dimensional version of Cramér's theorem.

Theorem 3.5 (Multi-dimensional Cramér's theorem) *Let $(X_k)_{k\geq 1}$ be i.i.d. $\mathbb{R}^d$-valued random variables with common law μ. Assume that the moment generating function $Z(\lambda)$ defined in (3.1) is finite for all $\lambda \in \mathbb{R}^d$. Then the probability measures*

$$\mu_n := \mathbb{P}\big[\frac{1}{n}\sum_{k=1}^{n} X_k \in \cdot\,\big] \qquad (n \geq 1)$$

satisfy the large deviation principle with speed n and rate function I given by

$$I(y) := \sup_{\lambda\in\mathbb{R}^d} \big[\langle \lambda, y\rangle - \log Z(\lambda)\big].$$

Proof We apply the Gärtner-Ellis theorem with $s_n = n$ and $\Gamma(\lambda) = Z(\lambda)$. Indeed, by the independence of the random variables $(X_k)_{k\geq 1}$, we have

$$\frac{1}{n}\log \mathbb{E}\big[e^{\sum_{k=1}^{n}\langle \lambda, X_k\rangle}\big] = Z(\lambda)$$

for each n, so the right-hand side of (3.5) is constant as a function of n.

Let μ denote the law of X_1. Then the covariance matrix $\mathrm{Cov}(\mu)$ is a symmetric, nonnegative definite matrix, which can be diagonalized with respect to a suitable orthonormal basis. Therefore, we can without loss of generality assume that $\mathrm{Cov}(\mu)$ is a diagonal matrix and that there exists a $0 \leq d' \leq d$ such that $\mathrm{Cov}_{ii}(\mu) > 0$ for $0 < i \leq d'$ and $\mathrm{Cov}_{ii}(\mu) = 0$ for $d' < i \leq d$. Reducing the dimension if necessary, we can assume without loss of generality that $d' = d$. Then Lemma 3.1 tells us that $Z \in \mathrm{Conv}_\infty(\mathbb{R}^d)$ and hence we can apply the Gärtner-Ellis theorem to conclude that the measures μ_n satisfy the large deviation principle with speed n and rate function I. ∎

Remark We recall that elementary properties of the rate function I are listed in Lemma 3.3.

Remark Our proof of Theorem 3.5 shows that the condition that $Z(\lambda)$ is finite for all $\lambda \in \mathbb{R}^d$ can be replaced by the weaker assumption that $\log Z \in \mathrm{Conv}_1(\mathbb{R}^d)$. In fact, it suffices if $Z(\lambda) < \infty$ for λ in some open environment of the origin, see [DZ98, Section 2.2.1], but this strongest version of Cramér's theorem cannot be derived from the Gärtner-Ellis theorem.

We next turn our attention to moderate deviations. The following theorem implies Theorem 0.4. For notational simplicity, we only state and prove the one-dimensional case. The multi-dimensional case is similar, with $I(y) = \frac{1}{2}\langle y, \mathrm{Cov}^{-1} y\rangle$, where Cov is the covariance matrix of X_1 and Cov^{-1} is its inverse.

Theorem 3.6 (Moderate deviations) *Let $(X_k)_{k\geq 1}$ be a sequence of i.i.d. absolutely integrable real random variables with mean $\mathbb{E}[|X_1|] = 0$ and variance $\sigma^2 = \mathrm{Var}(X_1) > 0$. Assume that there exists an $\varepsilon > 0$ such that $\mathbb{E}[e^{\lambda X_1}] < \infty$ for all $|\lambda| \leq \varepsilon$. Then, for each $\frac{1}{2} < \alpha < 1$, the probability measures*

$$\mu_n := \mathbb{P}\big[\frac{1}{n^\alpha}\sum_{k=1}^n X_k \in \cdot\,\big] \qquad (n \geq 1)$$

satisfy the large deviation principle with speed $n^{2\alpha-1}$ and rate function I given by

$$I(y) := \frac{1}{2\sigma^2}y^2 \qquad (y \in \mathbb{R}).$$

Proof We apply the Gärtner-Ellis theorem with $s_n = n^{2\alpha-1}$. Let $Z(\lambda) := \mathbb{E}[e^{\lambda X_1}]$ $(\lambda \in \mathbb{R})$. Then, as in the proof of Theorem 3.5, $\log \int e^{\langle \lambda, x\rangle}\mu_n(\mathrm{d}x) = n \log Z(\lambda)$. It follows that

$$\begin{aligned} n^{1-2\alpha}\log\int e^{n^{2\alpha-1}\lambda x}\mu_n(\mathrm{d}x) &= n^{1-2\alpha}\log\mathbb{E}\big[e^{n^{\alpha-1}\lambda\sum_{k=1}^n X_k}\big] \\ &= n^{2-2\alpha}\log\mathbb{E}\big[e^{n^{\alpha-1}\lambda X_1}\big] = n^{2-2\alpha}\log Z(n^{\alpha-1}\lambda). \end{aligned}$$

It follows from Lemma 3.1 that $\log Z$ is infinitely differentiable with $\log Z(0) = 0$, $(\log Z)'(0) = 0$, and $(\log Z)''(0) = \frac{1}{2}\sigma^2$, so approximately $\log Z(n^{\alpha-1}\lambda) \approx \frac{1}{2}\sigma^2 n^{2\alpha-2}\lambda^2$ when n is large and in this way we see that (3.5) is satisfied with

$$\Gamma(\lambda) = \tfrac{1}{2}\sigma^2\lambda^2 \qquad (\lambda \in \mathbb{R}).$$

Then clearly $\Gamma \in \mathrm{Conv}_\infty(\mathbb{R})$, so to complete the proof, it suffices to notice that by Exercise 2.12, the Legendre transform of Γ is the function I defined above. ∎

3.4 Relative entropy

We now start preparing for the proof of Sanov's theorem, which generalizes Theorem 0.7 from the introduction to general Polish spaces. Let E be a Polish space and let $\mathcal{M}_1(E)$ be the space of probability measures on E, equipped with the topology of weak convergence, under which $\mathcal{M}_1(E)$ is Polish. Recall that by the Radon-Nikodym theorem, if $\nu, \mu \in \mathcal{M}_1(E)$, then ν has a density w.r.t. μ if and only if ν is *absolutely continuous* w.r.t. μ, i.e., $\nu(A) = 0$ for all $A \in \mathcal{B}(E)$ such that $\mu(A) = 0$. We denote this as $\nu \ll \mu$ and let $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ denote the density of ν w.r.t. μ, which is uniquely defined up to a.s. equality w.r.t. μ. For any $\nu, \mu \in \mathcal{M}_1(E)$, we define the *relative entropy* $H(\nu|\mu)$ of ν w.r.t. μ as

$$H(\nu|\mu) := \begin{cases} \displaystyle\int \log\big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\big)\mathrm{d}\nu = \int \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \log\big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\big)\mathrm{d}\mu & \text{if } \nu \ll \mu, \\ \infty & \text{otherwise.} \end{cases}$$

Note that if $\nu \ll \mu$, then a.s. equality w.r.t. μ implies a.s. equality w.r.t. ν, which shows that the first formula for $H(\nu|\mu)$ is unambiguous.

The following lemma gives some more properties of the relative entropy.

Lemma 3.7 (Properties of the relative entropy) *For each $\mu \in \mathcal{M}_1(E)$, the function $H(\,\cdot\,|\mu)$ has the following properties.*

(i) *$H(\mu|\mu) = 0$ and $H(\nu|\mu) > 0$ for all $\nu \neq \mu$.*

(ii) *The map $\mathcal{M}_1(E) \ni \nu \mapsto H(\nu|\mu)$ is convex.*

(iii) *$H(\,\cdot\,|\mu)$ is a good rate function.*

Proof Define $\phi : [0, \infty) \to \mathbb{R}$ by

$$\phi(r) := \begin{cases} r \log r & (r > 0), \\ 0 & (r = 0). \end{cases}$$

Then ϕ is continuous at 0 and

$$\phi'(r) = \log r + 1 \quad \text{and} \quad \phi''(r) = r^{-1} \qquad (r > 0).$$

In particular, ϕ is strictly convex, so by Jensen's inequality

$$H(\nu|\mu) = \int \phi\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\mathrm{d}\mu \geq \phi\big(\int \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\mathrm{d}\mu\big) = 1 \log 1 = 0,$$

with equality if and only if $\mathrm{d}\nu/\mathrm{d}\mu$ is equal to a constant a.s. w.r.t. μ. This proves part (i).

To prove part (ii), fix $\nu_1, \nu_2 \in \mathcal{M}_1(E)$ and $0 \leq p \leq 1$. We wish to show that

$$H(p\nu_1 + (1-p)\nu_2|\mu) \geq pH(\nu_1|\mu) + (1-p)H(\nu_2|\mu).$$

If either $\nu_1 \not\ll \mu$ or $\nu_2 \not\ll \mu$ (or both), then the statement is obvious. Otherwise, setting $f_i = \mathrm{d}\nu_i/\mathrm{d}\mu$, we have

$$\begin{aligned} H(p\nu_1 + (1-p)\nu_2|\mu) &= \int \phi(pf_1 + (1-p)f_2)\mathrm{d}\mu \\ &\geq \int \big(p\phi(f_1) + (1-p)\phi(f_2)\big)\mathrm{d}\mu = pH(\nu_1|\mu) + (1-p)H(\nu_2|\mu) \end{aligned}$$

by the convexity of $\phi(r) = r\log r$.

To prove part (iii), finally, we must show that for each $r < \infty$, the level set

$$L_r := \big\{\nu \in \mathcal{M}_1(E) : H(\nu|\mu) \leq r\big\}$$

is a compact subset of $\mathcal{M}_1(E)$. Let $L^1(\mu)$ be the Banach space consisting of all equivalence classes of w.r.t. μ a.e. equal, absolutely integrable functions, equipped with the norm $\|f\|_1 := \int |f|\mathrm{d}\mu$. Then, identifying a measure with its density, we have

$$\{\nu \in \mathcal{M}(E) : \nu \ll \mu\} \cong \{f \in L^1(\mu) : f \geq 0\},$$

and we we may identify L_r with the set

$$L'_r := \big\{f \in L^1(\mu) : f \geq 0,\ \int f\mathrm{d}\mu = 1,\ \int f\log f\mathrm{d}\mu \leq r\big\}.$$

Recall that a set $C \subset L^1(\mu)$ is *uniformly integrable* if for each $\varepsilon > 0$ there exists a $K < \infty$ such that

$$\sup_{f\in C} \int 1_{\{|f|\geq K\}}|f|\mathrm{d}\mu \leq \varepsilon.$$

A sufficient condition for uniform integrability is the existence of a nonnegative, increasing, convex function $\psi : [0,\infty) \to [0,\infty)$ such that $\lim_{r\to\infty} \psi(r)/r = \infty$ and

$$\sup_{f\in C} \int \psi(|f|)\mathrm{d}\mu < \infty.$$

(In fact, by the De la Vallée-Poussin theorem, this condition is also necessary, but we will not need this deeper converse.) Applying this to $\psi = \phi + 1$, we see that the set L'_r is uniformly integrable.

By Prohorov's theorem (Proposition 1.19), to show that L_r is relatively compact, it suffices to show that for each $\varepsilon > 0$ there exists a compact set $D \subset E$ such that $\sup_{\nu \in L_r} \nu(E \backslash D) \leq \varepsilon$. Since L'_r is uniformly integrable, we can find a $K < \infty$ such that $\sup_{f \in L'_r} \int 1_{\{f \geq K\}} f \mathrm{d}\mu \leq \frac{1}{2}\varepsilon$. Moreover, since E is Polish, for each $\delta > 0$, we can find a compact set $D \subset E$ such that $\mu(E \backslash D) \leq \delta$. Applying this with $\delta = \varepsilon/(2K)$, we see that

$$\sup_{\nu \in L_r} \nu(E \backslash D) = \sup_{f \in L'_r} \Big\{ \int 1_{\{f<K\} \backslash D} f \mathrm{d}\mu + 1_{\{f \geq K\} \backslash D} f \mathrm{d}\mu \Big\} \leq K\mu(E \backslash D) + \tfrac{1}{2}\varepsilon \leq \varepsilon,$$

proving the relative compactness of L_r.

To complete the proof, we need to show that L_r is closed with respect to the topology on $\mathcal{M}_1(E)$. We start by showing that L'_r is closed with respect to the norm $\| \cdot \|_1$. Let $f_n \in L'_r$, $f \in L^1(\mu)$ be such that $\|f_n - f\|_1 \to 0$. By going to a subsequence if necessary, we may assume that $f_n \to f$ a.s. Since the function ϕ is bounded from below, it follows from Fatou's lemma that

$$\int \phi(f) \mathrm{d}\mu \leq \liminf_{n \to \infty} \int \phi(f_n) \mathrm{d}\mu \leq r,$$

which shows that $f \in L'_r$.

It seems that we are almost done, but there still remains a tough technical hurdle to pass. The assumption that $\nu_n \Rightarrow \nu$ with respect to the topology on $\mathcal{M}_1(E)$ is much weaker than the assumption that the densities f_n of ν_n converge to the density f of ν with respect to the norm $\| \cdot \|_1$, and hence the fact that L'_r is norm-closed is much weaker than what we really need, namely, that L_r is closed with respect to weak convergence of probability measures.

We recall that for any real Banach space $(V, \| \cdot \|)$, the *dual* V^* is the space of all continuous linear forms on V, i.e., the space of all continuous linear maps $l : V \to \mathbb{R}$. The *weak topology* on V is the weakest topology on V that makes all the maps $\{l : l \in V^*\}$ continuous, i.e., it is the topology on V generated by the open sets $\{l^{-1}(O) : O \subset \mathbb{R} \text{ open}, l \in V^*\}$. The weak topology is usually weaker than the norm topology on V. Some care is needed when dealing with weak topologies since they are often not metrizable.

It is known that the dual of $L^1(\mu)$ is isomorphic to the space $L^\infty(\mu)$ of equivalence classes of w.r.t. μ a.e. equal, bounded measurable functions, equipped with the essential supremumnorm $\|f\|_\infty := \inf\{R < \infty : |f| \leq R \text{ a.s.}\}$. In particular, this means that the weak topology on $L^1(\mu)$ is the weakest topology that makes the

linear forms

$$f \mapsto l_g(f) := \int fg \, \mathrm{d}\mu \quad \big(g \in B_b(E)\big)$$

continuous. We now need two facts from functional analysis.

1. Let V be a Banach space and let $C \subset V$ be convex and norm-closed. Then C is also closed with respect to the weak topology on V.
2. (Dunford-Pettis) A subset $C \subset L^1(\mu)$ is relatively compact in the weak topology if and only if C is uniformly integrable.

It follows from the convexity of the map $\nu \mapsto H(\nu \,|\, \mu)$ that L_r is convex, so we can use 1. and the fact that L'_r is norm closed to conclude that L'_r is also closed with respect to the weak topology on $L^1(\mu)$. Now assume that $\nu_n \in L_r$ converge in the topology[1] on $\mathcal{M}_1(E)$ to a limit ν, and let $f_n := \mathrm{d}\nu_n/\mathrm{d}\mu$ and $f := \mathrm{d}\nu/\mathrm{d}\mu$. Since L'_r is uniformly integrable, by the Dunford-Pettis theorem, we can select a subsequence so that $f_n \to f$ in the weak topology on $L^1(\mu)$. Using the fact that L'_r closed with respect to the weak topology on $L^1(\mu)$, we conclude that $\nu \in L_r$, completing our proof. ∎

Remark The most difficult part of the proof of Lemma 3.7 is showing that $\nu \mapsto H(\nu \,|\, \mu)$ is lower semi-continuous with respect to the topology on $\mathcal{M}_1(E)$. We have based our proof on [DZ93, Lemma 6.2.16], using also some minor ideas from [RS15]. Alternatively, one can first prove a variational formula for $H(\nu \,|\, \mu)$ from which lower semi-continuity then follows, which is the approach taken in [DS89, Lemma 3.2.13], [DE97, Lemma 1.4.3 (b) and (c)], and [RS15, Theorem 5.6 and Proposition 5.7], but this is also quite technical.

We will prove Sanov's theorem by deriving it from Cramér's theorem using a projective limit (Theorem 1.37). In order to do this, we need to show that Cramér's rate function I is the contraction of Sanov's rate function $H(\,\cdot\,|\mu)$ (see Figure 3.2). Note that if $E = \mathbb{R}^d$ and f is the identity function, then the measures μ_λ defined in (3.7) are the same as those defined in (3.2). The next lemma says that among all measures ν such that $\int f \, \mathrm{d}\nu = \int f \, \mathrm{d}\mu_\lambda$, the measure μ_λ stands out since it has the lowest relative entropy with respect to μ.

Lemma 3.8 (Minimizers of the entropy) *Let E be a Polish space, let μ be a probability measure on E, and let $f : E \to \mathbb{R}^d$ be a measurable function. Assume*

[1] In spite of the confusing terminology, this is not equivalent to saying that the densities $\mathrm{d}\nu_n/\mathrm{d}\mu$ converge to $\mathrm{d}\nu/\mathrm{d}\mu$ with respect to the weak topology on $L^1(\mu)$.

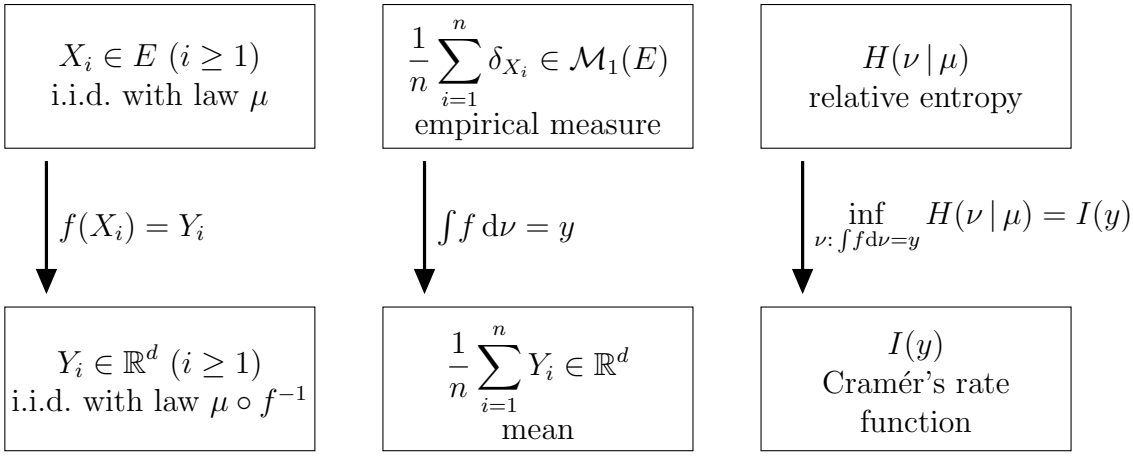

Figure 3.2: Two levels of large deviation principles: relation between Sanov's and Cramér's theorem.

that

$$Z(\lambda) := \int e^{\langle \lambda,\, f(x) \rangle} \mu(\mathrm{d}x) < \infty \qquad (\lambda \in \mathbb{R}^d),$$

and that the covariance matrix of $\mu \circ f^{-1}$ *is strictly positive. Let*

$$I(y) := \sup_{\lambda \in \mathbb{R}^d} \big[\langle y, \lambda \rangle - \log Z(\lambda) \big] \qquad (y \in \mathbb{R}^d)$$

be the Legendre transform of $\log Z$*. For each* $\lambda \in \mathbb{R}^d$*, let* μ_λ *be the probability measure on* E *defined by*

$$\mu_\lambda(\mathrm{d}x) = \frac{1}{Z(\lambda)} e^{\langle \lambda,\, f(x) \rangle} \mu(\mathrm{d}x). \tag{3.7}$$

Let

$$\mathcal{M}_1^f(E) := \{\nu \in \mathcal{M}_1(E) : \int |f| \,\mathrm{d}\nu < \infty\}.$$

Then, for all $y_\circ \in \mathcal{U}_I$*, the function* $H(\,\cdot\,|\mu)$ *assumes its minimum over the set* $\{\nu \in \mathcal{M}_1^f(E) : \int \nu(\mathrm{d}x) f(x) = y_\circ\}$ *in the unique point* $\mu_{\lambda_\circ}$ *given by the requirement that*

$$\int \mu_{\lambda_\circ}(\mathrm{d}x) f(x) = y_\circ.$$

Moreover, one has $H(\mu_{\lambda_\circ}|\mu) = I(y_\circ)$*.*

Proof We wish to find the minimum of $H(\,\cdot\,|\mu)$ over the set of all $\nu \in \mathcal{M}_1^f(E)$ subject to the constraint $\int f\,\mathrm{d}\nu = y_\circ$, which are really d constraints since $y_\circ \in \mathbb{R}^d$. We use the method of Lagrange multipliers: we first try to find the minimum of the function $H(\nu\,|\,\mu) - \langle \lambda, \int f\,\mathrm{d}\nu\rangle$ for general $\lambda \in \mathbb{R}^d$, and then try to choose λ in such a way that the minimizer satisfies the constraints.

We start by proving that for any $\lambda \in \mathbb{R}^d$ and $\nu \in \mathcal{M}_1^f(E)$,

$$H(\nu|\mu) \geq \int \nu(\mathrm{d}x)\langle \lambda, f(x)\rangle - \log Z(\lambda) \qquad \big(\nu \in \mathcal{M}_1^f(E)\big), \tag{3.8}$$

where equality holds for a given value of λ if and only if $\nu = \mu_\lambda$. The inequality is trivial if $H(\nu|\mu) = \infty$ so we may assume that $\nu \ll \mu$ and $H(\nu|\mu) = \int \log(\mathrm{d}\nu/\mathrm{d}\mu)\mathrm{d}\nu < \infty$. We can split the measure μ in an absolutely continuous and singular part w.r.t. ν, i.e., we can find a measurable set A and nonnegative measurable function h such that $\nu(A) = 0$ and

$$\mu(\mathrm{d}x) = 1_A(x)\mu(\mathrm{d}x) + h(x)\nu(\mathrm{d}x).$$

Weighting the measures on both sides of this equation with the density $\mathrm{d}\nu/\mathrm{d}\mu$, which is zero on A a.s. w.r.t. μ, we see that

$$\nu(\mathrm{d}x) = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)h(x)\nu(\mathrm{d}x),$$

which shows that $h(x) = (\mathrm{d}\nu/\mathrm{d}\mu)^{-1}$ a.s. with respect to ν. Since $r \mapsto \log(r)$ is a strictly concave function, Jensen's inequality gives

$$\begin{aligned}
\int \nu(\mathrm{d}x)\langle \lambda, f(x)\rangle - H(\nu|\mu) &= \int \nu(\mathrm{d}x)\Big(\log\big(e^{\langle \lambda,\, f(x)\rangle}\big) - \log\big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\big)\Big) \\
&= \int \nu(\mathrm{d}x)\log\Big(e^{\langle \lambda,\, f(x)\rangle}(\frac{\mathrm{d}\nu}{\mathrm{d}\mu})^{-1}(x)\Big) \leq \log\Big(\int \nu(\mathrm{d}x)\,e^{\langle \lambda,\, f(x)\rangle}h(x)\Big) \\
&\leq \log\Big(\int \mu(\mathrm{d}x)\,e^{\langle \lambda,\, f(x)\rangle}\Big) = \log Z(\lambda).
\end{aligned}$$

This proves (3.8). Since the logarithm is a strictly concave function, the first inequality here (which is an application of Jensen's inequality) is an equality if and only if the function $e^{\langle \lambda,\, f\rangle}(\frac{\mathrm{d}\nu}{\mathrm{d}\mu})^{-1}$ is a.s. constant w.r.t. ν. Since the logarithm is a strictly increasing function and $e^{\langle \lambda, f\rangle}$ is strictly positive, the second inequality is an equality if and only if $\mu = h\nu$, i.e., if $\mu \ll \nu$. Thus, we have equality in (3.8) if and only if $\mu \ll \nu$ and

$$\nu(\mathrm{d}x) = \frac{1}{Z}e^{\langle \lambda,\, f(x)\rangle}\mu(\mathrm{d}x),$$

where Z is some constant. Since ν is a probability measure, we must have $Z = Z(\lambda)$.

Our arguments so far imply that for each $y_\circ \in \mathbb{R}^d$, one has

$$H(\nu|\mu) \geq \langle \lambda, y_\circ \rangle - \log Z(\lambda) \qquad \forall \lambda \in \mathbb{R}^d,\ \nu \in \mathcal{M}_1^f(E) \text{ s.t. } \textstyle\int f \,\mathrm{d}\nu = y_\circ, \tag{3.9}$$

with equality if and only if λ has the property that $\int f \,\mathrm{d}\mu_\lambda = y_\circ$ and $\nu = \mu_\lambda$. To complete the proof, we must show that if $y_\circ \in \mathcal{U}_I$, then there exists a unique $\lambda_\circ$ such that $\int f \,\mathrm{d}\mu_{\lambda_\circ} = y_\circ$, and $H(\mu_{\lambda_\circ}|\mu) = I(y_\circ)$.

Note that $Z(\lambda)$ is the moment generating function of $\mu \circ f^{-1}$, i.e.,

$$Z(\lambda) = \int (\mu \circ f^{-1})(\mathrm{d}x) e^{\langle \lambda, x \rangle}.$$

Moreover, the image under f of the measure μ_λ defined in (3.7) is the measure

$$\mu_\lambda \circ f^{-1}(\mathrm{d}y) = \frac{1}{Z(\lambda)} e^{\langle \lambda,\, y \rangle} (\mu \circ f^{-1})(\mathrm{d}y),$$

i.e., this is $(\mu \circ f^{-1})_\lambda$ in the notation of formula (3.2). Note that we are assuming that the the covariance matrix of $\mu \circ f^{-1}$ is strictly positive, so Lemma 3.3 is applicable. Now, if $y_\circ \in \mathcal{U}_I$, then by Lemma 3.3 (vi), the supremum

$$I(y_\circ) = \sup_{\lambda \in \mathbb{R}^d} \big[\langle y_\circ, \lambda \rangle - \log Z(\lambda) \big]$$

is attained in a unique point $\lambda_\circ \in \mathbb{R}^d$ which is uniquely characterized by the requirement that $\int f \,\mathrm{d}\mu_{\lambda_\circ} = \langle (\mu \circ f^{-1})_{\lambda_\circ} \rangle = y_\circ$. Comparing with (3.9), we see that $I(y_\circ) = H(\mu_{\lambda_\circ} \,|\, \mu)$. ∎

Exercise 3.9 (Joint continuity of relative entropy) Let S be a finite set and let $\mathring{\mathcal{M}}_1(S) := \{\mu \in \mathcal{M}_1(S) : \mu(x) > 0\ \forall x \in S\}$. Prove the continuity of the map

$$\mathcal{M}_1(S) \times \mathring{\mathcal{M}}_1(S) \ni (\nu, \mu) \mapsto H(\nu|\mu).$$

Exercise 3.10 (Convexity of relative entropy) Let S be a finite set and let $\mu \in \mathcal{M}_1(S)$. Give a direct proof of the fact that

$$\mathcal{M}_1(S) \ni \nu \mapsto H(\nu|\mu)$$

is a lower semi-continuous, convex function.

3.5 Sanov's theorem

The aim of this section is to prove the following result, which (at least in the case $E = \mathbb{R}$) goes back to Sanov [San61]. As a simple application, we will also prove Theorem 0.7.

Theorem 3.11 (Sanov's theorem) *Let $(X_k)_{k\geq 0}$ be i.i.d. random variables taking values in a Polish space E, with common law μ, and let*

$$M_n := \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k} \qquad (n \geq 1)$$

be the empirical laws of the $(X_k)_{k\geq 0}$. Then the laws $\mu_n := \mathbb{P}[M_n \in \cdot\,]$, viewed as probability laws on the Polish space $\mathcal{M}_1(E)$ of probability measures on E, equipped with the topology of weak convergence, satisfy the large deviation principle with speed n and rate function $H(\,\cdot\,|\mu)$.

The proof of Theorem 3.11 depends on the following proposition, the proof of which we postpone till the end of this section.

Proposition 3.12 (Contracted rate function) *Let E be a Polish space, let μ be a probability measure on E, and let $f : E \to \mathbb{R}^d$ be a measurable function. Assume that*

$$Z(\lambda) := \int e^{\langle \lambda,\, f(x)\rangle}\mu(\mathrm{d}x) < \infty \qquad (\lambda \in \mathbb{R}^d),$$

and let

$$I(y) := \sup_{\lambda\in\mathbb{R}^d}\big[\langle y, \lambda\rangle - \log Z(\lambda)\big] \qquad (y \in \mathbb{R}^d)$$

be the Legendre transform of $\log Z$. Then

$$I(y) = \inf_{\substack{\nu\in\mathcal{M}_1^f(E)\\ \int f\,\mathrm{d}\nu = y}} H(\nu\,|\,\mu). \tag{3.10}$$

Proof of Theorem 3.11 We apply Theorem 1.37 about projective limits. We first consider the case that E is compact. In this case, $\mathcal{M}_1(E)$ is also compact so exponential tightness comes for free.

Since $\mathcal{C}(E)$ is separable, we may choose a countable dense set $\{f_i : i \in \mathbb{N}_+\} \subset \mathcal{C}(E)$. For each $i \in \mathbb{N}_+$, we define $\psi_i : \mathcal{M}_1(E) \to \mathbb{R}$ by $\psi_i(\nu) := \int f_i \mathrm{d}\nu$. The $(\psi_i)_{i\in\mathbb{N}_+}$ are

continuous by the definition of weak convergence of measures. We claim that they also separate points. To see this, imagine that $\nu, \nu' \in \mathcal{M}_1(E)$ and $\psi_i(\nu) = \psi_i(\nu')$ for all $i \geq 1$. Then $\int f \mathrm{d}\nu = \int f \mathrm{d}\nu'$ for all $f \in \mathcal{C}(E)$ by the fact that $\{f_i : i \in \mathbb{N}_+\}$ is dense, and therefore $\nu = \nu'$.

Let

$$\vec{f}_d(x) := \big(f_1(x), \ldots, f_d(x)\big) \qquad (x \in E,\ d \geq 1),$$

and

$$\vec{\psi}_d(\nu) := \big(\psi_1(\nu), \ldots, \psi_d(\nu)\big) = \int \vec{f}_d \,\mathrm{d}\nu \qquad \big(\nu \in \mathcal{M}_1(E)\big).$$

By Theorem 3.5, for each $d \geq 1$, the laws $\mu_n \circ \vec{\psi}_d^{-1}$ satisfy the large deviation principle with a good rate function I_d. By Proposition 3.12, this rate function is given by

$$I_d(y) = \inf_{\substack{\nu \in \mathcal{M}_1(E) \\ \int \vec{f}_d \,\mathrm{d}\nu = y}} H(\nu \,|\, \mu) \qquad (y \in \mathbb{R}^d).$$

Theorem 1.37 now implies that the measures μ_n satisfy the large deviation principle with rate function $H(\,\cdot\,|\mu)$. This completes the proof for compact E.

To prove the general statement, let $\overline{E}$ be a metrizable compactification of E. By Proposition 1.27, such a compactification exists and E is a G_δ-subset of E. By what we have already proved, the laws μ_n, viewed as probability laws on the Polish space $\mathcal{M}_1(\overline{E})$ of probability measures on $\overline{E}$, equipped with the topology of weak convergence, satisfy the large deviation principle with speed n and rate function $H(\,\cdot\,|\mu)$.

We view $\mathcal{M}_1(E)$ as a subset of $\mathcal{M}_1(\overline{E})$. By Exercise 1.29, the topology on $\mathcal{M}_1(E)$ is the induced topology from $\mathcal{M}_1(\overline{E})$. Since $\mathcal{M}_1(E)$ is Polish in this topology, it must be a G_δ-subset of $\mathcal{M}_1(\overline{E})$. By the restriction principle (Lemma 1.28), using the fact that $H(\,\cdot\,|\mu)$ is a good rate function (which has been proved in Lemma 3.7) and the fact that $H(\,\cdot\,|\,\mu) = \infty$ on $\mathcal{M}_1(\overline{E})\backslash\mathcal{M}_1(E)$, we conclude that the laws μ_n, viewed as probability laws on $\mathcal{M}_1(E)$, satisfy the large deviation principle with speed n and rate function $H(\,\cdot\,|\mu)$. ∎

Remark For some purposes, the topology of weak convergence on $\mathcal{M}_1(E)$ is too weak. With some extra work, it is possible to improve Theorem 3.11 by showing that the emperical measures satisfy the large deviation principle with respect to the (much stronger) topology of strong convergence of measures; see [DS89, Section 3.2].

Proof of Lemma 0.6 and Theorem 0.7 If in Theorem 3.11, $E = S$ is a finite set and $\mu(\{x\}) > 0$ for all $x \in S$, then the theorem and its proof simplify considerably. In this case, without loss of generality, we may assume that $S = \{0, \ldots, d\}$ for some $d \geq 1$. We may identify $\mathcal{M}_1(S)$ with the convex subset of $\mathbb{R}^d$ given by

$$\mathcal{M}_1(S) = \big\{x \in \mathbb{R}^d : x(i) \geq 0 \ \forall i = 1, \ldots, d, \ \sum_{i=1}^{d} x(i) \leq 1\big\},$$

where $x(0)$ is determined by the condition $\sum_{i=0}^{d} x(i) = 1$. Thus, we may apply Cramér's theorem (Theorem 3.5) to the $\mathbb{R}^d$-valued random variables M_n. The fact that the rate function from Cramér's theorem is in fact $H(\nu|\mu)$ follows from Lemma 3.8. Since $\mu(\{x\}) > 0$ for all $x \in S$, it is easy to see that the covariance condition of Lemma 3.3 is fulfilled, so Lemma 0.6 follows from Lemma 3.3 and the observation that $H(\nu|\mu) < \infty$ for all $\nu \in \mathcal{M}_1(S)$. ∎

Remark There exists a nice combinatorical proof of Sanov's theorem for finite spaces (Theorem 0.7), in the spirit of our Section 4.2 below. See [Hol00, Section II.1].

We still need to provide the proof of Proposition 3.12 Using Lemma 3.8, it is easy to prove (3.10) when the covariance matrix of $\mu \circ f^{-1}$ is positive and $y \in \mathcal{U}_I$ or $y \notin \overline{\mathcal{U}}_I$. By going to a suitable subspace, it is easy to get rid of the condition on the covariance matrix. Thus, it only remains to prove (3.10) when y lies on the boundary of $\mathcal{U}_I$. This seems to be surprisingly hard. One can try to use a continuity argument,[2] using that both sides of (3.10) are convex and lower-semicontinuous in y. Convexity is easy, but proving lower-semicontinuity for the right-hand side seems to be hard. If f is bounded, then this follows from the (nontrivial) fact that the level sets of $H(\,\cdot\,|\mu)$ are compact in the (non-Polish) topology of strong convergence of measures, but the general case seems hard. A different approach is to approximate μ with other, nicer measures, for which $\mathcal{U}_I = \mathbb{R}^d$. Again, one runs into the problem that convergence of the right-hand side of (3.10) seems to be difficult to prove. The proof below is by brute force, explicitly identifying the unique minimizer of the right-hand side of (3.10) for each value of y where the infimum is not ∞. This proof is probably best skipped at a first reading.

Proof of Proposition 3.12 Let us write $Z_\mu(\lambda)$ and $I_\mu(y)$ to make the dependence of these quantities on μ explicit. We observe that the formulas for $Z_\mu(\lambda)$, $I_\mu(y)$, and $H(\nu\,|\,\mu)$ still make sense if μ is a finite measure but not necessarily a probability

[2] I made such a claim in a previous version of the lecture notes, but the argument I used is not correct.

measure. Moreover, for any nonnegative constant r, one has $Z_{r\mu}(\lambda) = rZ_\mu(\lambda)$ and hence

$$I_{r\mu}(y) = I_\mu(y) - \log r \quad \text{and} \quad H(\nu \,|\, r\mu) = H(\nu \,|\, \mu) - \log r. \tag{3.11}$$

In view of this, if (3.10) holds for probability measures μ, then it holds more generally when μ is a finite measure, and vice versa. We will prove the statement immediately for finite measures.

Using the scaling relations (3.11), we see that (3.9) holds more generally if μ is a finite measure. Taking the supremum over $\lambda \in \mathbb{R}^d$, this implies that

$$I(y) \leq \inf_{\substack{\nu \in \mathcal{M}_1^f(E) \\ \int f \,\mathrm{d}\nu = y}} H(\nu \,|\, \mu).$$

To prove the opposite inequality, by the definition of $I(y)$, we must show that there exists $\lambda_n \in \mathbb{R}^d$ and $\nu \in \mathcal{M}_1^f(E)$ with $\int f \,\mathrm{d}\nu = y$ such that

$$\langle y, \lambda_n \rangle - \log Z(\lambda_n) \underset{n\to\infty}{\longrightarrow} H(\nu \,|\, \mu). \tag{3.12}$$

For any finite nonzero measure $\mu \in \mathcal{M}(E)$ and $\lambda \in \mathbb{R}^d$, we define μ_λ by (3.7), which is a probability measure even if μ is not. We have seen in (3.8) that

$$H(\mu_\lambda \,|\, \mu) = \int \mu_\lambda(\mathrm{d}x) \langle \lambda, f(x) \rangle - \log Z(\lambda). \tag{3.13}$$

In the proof of Lemma 3.8, we have seen that if y lies in the interior of the support of $\mu \circ f^{-1}$, then there exists a unique $\lambda_\circ \in \mathbb{R}^d$ such that $\int f \,\mathrm{d}\mu_{\lambda_\circ} = y$. By (3.13), we then see that (3.12) is satisfied for $\lambda_n := \lambda_\circ$ and $\nu := \mu_{\lambda_\circ}$.

For general y, we have to proceed more carefully. Consider the set

$$C := \{ \int f \,\mathrm{d}\nu : \nu \in \mathcal{M}_1^f(E),\ \nu \ll \mu \}.$$

It is not hard to see that C is a convex set. For $y \in C$, let

$$F_y := \big\{ z \in \mathbb{R}^d : \exists \varepsilon > 0 \text{ s.t. } y - \varepsilon z \in C \text{ and } y + \varepsilon z \in C \big\}.$$

It follows from the convexity of C that F_y is a linear subspace of $\mathbb{R}^d$ (possibly of dimension zero). For example, if C is a closed cube, then for a point y that lies in the interior of C, in the interior of a face of C, in the interior of an edge of C, or on a corner of C, the dimension of F_y is 3, 2, 1, or 0, respectively. Since C may in general be neither open nor closed, its structure can be quite complicated. For

example, it is possible that the closure of C is a cube, but for a given face of this cube, only a part of the face lies inside C.

It is clear that the right-hand side of (3.10) is ∞ if $y \notin C$. We will show that also $I(y) = \infty$ for $y \notin C$. On the other hand, we will show that for each $y \in C$, the infimum on the right-hand side of (3.10) is attained in a unique probability measure ν, and we will show that there exists λ_n such that (3.12) holds for this ν.

Let L_y denote the affine space $L_y := \{y + z : z \in F_y\}$, let $E' := f^{-1}(L_y)$ and let μ' denote the restriction of μ to E'. Then $\mu' \circ f^{-1}$ is the restriction of $\mu \circ f^{-1}$ to L_y. If $y \in C$, then $\mu' \circ f^{-1}$ must be nonzero and y lies in the interior of support$(\mu' \circ f^{-1})$, viewed as a subset of L_y. Since $\nu \ll \mu$ and $\int f \,\mathrm{d}\nu = y$ imply that $\nu \ll \mu'$, the right-hand side of (3.10) can be rewritten as

$$\inf_{\substack{\nu \in \mathcal{M}_1^f(E') \\ \int f \,\mathrm{d}\nu = y}} H(\nu \,|\, \mu'). \tag{3.14}$$

Note that μ' may fail to be a probability measure even if μ is one. Defining $f' : E' \to F_y$ by $f'(x) := f(x) - y$, we can rewrite (3.14) as

$$\inf_{\substack{\nu \in \mathcal{M}_1^f(E') \\ \int f' \,\mathrm{d}\nu = 0}} H(\nu \,|\, \mu'). \tag{3.15}$$

For each $\lambda' \in F_y$, we define $Z'(\lambda') := \int e^{\langle \lambda', f'(x) \rangle} \mu'(\mathrm{d}x)$ and we define tilted measures $\mu'_{\lambda'}$ as in (3.7). Since 0 lies in the interior of support$(\mu' \circ f'^{-1})$, viewed as a subset of F_y, the proof of Lemma 3.8 tells us that there exists a unique $\lambda'_\circ \in F_y$ such that $\int f' \,\mathrm{d}\mu'_{\lambda'_\circ} = 0$, and the infimum in (3.15) is attained in the unique point $\nu = \mu'_{\lambda'_\circ}$. By (3.13),

$$H(\mu'_{\lambda'_\circ} \,|\, \mu) = 0 - \log Z'(\lambda'_\circ).$$

We will show that (3.12) is satisfied for $\nu = \mu'_{\lambda'_\circ}$. By a change of basis, we can without loss of generality assume that $F_y = \{\lambda \in \mathbb{R}^d : \lambda(i) = 0 \ \forall i = d'+1, \dots, d\}$ and that

$$C \subset \big\{y + z : z \in \mathbb{R}^d : z(i) \le 0 \ \forall i = d'+1, \dots, d\big\}. \tag{3.16}$$

In (3.12), we choose λ_n in such a way that

$$\lambda_n(i) = \lambda'_\circ(i) \quad (i = 1, \dots, d'), \qquad \lambda_n(i) \to \infty \quad (i = d'+1, \dots, d).$$

Then

$$H(\mu'_{\lambda'_\circ} \,|\, \mu) = -\log Z'(\lambda'_\circ) = -\log \int_{\{x:\, f(x) \in L_y\}} e^{\langle \lambda'_\circ,\, f(x) - y \rangle} \mu(\mathrm{d}x).$$

On the other hand, the left-hand side of (3.12) can be written as

$$-\log \int_E e^{\langle \lambda_n,\, f(x)-y\rangle} \mu(\mathrm{d}x).$$

To prove (3.12), we need to show that

$$\int_{\mathbb{R}^d} e^{\langle \lambda_n,\, z\rangle} \mu \circ f'^{-1}(\mathrm{d}z) \underset{n\to\infty}{\longrightarrow} \int_{F_y} e^{\langle \lambda'_\circ,\, z\rangle} \mu \circ f'^{-1}(\mathrm{d}z).$$

By (3.16), the measure $\mu \circ f'^{-1}$ is concentrated on $\{z \in \mathbb{R}^d : z(i) \leq 0 \ \forall i = d'+1, \ldots, d\}$. Since $e^{\langle \lambda_n,\, z\rangle} \downarrow 0$ if $z(i) > 0$ for some $i \in \{d'+1, \ldots, d\}$, in the limit, only the integral over F_y remains and we see that (3.12) is satisfied.

To complete the proof, we must show that $I(y) = \infty$ for $y \notin C$. In this case, by a change of basis, we can without loss of generality assume that $\mu \circ f^{-1}$ is concentrated on $\{y + z : z(i) < 0 \ \forall i = 1, \ldots, d\}$. Choosing $\lambda_n(i) \to \infty$ for all $i = 1, \ldots, d$, setting $f'(x) := f(x) - y$ as before, one finds that

$$\langle y, \lambda_n\rangle - \log Z(\lambda_n) = -\log \int_{\mathbb{R}^d} e^{\langle \lambda_n,\, z\rangle} \mu \circ f'^{-1}(\mathrm{d}z) \underset{n\to\infty}{\longrightarrow} \infty,$$

proving that $I(y) = \infty$ ∎

Chapter 4

Markov chains

4.1 Basic notions

Let S be a finite set and let P be a *probability kernel* on S, i.e., $P : S \times S \to \mathbb{R}$ is a function such that

$$\begin{array}{lll} \text{(i)} & P(x,y) \geq 0 & (x,y \in S), \\ \text{(ii)} & \sum_{y \in S} P(x,y) = 1 & (x \in S). \end{array}$$

For any function $f : S \to \mathbb{R}$, we put

$$Pf(x) := \sum_{y \in S} P(x,y) f(y),$$

which defines a linear operator $P : \mathbb{R}^S \to \mathbb{R}^S$. For any measure μ on S we write $\mu(x) := \mu(\{x\})$ and for $f : S \to \mathbb{R}$, we let

$$\mu f(y) := \sum_{x \in S} \mu(x) f(x)$$

denote the expectation of f w.r.t. μ. Viewing a measure μ as a linear operator $\mu : \mathbb{R}^S \to \mathbb{R}$, we see that the composition of a probability kernel $P : \mathbb{R}^S \to \mathbb{R}^S$ and a probability measure $\mu : \mathbb{R}^S \to \mathbb{R}$ is an operator $\mu P : \mathbb{R}^S \to \mathbb{R}$ that corresponds to the probability measure $\mu P(y) = \sum_{x \in S} \mu(x) P(x,y)$.

A *Markov chain* with *state space* S, *transition kernel* P and *initial law* μ is a collection of S-valued random variables $(X_k)_{k \geq 0}$ whose finite-dimensional distributions

are characterized by

$$\mathbb{P}[X_0 = x_0, \ldots, X_n = x_n] = \mu(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n)$$

($n \geq 1$, $x_0, \ldots, x_n \in S$). Note that in particular, the law of X_n is given by μP^n, where P^n is the n-th power of the linear operator P. We also introduce the notation

$$\mu \otimes P(x_0, x_1) := \mu(x_0) \otimes P(x_0, x_1)$$

to denote the probability measure on S^2 that is the law of (X_0, X_1).

Write $x \overset{P}{\rightsquigarrow} y$ if there exist $n \geq 0$ such that $P^n(x, y) > 0$ or equivalently, there exist $x = x_0, \ldots, x_n = y$ such that $P(x_{k-1}, x_k) > 0$ for each $k = 1, \ldots, n$. Then P is called *irreducible* if $x \overset{P}{\rightsquigarrow} y$ for all $x, y \in S$. An *invariant law* of P is a probability measure μ on S such that $\mu P = \mu$. Equivalently, μ is invariant if the Markov chain $(X_k)_{k\geq 0}$ with transition kernel P and initial law μ is *stationary*, i.e. $(X_k)_{k\geq 0}$ is equal in law to $(Y_k)_{k\geq 0}$ defined as $Y_k := X_{k+1}$ ($k \geq 0$). The *period* of a state $x \in S$ is the greatest common divisor of the set $\{n \geq 1 : P^n(x, x) > 0\}$. If P is irreducible, then all states have the same period. If all states have period one, then we say that P is *aperiodic*. Basic results of Markov chain theory tell us that an irreducible Markov chain with a finite state space S has a unique invariant law μ, which has the property that $\mu(x) > 0$ for all $x \in S$. If P is moreover aperiodic, then νP^n converges to μ as $n \to \infty$, for each initial law ν.

For any Markov chain $X = (X_k)_{k\geq 0}$, we let

$$M_n^{(2)} := \frac{1}{n} N_n^{(2)}, \quad \text{where} \quad N_n^{(2)}(x) := \sum_{k=1}^{n} 1_{\{(X_{k-1}, X_k) = (x_1, x_2)\}} \tag{4.1}$$

($x \in S^2$, $n \geq 1$) be the *pair empirical distribution* of the first $n + 1$ random variables. The $M_n^{(2)}$ are random variables taking values in the space $\mathcal{M}_1(S^2)$ of probability measures on $S^2 := \{x = (x_1, x_2) : x_i \in S \ \forall i = 1, 2\}$. If X is irreducible, then the $M_n^{(2)}$ satisfy a strong law of large numbers.

Proposition 4.1 (SLLN for Markov chains) *Let $X = (X_k)_{k\geq 0}$ be an irreducible Markov chain with finite state space S, transition kernel P, and arbitrary initial law. Let $(M_n^{(2)})_{n\geq 1}$ be the pair empirical distributions of X and let μ be its invariant law. Then*

$$M_n^{(2)} \underset{n\to\infty}{\longrightarrow} \mu \otimes P \quad \text{a.s.} \tag{4.2}$$

Proof (sketch) It suffices to prove the statement for deterministic starting points $X_0 = z$. Let $\tau_0 := 0$ and $\tau_N := \inf\{k > \tau_{N-1} : X_k = z\}$ ($N \geq 1$) be the return

times of X to z and define random variables $(Y_N)_{N\geq 1}$ by

$$Y_N(x) := \sum_{k=\tau_{N-1}+1}^{\tau_N} 1_{\{(X_{k-1}, X_k) = (x_1, x_2)\}} \qquad (x \in S^2).$$

It is not hard to check that the $(Y_N)_{N\geq 1}$ are i.i.d. with finite mean $\mathbb{E}[Y_i(x_1, x_2)] = \mathbb{E}[\tau_1]\ \nu \otimes P(x_1, x_2)$ $((x_1, x_2) \in S^2)$, and the $(\tau_N - \tau_{N-1})_{N\geq 1}$ are i.i.d. with mean $\mathbb{E}[\tau_1]$. Therefore, by the ordinary strong law of large numbers

$$M^{(2)}_{\tau_N} = \frac{N}{\tau_N}\frac{1}{N}\sum_{M=1}^{N} Y_M \underset{N\to\infty}{\longrightarrow} \nu \otimes P \quad \text{a.s.}$$

The final part of the proof is a bit technical. For each $n \geq 0$, let $N(n) := \inf\{N \geq 1 : \tau_N \geq n\}$. Using Borel-Cantelli, one can check that for each $\varepsilon > 0$, the event

$$\{|M^{(2)}_n - M^{(2)}_{\tau_{N(n)}}| \geq \varepsilon\}$$

occurs only for finitely many n. Using this and the a.s. convergence of the $M^{(2)}_{\tau_{N(n)}}$ one obtains the a.s. convergence of the $M^{(2)}_n$. ∎

We will be interested in large deviations away from (4.2).

4.2 A LDP for Markov chains

In this section, we prove a basic large deviation result for the empirical pair distribution of irreducible Markov chains. For concreteness, for any finite set S, we equip the space $\mathcal{M}_1(S)$ of probability measures on S with the *total variation distance*

$$d(\mu, \nu) := \sup_{A\subset S} |\mu(A) - \nu(A)| = \tfrac{1}{2}\sum_{x\in S} |\mu(x) - \nu(x)|,$$

where for simplicity we write $\mu(x) := \mu(\{x\})$. Note that since S is finite, convergence in total variation norm is equivalent to weak convergence or pointwise convergence (and in fact any reasonable form of convergence one can think of).

For any $\nu \in \mathcal{M}_1(S^2)$, we let

$$\nu^1(x_1) := \sum_{x_2\in S} \nu(x_1, x_2) \quad \text{and} \quad \nu^2(x_2) := \sum_{x_1\in S} \nu(x_1, x_2)$$

denote the first and second marginals of ν, respectively, and we let

$$\mathcal{V} := \big\{\nu \in \mathcal{M}_1(S^2) : \nu^1 = \nu^2\big\}$$

denote the space of all probability measures on S^2 whose first and second marginals agree. The main result of this section is the following theorem.

Theorem 4.2 (LDP for Markov chains) *Let $X = (X_k)_{k\geq 0}$ be a Markov chain with finite state space S, irreducible transition kernel P, and arbitrary initial law. Let $(M_n^{(2)})_{n\geq 1}$ be the pair empirical distributions of X. Then the laws $\mathbb{P}[M_n^{(2)} \in \cdot\,]$ satisfy the large deviation principle with speed n and rate function $I^{(2)}$ given by*

$$I^{(2)}(\nu) := \begin{cases} H(\nu|\nu^1 \otimes P) & \text{if } \nu \in \mathcal{V}, \\ \infty & \text{otherwise,} \end{cases}$$

where $H(\,\cdot\,|\,\cdot\,)$ denotes the relative entropy of one measure w.r.t. another.

Remark By the contraction principle, Theorem 4.2 also gives us a large deviation principle for the 'usual' empirical distributions

$$M_n(x) := \frac{1}{n}\sum_{k=0}^{n-1} 1_{\{X_k = x\}} \qquad (x \in S,\ n \geq 1).$$

In this case, however, it is in general[1] not possible to write down a nice, explicit formula for the rate function. This is because pairs are the 'natural' object to look at for Markov processes.

The proof of Theorem 4.2 needs some preparations.

Lemma 4.3 (Characterization as invariant measures) *One has*

$$\mathcal{V} = \big\{\nu^1 \otimes P : \nu^1 \in \mathcal{M}_1(S),\ P \text{ a probability kernel on } S,\ \nu^1 P = \nu^1\big\}.$$

Proof If P is a probability kernel on S, and $\nu^1 \in \mathcal{M}_1(S)$ satisfies $\nu^1 P = \nu^1$ (i.e., ν^1 is an invariant law for the Markov chain with kernel P), then $(\nu^1 \otimes P)^2 = \nu^1 P = \nu^1$, which shows that $\nu^1 \otimes P \in \mathcal{V}$. On the other hand, for any $\nu \in \mathcal{V}$, we may define a kernel P by setting

$$P(x_1, x_2) := \frac{\nu(x_1, x_2)}{\nu^1(x_1)},$$

[1] An exception are continuous-time reversible Markov chains. See [Hol00, Thm. IV.14(b)].

whenever the denominator is nonzero, and choosing $P(x_1, \cdot\,)$ in some arbitrary way if $\nu^1(x_1) = 0$. Then $\nu^1 \otimes P(x_1, x_2) = \nu(x_1, x_2)$ and $\nu^1 P = (\nu^1 \otimes P)^2 = \nu^2 = \nu^1$ by the fact that $\nu \in \mathcal{V}$. ∎

For any $z \in S$, let us define

$$\mathcal{R}_{n,z} := \big\{ r \in \mathbb{N}^{S^2} : \exists (x_0, \ldots, x_n) \in S^{n+1},\ x_0 = z, \\ \text{s.t. } r(y_1, y_2) = \sum_{k=1}^{n} 1_{\{(x_{k-1}, x_k) = (y_1, y_2)\}}\ \forall y \in S^2 \big\}$$

and $\mathcal{R}_n := \bigcup_{z \in S} \mathcal{R}_{n,z}$. Then the random variables $N_n^{(2)}$ from (4.1) take values in $\mathcal{R}_n$. For the pair empirical distributions $M_n^{(2)}$, the relevant spaces are

$$\mathcal{V}_n := \{n^{-1} r : r \in \mathcal{R}_n\} \quad \text{and} \quad \mathcal{V}_{n,z} := \{n^{-1} r : r \in \mathcal{R}_{n,z}\}.$$

For any $U \subset S^2$, we identify the space $\mathcal{M}_1(U)$ of probability laws on U with the space

$$\big\{ \nu \in \mathcal{M}_1(S^2) : \nu(x_1, x_2) = 0\ \forall x \notin U \big\},$$

and we define

$$\mathcal{V}(U) := \mathcal{V} \cap \mathcal{M}_1(U), \quad \mathcal{V}_n(U) := \mathcal{V}_n \cap \mathcal{M}_1(U), \quad \text{and} \quad \mathcal{V}_{n,z}(U) := \mathcal{V}_{n,z} \cap \mathcal{M}_1(U).$$

We will need a lemma that says that for suitable $U \subset S^2$, the spaces $\mathcal{V}_n(U)$ approximate $\mathcal{V}(U)$ as $n \to \infty$. The typical example we have in mind is $U = \{(x_1, x_2) \in S^2 : P(x_1, x_2) > 0\}$ where P is an irreducible probability kernel on S or some subset of S. For any $U \subset S^2$, let us write

$$\begin{aligned} \overline{U} := &\{x_1 \in S : (x_1, x_2) \in U \text{ for some } x_2 \in S\} \\ &\cup \{x_2 \in S : (x_1, x_2) \in U \text{ for some } x_1 \in S\}. \end{aligned} \tag{4.3}$$

We will say that U is *irreducible* if for every $x, y \in \overline{U}$ there exist $n \geq 0$ and $x = x_0, \ldots, x_n = y$ such that $(x_{k-1}, x_k) \in U$ for all $k = 1, \ldots, n$.

Lemma 4.4 (Limiting space of pair empirical distribution) *One has*

$$\lim_{n \to \infty} \sup_{\nu \in \mathcal{V}_n} d(\nu, \mathcal{V}) = 0. \tag{4.4}$$

Moreover, for each $z \in S$ and $\nu \in \mathcal{V}$ there exist $\nu_n \in \mathcal{V}_{n,z}$ such that $d(\nu_n, \nu) \to 0$ as $n \to \infty$. If $U \subset S^2$ is irreducible, then moreover, for each $z \in \overline{U}$ and $\nu \in \mathcal{V}(U)$ there exist $\nu_n \in \mathcal{V}_{n,z}(U)$ such that $d(\nu_n, \nu) \to 0$ as $n \to \infty$.

Proof We leave formula (4.4) as an excercise to the reader (Excercise 4.5 below). To prove that for any $z \in S$ we can approximate arbitrary $\nu \in \mathcal{V}$ with $\nu_n \in \mathcal{V}_{n,z}$, by a simple diagonal argument (Exercise 4.6 below), we can without loss of generality assume that $\nu(x) > 0$ for all $x \in S^2$. By Lemma 4.3, there must exist some probability kernel P on S such that $\nu = \nu^1 \otimes P$ and $\nu^1 P = \nu^1$. Since $\nu(x) > 0$ for all $x \in S^2$, we must have $P(x_1, x_2) > 0$ for all $x \in S^2$. In particular, this implies that P is irreducible and ν^1 is the unique invariant law of P. Let $X = (X_k)_{k\geq 0}$ be a Markov chain with transition kernel P and initial state $X_0 = z$, and let $(M_n^{(2)})_{n\geq 1}$ be its pair empirical measures. Then $M_n^{(2)} \in \mathcal{V}_{n,z}$ for all $n \geq 1$ while $M_n^{(2)} \to \nu^1 \otimes P = \nu$ a.s. by Proposition 4.1. Since the empty set cannot have probability one, it follows that there must exist $\nu_n \in \mathcal{V}_{n,z}$ such that $d(\nu_n, \nu) \to 0$ as $n \to \infty$.

The same argument shows that if U is irreducible, then for any $z \in \overline{U}$, an arbitrary $\nu \in \mathcal{V}(U)$ can be approximated with $\nu_n \in \mathcal{V}_{n,z}(U)$. In this case, by a diagonal argument, we may assume without loss of generality that $\nu(x) > 0$ for all $x \in U$. By Lemma 4.3, there exists some probability kernel P on $\overline{U}$ such that $\nu = \nu^1 \otimes P$ and $\nu^1 P = \nu^1$. Since $\nu(x) > 0$ for all $x \in U$, we must have $P(x_1, x_2) > 0$ for all $x \in U$, hence P is irreducible. Using the strong law of large numbers for the Markov chain with transition kernel P, the argument then proceeds as before. ∎

Exercise 4.5 (Marginals almost agree) Prove formula (4.4).

Exercise 4.6 (Diagonal argument) Let (E, d) be a metric space, let $x_n, x \in E$ satisfy $x_n \to x$ and for each n, let $x_{n,m} \in E$ satisfy $x_{n,m} \to x_n$ as $m \to \infty$. Then there exist $m(n) \to \infty$ such that $x_{n,m'(n)} \to x$ for all $m'(n) \geq m(n)$.

Exercise 4.7 (Continuity of rate function) Let P be a probability kernel on S and let $U := \{(y_1, y_2) \in S^2 : P(y_1, y_2) > 0\}$. Prove the continuity of the map

$$\mathcal{M}_1(U) \ni \nu \mapsto H(\nu|\nu^1 \otimes P).$$

Show that if $U \neq S^2$, then the map $\mathcal{M}_1(S^2) \ni \nu \mapsto H(\nu|\nu^1 \otimes P)$ is *not* continuous.

Proof of Theorem 4.2 If μ_n, μ'_n both satisfy a large deviation principle with the same speed and rate function, then any convex combination of μ_n, μ'_n also satisfies this large deviation principle. In view of this, it suffices to prove the claim for Markov chains started in a deterministic initial state $X_0 = z$.

We observe that for any $r : S^2 \to \mathbb{N}$, the pair counting process defined in (4.1) satisfies

$$\mathbb{P}[N_n^{(2)} = r] = \mathcal{C}_{n,z}(r) \prod_{(x_1,x_2)\in S^2} P(x_1,x_2)^{r(x_1,x_2)}, \tag{4.5}$$

where

$$\mathcal{C}_{n,z}(r) := \Big|\{x \in S^{n+1} : x_0 = z,\ \sum_{k=1}^{n} 1_{\{(x_{k-1},x_k)=(y_1,y_2)\}} = r(y_1,y_2)\ \forall y \in S^2\}\Big|$$

is the number of different sequences $X_0, \ldots, X_n$ that give rise to the same pair frequencies $N_n^{(2)} = r$. In order to estimate $\mathcal{C}_{n,z}(r)$, for a given $r \in \mathcal{R}_{n,z}$, we draw a directed graph whose vertex set is S and that has $r(x_1,x_2)$ arrows pointing from x_1 to x_2. Let $\mathcal{W}_{n,z}(r)$ be the number of distinct walks in this graph that start at z and that use each arrow exactly once, where we distinguish between different arrows, i.e., if there are more arrows pointing from x_1 to x_2, then we do care about which arrow is used first, which arrow next, and so on. Then

$$\mathcal{C}_{n,z}(r) = \frac{\mathcal{W}_{n,z}(r)}{\prod_{(x_1,x_2)\in S^2} r(x_1,x_2)!}. \tag{4.6}$$

A simple combinatorical argument (see Lemma 4.8 below) shows that

$$\prod_{x_1:\, r^1(x_1)>0} (r^1(x_1)-1)! \le \mathcal{W}_{n,z}(r) \le \prod_{x_1\in S} r^1(x_1)! \qquad (r \in \mathcal{R}_n). \tag{4.7}$$

Combining (4.6), (4.7) and (4.5), we obtain the bounds

$$\begin{aligned} &\frac{\prod_{x_1:\, r(x_1)>0}(r^1(x_1)-1)!}{\prod_{(x_1,x_2)\in S^2} r(x_1,x_2)!} \prod_{(x_1,x_2)\in S^2} P(x_1,x_2)^{r(x_1,x_2)} \\ &\quad \le \mathbb{P}[N_n^{(2)} = r] \le \frac{\prod_{x_1\in S} r^1(x_1)!}{\prod_{(x_1,x_2)\in S^2} r(x_1,x_2)!} \prod_{(x_1,x_2)\in S^2} P(x_1,x_2)^{r(x_1,x_2)} \end{aligned} \tag{4.8}$$

$(r \in \mathcal{R}_{n,z})$. We recall that *Stirling's formula*[2] implies that

$$\log(n!) = n\log n - n + H(n) \qquad \text{as } n \to \infty,$$

where we use the convention that $0 \log 0 = 0$, and the error term $H(n)$ is of order $\log n$ and can in fact uniformly be estimated as

$$|H(n)| \le C \log n \qquad (n \ge 0),$$

[2]Recall that Stirling's formula says that $n! \sim \sqrt{2\pi n}(n/e)^n$.

with $C < \infty$ some constant. It follows that the logarithm of the right-hand side of (4.8) is given by

$$\begin{aligned}
&\sum_{x_1 \in S} \big(r^1(x_1) \log r^1(x_1) - r^1(x_1) + H(r^1(x_1))\big) \\
&\quad - \sum_{(x_1,x_2)\in S^2} \big(r(x_1,x_2) \log r(x_1,x_2) - r(x_1,x_2) + H(r(x_1,x_2))\big) \\
&\quad + \sum_{(x_1,x_2)\in S^2} r(x_1,x_2) \log P(x_1,x_2) \\
&\quad = \sum_{(x_1,x_2)\in S^2} r(x_1,x_2)\big(\log r^1(x_1) + \log P(x_1,x_2) - \log r(x_1,x_2)\big) + H'(r,n),
\end{aligned}$$

where we have used that $\sum_{x_1} r^1(x_1) = n = \sum_{(x_1,x_2)\in S^2} r(x_1,x_2)$ and $H'(r,n)$ is an error term that can be estimated uniformly in r as

$$\begin{aligned}
|H'(r,n)| &\leq \sum_{x_1\in S} C \log(r^1(x_1)) + \sum_{(x_1,x_2)\in S^2} C \log r(x_1,x_2) \\
&\leq C(|S| + |S|^2) \log n \qquad (n \geq 1,\ r \in \mathcal{R}_{n,z}),
\end{aligned}$$

with the same constant C as before. Dividing by n, we find that

$$\begin{aligned}
\frac{1}{n} \log \mathbb{P}[N_n^{(2)} = r] &\leq - \sum_{(x_1,x_2)\in S^2} \frac{r(x_1,x_2)}{n} \log \frac{r(x_1,x_2)}{r^1(x_1)P(x_1,x_2)} + \frac{1}{n} H'(r,n) \\
&= -H(\nu|\nu_r^1 \otimes P) + \frac{1}{n} H'(r,n),
\end{aligned}$$

where $\nu(x_1,x_2) := n^{-1} r(x_1,x_2)$. Treating the left-hand side of (4.8) in much the same way, we find that

$$\frac{1}{n} \log \mathbb{P}[M_n^{(2)} = \nu] = -H(\nu|\nu^1 \otimes P) + O(n^{-1} \log n) \tag{4.9}$$

for all $\nu \in \mathcal{V}_{n,z}$, where the error term is of order $n^{-1} \log n$ *uniformly* for all $\nu \in \mathcal{V}_{n,z}$.

We are now almost done. Let $U := \{(x_1,x_2) \in S^2 : P(x_1,x_2) > 0\}$. Then obviously $M_n^{(2)} \in \mathcal{M}_1(U)$ for all $n \geq 1$, hence by the restriction principle (Lemma 1.28) and the fact that $H(\nu|\nu^1 \otimes P) = \infty$ for all $\nu \notin \mathcal{M}_1(U)$, instead of proving the large deviation principle on $\mathcal{M}_1(S^2)$, we may equivalently prove the large deviation principle on $\mathcal{M}_1(U)$. By Excercise 4.7, the map

$$\mathcal{M}_1(U) \ni \nu \mapsto H(\nu|\nu^1 \otimes P)$$

is continuous. (Note that we need the space $\mathcal{M}_1(U)$ since the same is not true for $\mathcal{M}_1(S^2) \ni \nu \mapsto H(\nu|\nu^1 \otimes P)$.) Using the continuity of this map and Lemmas 1.15 and 1.18, we see that it suffices to show that the counting measures on $\mathcal{V}_{n,z}(U)$

$$\rho_n := \sum_{\nu \in \mathcal{V}_{n,z}(U)} \delta_\nu$$

satisfy the large deviation principle on $\mathcal{M}_1(U)$ with speed n and trivial rate function

$$J(\nu) := \begin{cases} 0 & \text{if } \nu \in \mathcal{V}(U), \\ \infty & \text{otherwise.} \end{cases}$$

We will prove the large deviations upper and lower bounds from Proposition 1.7. For the upper bound, we observe that if $C \subset \mathcal{M}_1(U)$ is closed and $C \cap \mathcal{V}(U) = \emptyset$, then, since $\mathcal{V}(U)$ is a compact subset of $\mathcal{M}_1(U)$, the distance $d(C, \mathcal{V}(U))$ must be strictly positive. By Lemma 4.4, it follows that $C \cap \mathcal{V}_n(U) = \emptyset$ for n sufficiently large and hence $\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[M_n^{(2)} \in C] = \infty$. If $C \cap \mathcal{V}(U) \neq \emptyset$, then we may use the fact that $|\mathcal{V}_n| \leq n^{|S|^2}$, to obtain the crude estimate

$$\limsup_{n\to\infty} \frac{1}{n} \log \rho_n(C) \leq \limsup_{n\to\infty} \frac{1}{n} \log \rho_n(\mathcal{V}) \leq \lim_{n\to\infty} \frac{1}{n} \log \left(n^{|S|^2}\right) = 0,$$

which completes our proof of the large deviations upper bound. To prove also the large deviations lower bound, let $O \subset \mathcal{M}_1(U)$ be open and let $O \cap \mathcal{V}(U) \neq \emptyset$ (otherwise the statement is trivial). Pick any $\nu \in O \cap \mathcal{V}(U)$. By Lemma 4.4, we can choose $\nu_n \in \mathcal{V}_{n,z}(U)$ such that $\nu_n \to \nu$. It follows that $\nu_n \in O$ for n sufficiently large, and hence

$$\liminf_{n\to\infty} \frac{1}{n} \log \rho_n(O) \geq \lim_{n\to\infty} \frac{1}{n} \log \rho_n(\{\nu_n\}) = 0,$$

as required. ∎

We still need to prove the estimates (4.7). Let $G = (V, E)$ be a finite directed graph with vertex set V and set of directed edges E. For each edge $e \in E$ there is defined a starting vertex $e^- \in V$ and endvertex $e^+ \in V$. We allow for the case that $e^- = e^+$ (in this case, e is called a loop). We write

$$E_{x,\bullet} := \{e \in E : e^- = x\}, \quad E_{\bullet,y} := \{e \in E : e^+ = y\}, \quad \text{and} \quad E_{x,y} := E_{x,\bullet} \cap E_{\bullet,y}$$

for the sets of all edges with a specified starting vertex, or endvertex, or both. We allow for the case that $r(x, y) := |E_{x,y}|$ is larger than one.

By definition, a *walk* is an ordered collection of edges $(e_1, \ldots, e_n)$ such that $e_k^+ = e_{k+1}^-$ for $k = 1, \ldots, n-1$. We call e_1^- and e_n^+ the starting vertex and endvertex of the walk. For any subset of edges $F \subset E$, we write $x \rightsquigarrow_F y$ if $x = y$ or there exists a walk using only edges from F with starting vertex x and endvertex y. By definition, a (directed) *spanning tree* rooted at $z \in V$ is a collection of edges $T \subset E$ such that $|T \cap E_{x,\bullet}| = 1$ and $x \rightsquigarrow_T z$ for all $x \in V$, i.e., from each vertex there is a unique directed path to the root.

Lemma 4.8 (Walks that use all edges) *Let $G = (V, E)$ be a finite directed graph and let $y, z \in V$. Write $r(x_1, x_2) := |E_{x_1,x_2}|$, $r^1(x_1) := |E_{x_1,\bullet}|$, and $r^2(x_2) := |E_{\bullet,x_2}|$ $(x_1, x_2 \in S)$. Assume that $r^1(x) > 0$ for each $x \in V$ and that*

$$r^1(x) - r^2(x) = 1_{\{x=y\}} - 1_{\{x=z\}} \qquad (x \in V). \tag{4.10}$$

Let $\mathcal{W}$ denote the number of walks in G that end in z and use each edge exactly once. Let $\mathcal{T}$ denote the number of spanning trees rooted at z. Then

$$\mathcal{W} = \mathcal{T} r^1(z) \prod_{x \in V} \big(r^1(x) - 1\big)! \tag{4.11}$$

In particular, one has the estimates (4.7).

Proof Let W denote the set of all walks w in G that end in z and use each edge exactly once. It follows from (4.10) that each $w \in W$ must start in y. We can encode such a walk by numbering, for each $x \in V$, the set of outgoing edges $E_{x,\bullet}$ at x according to which edges is used first, second etc. Let Π be the collection of all functions $\pi : E \to \mathbb{N}_+$ such that $\pi : E_{x,\bullet} \to \{1, \ldots, r^1(x)\}$ is a bijection for each $x \in V$. We say that such a function π encodes a walk $w \in W$ if for each $x \in V$ and $e \in E_{x,\bullet}$, one has $\pi(e) = k$ iff w leaves x for the k-th time using the edge e. Clearly, $\mathcal{W} = |W| \leq |\Pi|$ which yields the upper bound in (4.7). For any $\pi \in \Pi$, let $T_\pi := \bigcup_{x \in V \setminus \{z\}} \{e \in E_x : \pi(e) = r^1(x)\}$. In particular, if π encodes a walk $w \in W$, then these are the arrows used when the walk leaves a vertex $\neq z$ for the last time. We claim that:

- A function $\pi \in \Pi$ encodes a walk $w \in W$ if and only if T_π is a spanning tree rooted at z.

Indeed, given a walk $w \in W$, if for a vertex $\neq z$, we follow the arrow used when w last leaves this vertex, and so on for the next vertex, then we end up in z, proving that T_π is a spanning tree rooted at z. Conversely, if a function $\pi \in \Pi$ has the

property that T_π is a spanning tree rooted at z, then, starting at y, we can walk around through the graph in such a way that if at a given moment we are in a vertex x, then we leave x using the outgoing edge $e \in E_{x,\bullet}$ with the lowest number $\pi(e)$ that has not yet been used. This process stops when we arrive at a vertex such that all outgoing edges at this vertex have been used. By (4.10), it follows that all incoming arrows have also been used, which is possible only if we are in z. We observe that if $e \in T_\pi$ has been used, then all arrows in $E_{e^-,\bullet}$ have been used and hence by (4.10) also all arrows in $E_{\bullet,e^-}$ have been used. Since all arrows in $E_{\bullet,z}$ have been used and T_π is a spanning tree rooted at z, it follows that all arrows in T_π have been used, which implies that all arrows in E have been used, i.e., $w \in W$.

This completes the proof of (4.11). In particular, fixing one spanning tree rooted at z, in each vertex $x \neq z$ we have $(r^1(x) - 1)!$ ways to choose the order of the outgoing edges except for the one that is used last, which yields the lower bound in (4.7). (Note that in (4.7), we apply Lemma 4.8 to the subgraph consisting of all vertices of G that have been visited at least once.) ∎

The proof of Theorem 4.2 yields a useful corollary. Below, we use the notation

$$H(\nu|\mu) := \sum_{x\in S} \nu(x) \log \frac{\nu(x)}{\mu(x)} = \sum_{x\in S} \mu(x) \frac{\nu(x)}{\mu(x)} \log \frac{\nu(x)}{\mu(x)},$$

even if μ is not a probability measure. Note that below, the transition kernel P need not be irreducible!

Corollary 4.9 (Restricted Markov process) *Let $X = (X_k)_{k\geq 0}$ be a Markov chain with finite state space S, transition kernel P, and arbitrary initial law. Let*

$$U \subset \{(x_1, x_2) \in S^2 : P(x_1, x_2) > 0\}$$

be irreducible and let $X_0 \in \overline{U}$ a.s. Let $(M_n^{(2)})_{n\geq 1}$ be the pair empirical distributions of X and let $\tilde{P}$ denote the restriction of P to U. Then the restricted measures

$$\mathbb{P}[M_n^{(2)} \in \cdot\,]\big|_{\mathcal{M}_1(U)}$$

satisfy the large deviation principle with speed n and rate function $I^{(2)}$ given by

$$\tilde{I}^{(2)}(\nu) := \begin{cases} H(\nu|\nu^1 \otimes \tilde{P}) & \text{if } \nu \in \mathcal{V}(U), \\ \infty & \text{otherwise.} \end{cases}$$

Proof The restricted measures $\mathbb{P}[M_n^{(2)} \in \cdot\,]\big|_{\mathcal{M}_1(U)}$ are no longer probability measures, but we have never used this in the proof of Theorem 4.2. In fact, a careful inspection reveals that the proof carries over without a change, where we only need the irreducibility of U (but not of P). In particular, formula (4.9) also holds for the restricted measures and the arguments below there work for any irreducible $U \subset \{(x_1, x_2) \in S^2 : P(x_1, x_2) > 0\}$. ∎

Exercise 4.10 (Relative entropy and conditional laws) Let S be a finite space, let ν, μ be probability measures on S and let Q, P be probability kernels on S. Show that

$$H(\nu \otimes Q|\mu \otimes P) = H(\nu|\mu) + \sum_{x_1 \in S} \nu(x_1) H(Q_{x_1}|P_{x_1}),$$

where $Q_{x_1}(x_2) := Q(x_1, x_2)$ and $P_{x_1}(x_2) := P(x_1, x_2)$ ($(x_1, x_2) \in S^2$). In particular, if Q is a probability kernel such that $\nu = \nu^1 \otimes Q$, then

$$H(\nu|\nu^1 \otimes P) = \sum_{x_1 \in S} \nu^1(x_1) H(Q_{x_1}|P_{x_1}).$$

Exercise 4.11 (Minimizer of the rate function) Let P be irreducible. Show that the unique minimizer of the function $\mathcal{V} \ni \nu \mapsto H(\nu|\nu^1 \otimes P)$ is given by $\nu = \mu \otimes P$, where μ is the invariant law of P.

By definition, a *cycle* in S is an ordered collection $C = (x_1, \ldots, x_n)$ of points in S such that $x_1, \ldots, x_n$ are all different. We call two cycles equal if they differ only by a cyclic permutation of their points and we call $|C| = n \geq 1$ the *length* of a cycle $C = (x_1, \ldots, x_n)$. We write $(y_1, y_2) \in C$ if $(y_1, y_2) = (x_{k-1}, x_k)$ for some $k = 1, \ldots, n$, where $x_0 := x_n$.

Recall that an element x of a convex set K is an *extremal element* if x cannot be written as a nontrivial convex combination of other elements of K, i.e., there do not exist $y, z \in K$, $y \neq z$ and $0 < p < 1$ such that $x = py + (1-p)z$. If $K \subset \mathbb{R}^d$ is convex and compact, then it is known that for each element $x \in K$ there exists a unique probability measure ρ on the set K_{e} of extremal elements of K such that $x = \int y \rho(\mathrm{d}y)$.

Exercise 4.12 (Cycle decomposition) Prove that the extremal elements of the space $\mathcal{V}$ are the probability measures of the form

$$\nu_C(y_1, y_2) := \frac{1}{|C|} 1\{(y_1, y_2) \in C\},$$

where $C = (x_1, \ldots, x_n)$ is a cycle in S. Hint: show that for each $\nu \in \mathcal{V}$ and $(y_1, y_2) \in S^2$ such that $\nu(y_1, y_2) > 0$, one can find a cycle $C \in \mathcal{C}(S^2)$ and a constant $c > 0$ such that $(y_1, y_2) \in C$ and $c\nu_C \leq \nu$. Use this to show that for each $\nu \in \mathcal{V}_{\rm e}$ there exists a cycle C such that $\nu(y_1, y_2) = 0$ for all $(y_1, y_2) \notin C$.

Note Since $\mathcal{V}$ is a finite dimensional, compact, convex set, Excercise 4.12 shows that for each $\nu \in \mathcal{V}$, there exists a unique probability law ρ on the set of all cycles in S such that

$$\nu(y_1, y_2) = \sum_C \rho(C)\nu_C(y_1, y_2),$$

where the sum rums over al cycles in S. Note that in Excercise 4.12, you are not asked to give an explicit formula for ρ.

Exercise 4.13 (Convexity of rate function (!)) Let P be a probability kernel on S. Prove that

$$\mathcal{M}_1(S^2) \ni \nu \mapsto H(\nu|\nu^1 \otimes P)$$

is a convex, lower semi-continuous function.

Important note I do not know an elegant solution to this exercise. I originally copied this from [Hol00], who first gives the special case that $P(x, y) = \mu(y)$ does not depend on x as Exercise II.12, and then in his Lemma IV.5 shows that the general case can easily be derived from this special case. Den Hollander probably based himself on Problems IX 6.1 and 6.2 from [Ell85]. These problems, however, are meant to be solved using deep theory from Chapter IX of [Ell85] that is not available here or in [Hol00].

Recall that if $(X_k)_{k\geq 0}$ is a Markov chain with initial law μ and transition kernel P, then $\mu \otimes P$ is the joint law of (X_0, X_1). More generally, let $\mu \otimes^n P$ denote the joint law of $(X_0, \ldots, X_{n-1})$. Let μ, ν be invariant laws of probability kernels P, Q, respectively. It seems that if P is irreducible, then

$$H(\nu \otimes Q|\nu \otimes P) = \lim_{n\to\infty} \frac{1}{n} H(\nu \otimes^n Q|\mu \otimes^n P).$$

Now if ν' is an invariant law of Q', and $p \in [0, 1]$, then

$$\begin{aligned} &pH\big(\nu \otimes Q \,\big|\, \nu \otimes P\big) + (1-p)H\big(\nu' \otimes Q' \,\big|\, \nu' \otimes P\big) \\ &\quad = \lim_{n\to\infty} \frac{1}{n} H\big(p\nu \otimes^n Q + (1-p)\nu' \otimes^n Q' \,\big|\, \mu \otimes^n P\big). \end{aligned} \tag{4.12}$$

On the other hand, writing

$$\nu'' \otimes Q'' := p\nu \otimes Q + (1-p)\nu' \otimes Q',$$

we have that

$$H\big(p\nu\otimes Q+(1-p)\nu'\otimes Q' \,\big|\, (p\nu+(1-p)\nu')\otimes P\big) = \lim_{n\to\infty} \frac{1}{n} H\big(\nu''\otimes^n Q'' \,\big|\, \mu\otimes^n P\big). \quad (4.13)$$

If $\nu'' \otimes^n Q'' \neq p\nu \otimes^n Q + (1-p)\nu' \otimes^n Q'$, then it seems that the expression in (4.13) is strictly smaller than the expression in (4.12).

Results in this spirit are proved in Chapter IX of [Ell85]. In particular, it is shown there that if one wishes to minimize the relative entropy density of a stationary measure with respect to a product measure, under the condition that the two-dimensional marginals are given by some $\nu \in \mathcal{M}_1(S^2)$, then the minimum is attained by the Markov chain that has these two-dimensional marginals. In Problem IX 6.1 of [Ell85], this and the contraction principle are used to prove the convexity and lower semi-continuity of of $\nu \mapsto H(\nu|\nu^1 \otimes P)$.

Exercise 4.14 (Not strictly convex) Let P be any probability kernel on $S = \{1,2\}$. Define $\mu, \nu \in \mathcal{M}_1(S^2)$ by

$$\begin{pmatrix} \mu(1,1) & \mu(1,2) \\ \mu(2,1) & \mu(2,2) \end{pmatrix} := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \nu(1,1) & \nu(1,2) \\ \nu(2,1) & \nu(2,2) \end{pmatrix} := \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Define $\nu_p := p\mu + (1-p)\nu$. Show that

$$[0,1] \ni p \mapsto H(\nu_p|\nu_p^1 \otimes P)$$

is an affine function. Prove the same statement for

$$\mu := \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \nu := \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \end{pmatrix}.$$

These examples show that $\mathcal{M}_1(S^2) \ni \nu \mapsto H(\nu|\nu^1 \otimes P)$ is not strictly convex. Do you see a general pattern how to create such examples? Hint: Excercise 4.10.

Exercise 4.15 (Probability to stay inside a set) Let P be a probability kernel on $\{0, 1, \ldots, n\}$ $(n \geq 1)$ such that $P(x, y) > 0$ for all $1 \leq x \leq n$ and $0 \leq y \leq n$

but $P(0,y) = 0$ for all $1 \le y \le n$. (In particular, 0 is a *trap* of the Markov chain with transition kernel P.) Show that there exists a constant $0 < \lambda < \infty$ such that the Markov chain $(X_k)_{k\ge 0}$ with transition kernel P and initial state $X_0 = z \ge 1$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[X_n \ge 1] = -\lambda.$$

Give a (formal) expression for λ and show that λ does not depend on z. Hint: Corollary 4.9.

4.3 The empirical process

In this section, we return to the i.i.d. setting, but rather than looking at the (standard) empirical distributions as we did in Section 3.4, we will look at pair empirical distributions and more general at empirical distributions of k-tuples. Since i.i.d. sequences are a special case of Markov processes, our results from the previous section immediately give us the following theorem.

Theorem 4.16 (Sanov for pair empirical distributions)
(a) *Let S be a finite set and let μ be a probability measure on S such that $\mu(x) > 0$ for all $x \in S$. Let $(X_k)_{k\ge 0}$ be i.i.d. with common law μ and let $M_n^{(2)}$ be their pair empirical distributions as defined in (4.1). Then the laws $\mathbb{P}[M_n^{(2)} \in \cdot\,]$ satisfy the large deviation principle with speed n and rate function $I^{(2)}$ given by*

$$I^{(2)}(\nu) := \begin{cases} H(\nu|\nu^1 \otimes \mu) & \text{if } \nu^1 = \nu^2, \\ \infty & \text{otherwise,} \end{cases}$$

where ν^1 and ν^2 denote the first and second marginal of ν, respectively, and $H(\,\cdot\,|\,\cdot\,)$ denotes the relative entropy of one measure w.r.t. another.

(b) *More generally, if $U \subset S^2$ is irreducible, then the restricted measures*

$$\mathbb{P}[M_n^{(2)} \in \cdot\,]\big|_{\mathcal{M}_1(U)}$$

satisfy the large deviation principle with speed n and rate function $I^{(2)}$ given by

$$I^{(2)}(\nu) := \begin{cases} H\big(\nu\big|\ [\nu^1 \otimes \mu]_U\big) & \text{if } \nu^1 = \nu^2, \\ \infty & \text{otherwise,} \end{cases}$$

where $[\nu^1 \otimes \mu]_U$ denotes the restriction of the product measure $\nu^1 \otimes \mu$ to U.

Proof Immediate from Theorem 4.2 and Corollary 4.9. ∎

Exercise 4.17 (Sanov's theorem) Show that through the contraction principle, Theorem 4.16 (a) implies Sanov's theorem (Theorem 3.11) for finite state spaces.

Although Theorem 4.16, which is a statement about i.i.d. sequences only, looks more special that Theorem 4.2 and Corollary 4.9 which apply to general Markov chains, the two results are in fact more or less equivalent.

Derivation of Theorem 4.2 from Theorem 4.16 We first consider the special case that $P(x_1, x_2) > 0$ for all $(x_1, x_2) \in S^2$. Let ρ be the initial law of X, let μ be any probability measure on S satisfying $\mu(x) > 0$ for all $x \in S$, and let $\hat{X} = (\hat{X}_k)_{k\geq 0}$ be independent random variables such that $\hat{X}_0$ has law ρ and $\hat{X}_k$ has law μ for all $k \geq 1$. For any $x = (x_k)_{k\geq 0}$ with $x_k \in S$ $(k \geq 0)$, let us define $M_n^{(2)}(x) \in \mathcal{M}_1(S^2)$ by

$$M_n^{(2)}(x)(y_1, y_2) := \frac{1}{n}\sum_{k=1}^{n} 1_{\{(x_{k-1}, x_k) = (y_1, y_2)\}}.$$

We observe that

$$\begin{aligned}\mathbb{P}[X_0 = x_0, \ldots, X_n = x_n] &= \rho(x_0)\, e^{\sum_{k=1}^{n} \log P(x_{k-1}, x_k)} \\ &= \rho(x_0)\, e^{n \sum_{(y_1,y_2)\in S^2} \log P(y_1, y_2) M_n^{(2)}(x)(y_1, y_2)},\end{aligned}$$

while

$$\begin{aligned}\mathbb{P}[\hat{X}_0 = x_0, \ldots, \hat{X}_n = x_n] &= \rho(x_0)\, e^{\sum_{k=1}^{n} \log \mu(x_k)} \\ &= \rho(x_0)\, e^{n \sum_{(y_1,y_2)\in S^2} \log \mu(y_2) M_n^{(2)}(x)(y_1, y_2)}.\end{aligned}$$

It follows that the Radon-Nikodym derivative of $\mathbb{P}[M_n^{(2)}(X) \in \cdot\,]$ with respect to $\mathbb{P}[M_n^{(2)}(\hat{X}) \in \cdot\,]$ is given by

$$\frac{\mathbb{P}[M_n^{(2)}(X) = \nu]}{\mathbb{P}[M_n^{(2)}(\hat{X}) = \nu]} = e^{n \sum_{(y_1,y_2)\in S^2} \big(\log P(y_1, y_2) - \log \mu(y_2)\big)\nu(y_1, y_2)}.$$

By Theorem 4.16 (a), the laws $\mathbb{P}[M_n^{(2)}(\hat{X}) \in \cdot\,]$ satisfy the large deviation principle with speed n and rate function $\hat{I}^{(2)}$ given by

$$\hat{I}^{(2)}(\nu) = \begin{cases} H(\nu|\nu^1 \otimes \mu) & \text{if } \nu^1 = \nu^2, \\ \infty & \text{if } \nu^1 \neq \nu^2. \end{cases}$$

Applying Lemma 1.15 to the function

$$F(\nu) := \sum_{(y_1,y_2)\in S^2} \big(\log P(y_1,y_2) - \log \mu(y_2)\big)\nu(y_1,y_2),$$

which is continuous by our assumption that $P(y_1,y_2) > 0$ for all $y_1, y_2 \in S$, we find that the laws $\mathbb{P}[M_n^{(2)}(\hat{X}) \in \cdot\,]$ satisfy the large deviation principle with speed s_n and rate function $I^{(2)} = \hat{I}^{(2)} - F$. Since

$$\begin{aligned}
&H(\nu|\nu^1 \otimes \mu) - F(\nu) \\
&\quad = \sum_{(y_1,y_2)\in S^2} \nu(y_1,y_2)\big(\log \frac{\nu(y_1,y_2)}{\nu^1(y_1)\mu(y_2)} + \log \mu(y_2) - \log P(y_1,y_2)\big) \\
&\quad = \sum_{(y_1,y_2)\in S^2} \nu(y_1,y_2) \log \frac{\nu(y_1,y_2)}{\nu^1(y_1)P(y_1,y_2)} = H(\nu|\nu^1 \otimes P),
\end{aligned}$$

this proves the theorem.

In the general case, when P is irreducible but not everywhere positive, the argument is the same but we need to apply Theorem 4.16 (b) to $U := \{(x_1,x_2) \in S^2 : P(x_1,x_2) > 0\}$, and we use that the function F restricted to $\mathcal{M}_1(U)$ is continuous, hence Lemma 1.15 is applicable. ∎

Exercise 4.18 (Periodic boundary conditions) Let $(X_k)_{k\geq 0}$ be i.i.d. with common law $\mu \in \mathcal{M}_1(S)$. Let $M_n^{(2)}$ be the pair empirical distributions defined in (4.1) and set

$$\begin{aligned}
&\tilde{M}_n^{(2)} := \frac{1}{n}\tilde{N}_n^{(2)}, \quad \text{where} \\
&\quad \tilde{N}_n^{(2)}(x) := 1_{\{(X_n,X_1) = (x_1,x_2)\}} + \sum_{k=2}^{n} 1_{\{(X_{k-1},X_k) = (x_1,x_2)\}}
\end{aligned} \tag{4.14}$$

Show that the random variables $M_n^{(2)}$ and $\tilde{M}_n^{(2)}$ are exponentially close in the sense of (1.8), hence by Proposition 1.17, proving a large deviation principle for the $M_n^{(2)}$ is equivalent to proving one for the $\tilde{M}_n^{(2)}$.

Remark Den Hollander [Hol00, Thm II.8] who again follows [Ell85, Sect. I.5], gives a very nice and short proof of Sanov's theorem for the pair empirical distributions using periodic boundary conditions. The advantage of this approach is that the pair empirical distributions $\tilde{M}_n^{(2)}$ defined in (4.14) automatically have the property

that their first and second marginals agree, which means that one does not need to prove formula (4.4).

Based on this, along the lines of the proof above, Den Hollander [Hol00, Thm IV.3] then derives Theorem 4.2 in the special case that the transition kernel P is everywhere positive. In [Hol00, Comment (4) from Section IV.3], it is then claimed that the theorem still applies when P is not everywhere positive but irreducible and S^2 is replaced by $U := \{(x_1, x_2) \in S^2 : P(x_1, x_2) > 0\}$, and 'the proof is easily adapted'. This last comment seems to be quite far from the truth. At least, I do not see any *easy* way to adapt his proof. The reason is that periodic boundary conditions do not work well anymore if S^2 is replaced by a more general subset $U \subset S^2$. As a result, the technicalities needed to prove the analogue of Lemma 4.4 in a set-up with periodic boundary conditions become very unpleasant. Although a proof along these lines is possible, this seems to be more complicated than the approach used in these lecture notes.

The fact that Theorem 4.2 can rather easily be derived from Theorem 4.16 shows that the point of view that Chapter 3 is about large deviations of independent random variables while the present chapter is about large deviations of Markov chains is naive. With equal right, we might say that both chapters are concerned with large deviations of functions of i.i.d. random variables. The essential difference is in what kind of functions we consider. In Chapter 3, we considered the empirical distributions and functions thereof (such as the mean), while in the present chapter we consider the pair empirical distributions. By looking at yet different functions of i.i.d. random variables one can obtain a lot of very different, often difficult, but interesting large deviation principles.

There is no need to restrict ourselves to pairs. In fact, once we have a theorem for pairs, the step to general m-tuples is easy. (In contrast, there seems to be no easy way to derive the result for pairs from the large deviation principle for singletons.)

Theorem 4.19 (Sanov for empirical distributions of m-tuples) *Let S be a finite set and let μ be a probability measure on S such that $\mu(x) > 0$ for all $x \in S$. Let $(X_k)_{k\geq 1}$ be i.i.d. with common law μ and for fixed $m \geq 1$, define*

$$M_n^{(m)}(x) := \frac{1}{n}\sum_{k=0}^{n-1} 1_{\{(X_{k+1},\ldots,X_{k+m}) = x\}} \qquad (x \in S^m,\ n \geq 1).$$

Then the laws $\mathbb{P}[M_n^{(m)} \in \cdot\,]$ satisfy the large deviation principle with speed n and

rate function $I^{(m)}$ *given by*

$$I^{(m)}(\nu) := \begin{cases} H(\nu|\nu^{\{1,\ldots,m-1\}} \otimes \mu) & \text{if } \nu^{\{1,\ldots,m-1\}} = \nu^{\{2,\ldots,m\}}, \\ \infty & \text{otherwise,} \end{cases}$$

where $\nu^{\{1,\ldots,m-1\}}$ *and* $\nu^{\{2,\ldots,m\}}$ *denote the projections of* ν *on its first* $m-1$ *and last* $m-1$ *coordinates, respectively.*

Proof The statement for $m = 1, 2$ has already been proved in Theorems 3.11 and 4.2, respectively, so we may assume that $m \geq 3$. Define a probability kernel $P : S^{m-1} \to S^{m-1}$ by

$$P(x, y) := 1_{\{(x_2, \ldots, x_{m-1}) = (y_1, \ldots, y_{m-2})\}} \mu(y_{m-1}) \qquad (x, y \in S^{m-1}),$$

and set

$$\vec{X}_k := \big(X_{k+1}, \ldots, X_{k+m-1}\big) \qquad (k \geq 0).$$

Then $\vec{X} = (\vec{X}_k)_{k\geq 0}$ is a Markov chain with irreducible transition kernel P. By Theorem 4.2, the pair empirical distributions $\vec{M}_n^{(2)}$ of $\vec{X}$ satisfy a large deviation principle. Here the $\vec{M}_n^{(2)}$ take values in the space $\mathcal{M}_1(S^{m-1} \times S^{m-1})$ and the rate function is given by

$$\vec{I}^{(2)}(\rho) := \begin{cases} H(\rho|\rho^1 \otimes P) & \text{if } \rho^1 = \rho^2, \\ \infty & \text{otherwise,} \end{cases}$$

where ρ^1 and ρ^2 denote the first and second marginals of ρ, respectively. (Note that ρ is a probability measure on $S^{m-1} \times S^{m-1}$, hence ρ^1 and ρ^2 are probability measures on S^{m-1}.)

Define a map $\psi : S^m \to S^{m-1} \times S^{m-1}$ by

$$\psi(x_1, \ldots, x_m) := \big((x_1, \ldots, x_{m-1}), (x_2, \ldots, x_m)\big).$$

The image of S^m under ψ is the set

$$\begin{aligned} U &:= \big\{(x, y) \in S^{m-1} \times S^{m-1} : (x_2, \ldots, x_{m-1}) = (y_1, \ldots, y_{m-2})\big\} \\ &= \big\{(x, y) \in S^{m-1} \times S^{m-1} : P(x, y) > 0\big\}. \end{aligned}$$

It follows that $\vec{I}^{(2)}(\rho) = \infty$ unless $\rho \in \mathcal{M}_1(U)$. Since $\psi : S^m \to U$ is a bijection, each $\rho \in \mathcal{M}_1(U)$ is the image under ψ of a unique $\nu \in \mathcal{M}_1(S^m)$. Moreover,

$\rho^1 = \rho^2$ if and only if $\nu^{\{1,\dots,m-1\}} = \nu^{\{2,\dots,m\}}$. Thus, by the contraction principle (Proposition 1.14), our claim will follow provided we show that if $\nu \in \mathcal{M}_1(S^m)$ satisfies $\nu^{\{1,\dots,m-1\}} = \nu^{\{2,\dots,m\}}$ and $\rho = \nu \circ \psi^{-1}$ is the image of ν under ρ, then

$$H(\nu|\nu^{\{1,\dots,m-1\}} \otimes \mu) = H(\rho|\rho^1 \otimes P).$$

Here

$$\begin{aligned} H(\rho|\rho^1 \otimes P) = & \sum_{\substack{x_1,\dots,x_{m-1} \\ y_1,\dots,y_{m-1}}} \rho(x_1,\dots,x_{m-1},y_1,\dots,y_{m-1}) \\ & \times \Big(\log \rho(x_1,\dots,x_{m-1},y_1,\dots,y_{m-1}) - \log \rho^1(x_1,\dots,x_{m-1}) \\ & \quad - \log P(x_1,\dots,x_{m-1},y_1,\dots,y_{m-1}) \Big), \end{aligned}$$

where

$$\rho(x_1,\dots,x_{m-1},y_1,\dots,y_{m-1}) = 1_{\{(x_2,\dots,x_{m-1})=(y_1,\dots,y_{m-2})\}} \nu(x_1,\dots,x_{m-1},y_{m-1}),$$

$$\rho^1(x_1,\dots,x_{m-1}) = \nu^{\{1,\dots,m-1\}}(x_1,\dots,x_{m-1}),$$

and

$$P(x_1,\dots,x_{m-1},y_1,\dots,y_{m-1}) = 1_{\{(x_2,\dots,x_{m-1})=(y_1,\dots,y_{m-2})\}} \mu(y_{m-1}).$$

It follows that

$$\begin{aligned} H(\rho|\rho^1 \otimes P) = & \sum_{x_1,\dots,x_{m-1},y_{m-1}} \nu(x_1,\dots,x_{m-1},y_{m-1}) \\ & \times \Big(\log \nu(x_1,\dots,x_{m-1},y_{m-1}) - \log \nu^{\{1,\dots,m-1\}}(x_1,\dots,x_{m-1}) - \log \mu(y_{m-1}) \Big) \\ = & H(\nu|\nu^{\{1,\dots,m-1\}} \otimes \mu). \end{aligned}$$

∎

It is even possible to go one step further than Theorem 4.19 and prove a large deviations result for 'm-tuples' with $m = \infty$. Let $S^{\mathbb{N}}$ be the space of all infinite sequence $x = (x_k)_{k \geq 0}$ with $x \in S$. Note that $S^{\mathbb{N}}$, equipped with the product topology, is a compact metrizable space. Define a shift operator $\theta : S^{\mathbb{N}} \to S^{\mathbb{N}}$ by

$$(\theta x)_k := x_{k+1} \qquad (k \geq 0).$$

Let $X = (X_k)_{k \geq 0}$ be i.i.d. random variables with values in S and common law μ satisfying $\mu(x) > 0$ for all $x \in S$. For each $n \geq 1$, we define a random measure $M_n^{(\infty)}$ on $S^{\mathbb{N}}$ by

$$M_n^{(\infty)} := \frac{1}{n} \sum_{k=0}^{n-1} \delta_{\theta^k X},$$

where δ_x denotes the delta measure at a point x. We call $M_n^{(\infty)}$ the *empirical process*.

Exercise 4.20 (Empirical process) Sketch a proof of the fact that the laws $\mathbb{P}[M_n^{(\infty)} \in \cdot\,]$ satisfy a large deviation principle. Hint: projective limit.

Exercise 4.21 (First occurrence of a pattern) Let $(X_k)_{k\geq 0}$ be i.i.d. random variables with $\mathbb{P}[X_k = 0] = \mathbb{P}[X_k = 1] = \frac{1}{2}$. Give a formal expression for the limits

$$\lambda_{001} := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[(X_k, X_{k+1}, X_{k+2}) \neq (0,0,1)\ \forall k = 1, \ldots, n\big]$$

and

$$\lambda_{000} := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[(X_k, X_{k+1}, X_{k+2}) \neq (0,0,0)\ \forall k = 1, \ldots, n\big].$$

4.4 Perron-Frobenius eigenvalues

In excercises such as Excercise 4.21, we need an explicit way to determine the exponential rates associated with certain events or expectations of exponential functions in the spirit of Varadhan's lemma. In this section, we will see that such rates are given by the Perron-Frobenius eigenvalue of a suitably chosen irreducible, nonnegative matrix.

We start by recalling the classical Perron-Frobenius theorem. Let S be a finite set ($S = \{1, \ldots, n\}$ in the traditional formulation of the Perron-Frobenius theorem) and let $A : S \times S \to \mathbb{R}$ be a function. We view such functions a matrices, equipped with the usual matrix product, or equivalently we identify A with the linear operator $A : \mathbb{R}^S \to \mathbb{R}^S$ given by $Af(x) := \sum_{y\in S} A(x,y)f(y)$. We say that A is *nonnegative* if $A(x,y) \geq 0$ for all $x, y \in S$. A nonnegative matrix A is called *irreducible* if for each $x, y \in S$ there exists an $n \geq 1$ such that $A^n(x,y) > 0$. Note that for probability kernels, this coincides with our earlier definition of irreducibility. We let $\sigma(A)$ denote the *spectrum* of A, i.e., the collection of (possibly complex) eigenvalues of A, and we let $\rho(A)$ denote its *spectral radius*

$$\rho(A) := \sup\{|\lambda| : \lambda \in \sigma(A)\}.$$

If $\|\cdot\|$ is any norm on $\mathbb{R}^S$, then we define the associated *operator norm* $\|A\|$ of A as

$$\|A\| := \sup\{\|Af\| : f \in \mathbb{R}^S,\ \|f\| = 1\}.$$

It is well-known that for any such operator norm

$$\rho(A) = \lim_{n\to\infty} \|A^n\|^{1/n}. \tag{4.15}$$

We cite the following version of the Perron-Frobenius theorem from [Gan00, Section 8.3] (see also, e.g., [Sen73, Chapter 1]).

Theorem 4.22 (Perron-Frobenius) *Let S be a finite set and let $A : \mathbb{R}^S \to \mathbb{R}^S$ be a linear operator whose matrix is nonnegative and irreducible. Then*

(i) *There exist an $f : S \to \mathbb{R}$, unique up to multiplication by positive constants, and a unique $\alpha \in \mathbb{R}$ such that $Af = \alpha f$ and $f(x) \geq 0$.*

(ii) *$f(x) > 0$ for all $x \in S$.*

(iii) *$\alpha = \rho(A) > 0$.*

(iv) *The algebraic multiplicity of α is one. In particular, if A is written in its Jordan normal form, then α corresponds to a block of size 1×1.*

Remark If A is moreover aperiodic, then there exists some $n \geq 1$ such that $A^n(x, y) > 0$ for all $x, y \in S$. Now Perron's theorem [Gan00, Section 8.2] implies that all other eigenvalues λ of A satisfy $|\lambda| < \alpha$. If A is not aperiodic, then it is easy to see that this statement fails in general. (This is stated incorrectly in [DZ98, Thm 3.1.1 (b)].)

We call the constant α and function f from Theorem 4.22 the *Perron-Frobenius eigenvalue* and *eigenfunction* of A, respectively. We note that if $A^\dagger(x, y) := A(y, x)$ denotes the *transpose* of A, then $A^\dagger$ is also nonnegative and irreducible. It is well-known that the spectra of a matrix and its transpose agree: $\sigma(A) = \sigma(A^\dagger)$, and therefore also $\rho(A) = \rho(A^\dagger)$, which implies that the Perron-Frobenius eigenvalues of A and $A^\dagger$ are the same. The same is usually not true for the corresponding Perron-Frobenius eigenvectors. We call eigenvectors of A and $A^\dagger$ also *right* and *left eigenvectors*, respectively.

The main aim of the present section is to prove the following result.

Theorem 4.23 (Exponential rate as eigenvalue) *Let $X = (X_k)_{k\geq 0}$ be a Markov chain with finite state space S, irreducible transition kernel P, and arbitrary initial law. Let $\phi : S^2 \to [-\infty, \infty)$ be a function such that*

$$U := \{(x, y) \in S^2 : \phi(x, y) > -\infty\} \subset \{(x, y) \in S^2 : P(x, y) > 0\}$$

is irreducible, and let $\overline{U}$ be as in (4.3). Then, provided that $X_0 \in \overline{U}$ a.s., one has

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\big[e^{\sum_{k=1}^n \phi(X_{k-1},X_k)}\big] = r,$$

where e^r is the Perron-Frobenius eigenvalue of the nonnegative, irreducible matrix A defined by

$$A(x,y) := P(x,y)\,e^{\phi(x,y)} \qquad (x,y \in \overline{U}). \tag{4.16}$$

We start with some preparatory lemmas. The next lemma shows that there is a close connection between Perron-Frobenius theory and Markov chains.

Lemma 4.24 (Perron-Frobenius Markov chain) *Let S be a finite set and let $A : \mathbb{R}^S \to \mathbb{R}^S$ be a linear operator whose matrix is nonnegative and irreducible. Let α, η and h be its associated Perron-Frobenius eigenvalue and left and right eigenvectors, respectively, i.e., $\eta A = \alpha\eta$, $Ah = \alpha h$, $\eta, h > 0$. Choose any normalization such that $\sum_x h(x)\eta(x) = 1$. Then the matrix*

$$A_h(x,y) := \frac{A(x,y)h(y)}{\alpha h(x)} \qquad (x,y \in S) \tag{4.17}$$

is an irreducible probability kernel on S and $h\eta$ is its unique invariant law.

Proof Recall from Theorem 4.22 that h is strictly positive, hence A_h is well-defined. Since

$$\sum_{y\in S} A_h(x,y) = \sum_{y\in S} \frac{A(x,y)h(y)}{\alpha h(x)} = \frac{\alpha h(x)}{\alpha h(x)} = 1 \quad (x \in S),$$

we see that A_h is a probability kernel. Since $A_h(x,y) > 0$ if and only if $A(x,y) > 0$, the kernel A_h is irreducible. Since

$$\begin{aligned}\sum_{x\in S} h(x)\eta(x)A_h(x,y) &= \sum_{x\in S} h(x)\eta(x)\frac{A(x,y)h(y)}{\alpha h(x)}\\ &= \alpha^{-1}\sum_{x\in S}\eta(x)A(x,y)h(y) = \eta(y)h(y),\end{aligned}$$

we see that $h\eta$ is an invariant law for A_h, and the only such invariant law by the irreducibility of the latter. ∎

The following lemma is not only the key to proving Theorem 4.23, it also provides a link between Perron-Frobenius eigenvectors and entropy. In particular, in some special cases (such as Excercise 4.27), the following lemma can actually be used to obtain Perron-Frobenius eigenvectors by minimizing a suitable functional.

Lemma 4.25 (Minimizer of weighted entropy) *Let S be a finite set, let P be a probability kernel on S and let $\phi : S^2 \to [-\infty, \infty)$ be a function such that*

$$U := \{(x, y) \in S^2 : \phi(x, y) > -\infty\} \subset \{(x, y) \in S^2 : P(x, y) > 0\}$$

is irreducible. Let $\overline{U}$ be as in (4.3), define A as in (4.16), let $\alpha = e^r$ be its Perron-Frobenius eigenvalue and let $\eta, h > 0$ be the associated left and right eigenvectors, normalized such that $\sum_{x\in\overline{U}} h(x)\eta(x) = 1$. Let A_h be the probability kernel defined in (4.17) and let $\pi := h\eta$ be its unique invariant law. Let $\mathcal{V} := \{\nu \in \mathcal{M}_1(S^2) : \nu^1 = \nu^2\}$. Then the function

$$G_\phi(\nu) := \nu\phi - H(\nu|\nu^1 \otimes P)$$

satisfies $G_\phi(\nu) \leq r$ $(\nu \in \mathcal{V})$, with equality if and only if $\nu = \pi \otimes A_h$.

Proof We have $G_\phi(\nu) = -\infty$ if $\nu(x_1, x_2) > 0$ for some $(x_1, x_2) \notin U$. On the other hand, for $\nu \in \mathcal{V}(U)$, we observe that

$$\begin{aligned}
&\nu\phi - H(\nu|\nu^1 \otimes P) \\
&\quad = \sum_{(x_1,x_2)\in U} \nu(x_1, x_2)\phi(x_1, x_2) - \sum_{(x_1,x_2)\in U} \nu(x_1, x_2) \log \frac{\nu(x_1, x_2)}{\nu^1(x_1)P(x_1, x_2)} \\
&\quad = \sum_{(x_1,x_2)\in U} \nu(x_1, x_2)\Big(\phi(x_1, x_2) - \log \nu(x_1, x_2) + \log \nu^1(x_1) + \log P(x_1, x_2)\Big) \\
&\quad = \sum_{(x_1,x_2)\in U} \nu(x_1, x_2)\Big(- \log \nu(x_1, x_2) + \log \nu^1(x_1) + \log A(x_1, x_2)\Big) \\
&\quad = \sum_{(x_1,x_2)\in U} \nu(x_1, x_2)\Big(- \log \nu(x_1, x_2) + \log \nu^1(x_1) + \log A_h(x_1, x_2) \\
&\qquad\qquad\qquad + \log \alpha + \log h(x_1) - \log h(x_2)\Big) \\
&\quad = \log \alpha - H(\nu|\nu^1 \otimes A_h),
\end{aligned}$$

where in the last step we have used that $\nu^1 = \nu^2$. Now the statement follows from Excercise 4.11. ∎

Proof of Theorem 4.23 We will deduce the claim from our basic large deviations results for Markov chains (Theorem 4.2 and Corollary 4.9). A direct proof (using a bit of matrix theory) is also possible, but our aim is to exhibit the links with our earlier results. In fact, the calculations below can be reversed, i.e., a direct proof of Theorem 4.23 can be used as the basis for an alternative proof of Theorem 4.2; see [Hol00, Section V.4].

Let $M_n^{(2)}$ be the pair empirical distributions associated with X, defined in (4.1). Let

$$\mathcal{M}_1(U) \ni \nu \mapsto F(\nu) \in [-\infty, \infty)$$

be the continuous and bounded from above. Then, by Varadhan's lemma (Lemma 1.12) and Corollary 4.9,

$$\lim_{n\to\infty} \frac{1}{n} \log \int \mathbb{P}[M_n^{(2)} \in \mathrm{d}\nu]\big|_{\mathcal{M}_1(U)} e^{nF(\nu)} = \sup_{\nu\in\mathcal{M}_1(U)} \big[F(\nu) - \tilde{I}^{(2)}(\nu)\big],$$

where $\tilde{I}^{(2)}$ is the rate function from Corollary 4.9. A simpler way of writing this formula is

$$\lim_{n\to\infty} \frac{1}{n} \log \int \mathbb{E}\big[e^{nF(M_n^{(2)})}\big] = \sup_{\nu\in\mathcal{M}_1(S^2)} \big[F(\nu) - I^{(2)}(\nu)\big], \tag{4.18}$$

where $I^{(2)}$ is the rate function from Theorem 4.2 and we have extended F to a function on $\mathcal{M}_1(S^2)$ by setting $F(\nu) := -\infty$ if $\nu \in \mathcal{M}_1(S^2)\backslash\mathcal{M}_1(U)$.

Applying this to the 'linear' function F defined by

$$F(\nu) := \nu\phi = \sum_{x\in S} \nu(x)\phi(x) \qquad \big(\nu \in \mathcal{M}_1(S^2)\big),$$

formula (4.18) tells us that

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\big[e^{\sum_{k=1}^n \phi(X_{k-1},X_k)}\big] &= \sup_{\nu\in\mathcal{M}_1(S^2)} \big[\nu\phi - I^{(2)}(\nu)\big] \\ &= \sup_{\nu\in\mathcal{V}} \big[\nu\phi - I^{(2)}(\nu)\big] = r,\end{aligned}$$

where we have used that $I^{(2)}(\nu) = H(\nu|\nu^1 \otimes P)$ for $\nu \in \mathcal{V}$ and $I^{(2)}(\nu) = \infty$ otherwise, and the final equality follows from Lemma 4.25. ∎

Exercise 4.26 (First occurrence of a pattern: part 2) Let $(X_k)_{k\geq 0}$ be i.i.d. random variables with $\mathbb{P}[X_k = 0] = \mathbb{P}[X_k = 1] = \frac{1}{2}$. Let λ_{001} be defined as in Excercise 4.21 and let

$$\lambda_{00} := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[(X_k, X_{k+1}) \neq (0,0)\ \forall k = 1,\dots,n\big]$$

Prove that $\lambda_{001} = \lambda_{00}$.

Exercise 4.27 (First occurrence of a pattern: part 3) Consider a Markov chain $Z = (Z_k)_{k\geq 0}$ taking values in the space

$$S := \{\underline{1}, \underline{1}0, 1\underline{0}, \underline{1}00, 1\underline{0}0, 10\underline{0}, \dagger\},$$

that evolves according to the following rules:

$$\left.\begin{array}{c} \underline{1}0 \mapsto 1\underline{0} \\ \underline{1}00 \mapsto 1\underline{0}0 \mapsto 10\underline{0} \end{array}\right\} \text{ with probability one,}$$

and

$$\left.\begin{array}{r} \underline{1} \\ 1\underline{0} \\ 10\underline{0} \end{array}\right\} \mapsto \left\{\begin{array}{ll} \underline{1} & \text{with probability } 2^{-1}, \\ \underline{1}0 & \text{with probability } 2^{-2}, \\ \underline{1}00 & \text{with probability } 2^{-3}, \\ \dagger & \text{with probability } 2^{-3}, \end{array}\right.$$

i.e., from each of the states $\underline{1}, 1\underline{0}, 10\underline{0}$, we jump with probability $\frac{1}{2}$ to $\underline{1}$, with probability $\frac{1}{4}$ to $\underline{1}0$, with probability $\frac{1}{8}$ to $\underline{1}00$, and with probability $\frac{1}{8}$ to $\dagger$. The state $\dagger$, finally, is a trap:

$$\dagger \mapsto \dagger \quad \text{with probability one.}$$

Define $\phi : S \times S \to [-\infty, \infty)$ by

$$\phi(x, y) := \left\{\begin{array}{ll} 0 & \text{if } P(x, y) > 0 \text{ and } y \neq \dagger, \\ -\infty & \text{otherwise.} \end{array}\right.$$

Let θ be the unique solution in the interval $[0, 1]$ of the equation

$$\theta + \theta^2 + \theta^3 = 1,$$

and let $\tilde{Z} = (\tilde{Z}_k)_{k\geq 0}$ be a Markov chain with state space $S\backslash\{\dagger\}$ that evolves in the same way as Z, except that

$$\left.\begin{array}{r} \underline{1} \\ 1\underline{0} \\ 10\underline{0} \end{array}\right\} \mapsto \left\{\begin{array}{ll} \underline{1} & \text{with probability } \theta, \\ \underline{1}0 & \text{with probability } \theta^2, \\ \underline{1}00 & \text{with probability } \theta^3. \end{array}\right.$$

Let P and Q be the transition kernels of Z and $\tilde{Z}$, respectively. Set $U := \{(x, y) \in S^2 : \phi(x, y) > -\infty\}$. Prove that for any $\nu \in \mathcal{V}(U)$

$$\nu\phi - H(\nu|\nu^1 \otimes P) = \log(\tfrac{1}{2}) - \log\theta - H(\nu|\nu^1 \otimes Q). \tag{4.19}$$

Hint: Do a calculation as in the proof of Lemma 4.25, and observe that for any $\nu \in \mathcal{V}(U)$

$$\nu^1(\underline{1}1) = \nu^1(1\underline{1}) \quad \text{and} \quad \nu^1(\underline{1}11) = \nu^1(1\underline{1}1) = \nu^1(11\underline{1}),$$

hence $\nu^1(\underline{1}) + 2\nu^1(\underline{1}1) + 3\nu^1(\underline{1}11) = 1$.

Exercise 4.28 (First occurrence of a pattern: part 4) Let $(X_k)_{k\geq 0}$ be i.i.d. random variables with $\mathbb{P}[X_k = 0] = \mathbb{P}[X_k = 1] = \frac{1}{2}$ and let λ_{000} be defined as in Excercise 4.21. Prove that $\lambda_{000} = \log(\frac{1}{2}) - \log(\theta)$, where θ is the unique root of the equation $\theta + \theta^2 + \theta^3 = 1$ in the interval $[0, 1]$. Hint: use formula (4.19).

Exercise 4.29 (Percolation on a ladder) Consider an infinite graph with vertex set $\{1, 2\} \times \mathbb{N}$. For each $k \geq 1$, draw an edges between $(1, k-1)$ and $(1, k)$, between $(2, k-1)$ and $(2, k)$, and between $(1, k)$ and $(2, k)$. Let $0 < p < 1$. If edges are independently open with probability p and closed otherwise, then how far can we walk into this graph along open edges? To investigate this question, let $(Y_{i,k})_{i=1,2,3,\ k\geq 1}$ be i.i.d. Bernoulli random variables with $\mathbb{P}[Y_{i,k} = 1] = p$ and $\mathbb{P}[Y_{i,k} = 0] = 1 - p$. Define inductively a Markov chain $(X_k)_{k\geq 0}$ with state space $\{0, 1\}^2$ and initial state $(X_0(1), X_0(2)) = (1, 1)$ by first setting, for each $k \geq 1$,

$$\tilde{X}_k(1) := Y_{k,1}X_{k-1}(1) \quad \text{and} \quad \tilde{X}_k(2) := Y_{k,2}X_{k-1}(2),$$

and then

$$X_k(1) := \tilde{X}_k(1) \vee Y_{3,k}\tilde{X}_k(2) \quad \text{and} \quad X_k(2) := \tilde{X}_k(2) \vee Y_{3,k}\tilde{X}_k(1).$$

Calculate the limit

$$r := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big[X_n \neq (0, 0)\big].$$

Hint: find the transition kernel of X and calculate the relevant Perron-Frobenius eigenvalue. You can reduce the dimensionality of the problem by exploiting the symmetry between $(1, 0)$ and $(0, 1)$. Don't worry if the formula for r looks somewhat complicated.

4.5 Continuous time

Recall from Section 0.4 the definition of a continuous-time Markov process $X = (X_t)_{t\geq 0}$ with finite state space S, initial law μ, transition probabilities $P_t(x, y)$, semigroup $(P_t)_{t\geq 0}$, generator G, and transition rates $r(x, y)$ $(x \neq y)$. To simplify notation, we set $r(x, x) := 0$.

By definition, an *invariant law* is a probability measure ρ on S such that $\rho P_t = \rho$ for all $t \geq 0$, or, equivalently, $\rho G = 0$. This latter formula can be written more explicitly in terms of the rates $r(x, y)$ as

$$\sum_{y\in S} \rho(y)r(y, x) = \rho(x) \sum_{y\in S} r(x, y) \qquad (x \in S),$$

i.e., in equilibrium, the frequency of jumps to x equals the frequency of jumps from x. Basic results about Markov processes with finite state spaces tell us that if the transition rates $r(x,y)$ are irreducible, then the corresponding Markov process has a unique invariant law ρ, and $\mu P_t \Rightarrow \rho$ as $t \to \infty$ for every initial law μ. (For continuous-time processes, there is no such concept as (a)periodicity.)

We let

$$M_T(x) := \frac{1}{T}\int_0^T 1_{\{X_t = x\}}\mathrm{d}t \qquad (T > 0)$$

denote the *empirical distribution* of X up to time T. We denote the set of times when X makes a jump up to time T by

$$\Delta_T := \{t \in (0,T] : X_{t-} \neq X_t\}$$

and we set

$$W_T(x,y) := \frac{1}{T}\sum_{t\in\Delta_T} 1_{\{X_{t-} = x,\ X_t = y\}} \qquad (T > 0),$$

i.e., $W_T(x,y)$ is the *empirical frequency* of jumps from x to y. If the transition rates $r(x,y)$ are irreducible, then, for large T, we expect M_T to be close to the (unique) invariant law ρ of X and we expect $W_T(x,y)$ to be close to $\rho(x)r(x,y)$. We observe that (M_T, W_T) is a random variable taking values in the space $\mathcal{M}_1(S) \times [0,\infty)^{S^2}$. For any $w \in [0,\infty)^{S^2}$, we let

$$w^1(x_1) := \sum_{x_2\in S} w(x_1,x_2) \quad \text{and} \quad w^2(x_2) := \sum_{x_1\in S} w(x_1,x_2)$$

denote the first and second marginal of w, and we set

$$\mathcal{W} := \big\{(\rho,w) : \rho \in \mathcal{M}_1(S),\ w \in [0,\infty)^{S^2},\ w^1 = w^2,$$
$$w(x,y) = 0 \text{ whenever } \rho(x)r(x,y) = 0\big\}.$$

The aim of the present section is to prove the following analogue of Theorem 4.2. Note that the function ψ below satisfies $\psi'(z) = \log z$ and $\psi''(z) = 1/z$, is strictly convex and assumes its minimum in the point $z = 1$ where $\psi(1) = 0$.

Theorem 4.30 (LDP for Markov processes) *Let $(X_t)_{t\geq 0}$ be a continuous-time Markov process with finite state space S, irreducible transition rates $r(x,y)$, and arbitrary initial law. Let M_T and W_T $(T > 0)$ denote its empirical distributions and empirical frequencies of jumps, respectively, as defined above. Then the laws*

$\mathbb{P}[(M_T, W_T) \in \cdot\,]$ *satisfy the large deviation principle on* $\mathcal{M}_1(S) \times [0,\infty)^{S^2}$ *with speed* T *and good rate function* I *given by*

$$I(\rho, w) := \begin{cases} \displaystyle\sum_{x,y\in S} \rho(x)r(x,y)\psi\Big(\frac{w(x,y)}{\rho(x)r(x,y)}\Big) & \text{if } (\rho, w) \in \mathcal{W}, \\ \infty & \text{otherwise,} \end{cases}$$

where $\psi(z) := 1 - z + z\log z$ $(z > 0)$ *and* $\psi(0) := 1$ *and we set* $0\,\psi(a/b) := 0$, *regardless of the values of* $a, b \geq 0$.

Remark So far, we have only considered large deviation principles for *sequences* of measures μ_n. The theory for families of measures $(\mu_T)_{T>0}$ depending on a continuous parameter is completely analogous. Indeed, if the μ_T are finite measures on a Polish space E and I is a good rate function, then one has

$$\lim_{T\to\infty} \|f\|_{T,\mu_T} = \|f\|_{\infty,I} \qquad \big(f \in \mathcal{C}_{b,+}(E)\big)$$

if and only if for each $T_n \to \infty$,

$$\lim_{n\to\infty} \|f\|_{T_n,\mu_{T_n}} = \|f\|_{\infty,I} \qquad \big(f \in \mathcal{C}_{b,+}(E)\big).$$

A similar statement holds for the two conditions in Proposition 1.7. In other words: measures μ_T depending on a continuous parameter $T > 0$ satisfy a large deviation principle with speed T and good rate function I if and only if for each $T_n \to \infty$, the measures μ_{T_n} satisfy the large deviation principle with speed T_n and rate function I.

Exercise 4.31 (Properties of the rate function) Show that the function I from Theorem 4.30 is a good rate function and that $I(\rho, w) \geq 0$ with equality if and only if ρ is the unique invariant law of the Markov process X and $w(x,y) = \rho(x)r(x,y)$ $(x, y \in S)$.

Our strategy is to derive Theorem 4.30 from Theorem 4.2 using approximation. We start with an abstract lemma.

Lemma 4.32 (Diagonal argument) *Let* $(\mu_{m,n})_{m,n\geq 1}$ *be finite measures on a Polish space* E, *let* s_n *be positive constants, tending to infinity, and let* I_m, I *be good rate functions on* E. *Assume that for each fixed* $m \geq 1$, *the* $\mu_{m,n}$ *satisfy the*

large deviation principle with speed s_n and rate function I_m. Assume moreover that

$$\lim_{m\to\infty} \|f\|_{\infty,I_m} = \|f\|_{\infty,I} \qquad \big(f \in \mathcal{C}_{b,+}(E)\big).$$

Then there exist $n(m) \to \infty$ such that for all $n'(m) \geq n(m)$, the measures $\mu_{m,n'(m)}$ satisfy the large deviation principle with speed $s_{n'(m)}$ and rate function I.

Proof Let $\overline{E}$ be a metrizable compactification of E. We view the $\mu_{m,n}$ as measures on $\overline{E}$ such that $\mu_{m,n}(\overline{E}\backslash E) = 0$ and we extend the rate fuctions I_m, I to $\overline{E}$ by setting $I_m, I := \infty$ on $\overline{E}\backslash E$. Then

$$\lim_{m\to\infty} \|f\|_{\infty,I_m} = \|f\|_{\infty,I} \qquad \big(f \in \mathcal{C}(\overline{E})\big).$$

Let $\{f_i : i \geq 1\}$ be a countable dense subset of the separable Banach space $\mathcal{C}(\overline{E})$ of continuous real functions on E, equipped with the supremumnorm. Choose $n(m) \to \infty$ such that

$$\big|\|f_i\|_{s_{n'},\mu_{m,n'}} - \|f_i\|_{\infty,I_m}\big| \leq 1/m \qquad \big(n' \geq n(m),\ i \leq m\big).$$

Then, for any $n'(m) \geq n(m)$, one has

$$\begin{aligned}
&\limsup_{m\to\infty} \big|\|f_i\|_{s_{n'(m)},\mu_{m,n'(m)}} - \|f_i\|_{\infty,I}\big| \\
&\quad \leq \limsup_{m\to\infty} \big|\|f_i\|_{s_{n'(m)},\mu_{m,n'(m)}} - \|f_i\|_{\infty,I_m}\big| + \limsup_{m\to\infty} \big|\|f_i\|_{\infty,I_m} - \|f_i\|_{\infty,I}\big| = 0
\end{aligned}$$

for all $i \geq 1$. By Lemma 1.33 (b), the functions $|f_i|$ are rate function determining, hence by Proposition 1.32, the measures $\mu_{m,n'(m)}$ satisfy the large deviation principle on $\overline{E}$ with speed $s_{n'(m)}$ and rate function I. By the restriction principle (Lemma 1.28), they also satisfy the large deviation principle on E. ∎

Proposition 4.33 (Approximation of LDP's) *Let E be a Polish space and let $X_n, X_{m,n}$ $(m, n \geq 1)$ be random variables taking values in E. Assume that for each fixed $m \geq 1$, the laws $\mathbb{P}[X_{m,n} \in \cdot\,]$ satisfy a large deviation principle with speed s_n and good rate function I_m. Assume moreover that there exists a good rate function I such that*

$$\lim_{m\to\infty} \|f\|_{\infty,I_m} = \|f\|_{\infty,I} \qquad \big(f \in \mathcal{C}_{b,+}(E)\big), \tag{4.20}$$

and that there exists a metric d generating the topology on E such that for each $n(m) \to \infty$,

$$\lim_{m\to\infty} \frac{1}{s_{n(m)}} \log \mathbb{P}[d(X_{n(m)}, X_{m,n(m)}) \geq \varepsilon] = -\infty \qquad (\varepsilon > 0), \tag{4.21}$$

i.e., $X_{n(m)}$ and $X_{m,n(m)}$ are exponentially close in the sense of (1.8). Then the laws $\mathbb{P}[X_n \in \cdot\,]$ satisfy the large deviation principle with speed s_n and good rate function I.

Proof By the argument used in the proof of Proposition 1.32, it suffices to show that each subsequence $n(m) \to \infty$ contains a further subsequence $n'(m) \to \infty$ such that the laws $\mathbb{P}[X_{n'(m)} \in \cdot\,]$ satisfy the large deviation principle with speed $s_{n'(m)}$ and good rate function I. By (4.20) and Lemma 4.32, we can choose $n'(m) \to \infty$ such that the laws $\mathbb{P}[X_{m,n'(m)} \in \cdot\,]$ satisfy the large deviation principle with speed $s_{n'(m)}$ and good rate function I. By (4.21), the random variables $X_{n'(m)}$ and $X_{m,n'(m)}$ are exponentially close in the sense of Proposition 1.17, hence the large deviation principle for the laws of the $X_{m,n'(m)}$ implies the large deviation principle for the laws of the $X_{n'(m)}$. ∎

The following lemma gives sufficient conditions for the type of convergence in (4.20).

Lemma 4.34 (Convergence of rate functions) *Let E be a Polish space and let I, I_m be good rate functions on E such that*

(i) *For each $a \in \mathbb{R}$, there exists a compact set $K \subset E$ such that $\{x \in E : I_m(x) \leq a\} \subset K$ for all $m \geq 1$.*

(ii) *$\forall x_m, x \in E$ with $x_m \to x$, one has $\liminf_{m\to\infty} I_m(x_m) \geq I(x)$.*

(iii) *$\forall x \in E\ \exists x_m \in E$ such that $x_m \to x$ and $\limsup_{m\to\infty} I_m(x_m) \leq I(x)$.*

Then the I_m converge to I in the sense of (4.20).

Proof Formula (4.20) is equivalent to the statement that

$$\inf_{x\in E}[I_m(x) - F(x)] \underset{m\to\infty}{\longrightarrow} \inf_{x\in E}[I(x) - F(x)]$$

for any continuous $F : E \to [-\infty, \infty)$ that is bounded from above. If I_m, I satisfy conditions (i)–(iii), then the same is true for $I' := I - F$, $I'_m := I_m - F$, so it suffices to show that conditions (i)–(iii) imply that

$$\inf_{x\in E} I_m(x) \underset{m\to\infty}{\longrightarrow} \inf_{x\in E} I(x).$$

Since I is a good rate function, it achieves its minimum, i.e., there exists some $x_\circ \in E$ such that $I(x_\circ) = \inf_{x\in E} I(x)$. By condition (iii), there exist $x_m \in E$ such that $x_m \to x$ and

$$\limsup_{m\to\infty} \inf_{x\in E} I_m(x) \leq \limsup_{m\to\infty} I_m(x_m) \leq I(x_\circ) = \inf_{x\in E} I(x).$$

To prove the other inequality, assume that

$$\liminf_{m\to\infty} \inf_{x\in E} I_m(x) < \inf_{x\in E} I(x).$$

Then, by going to a subsequence if necessary, we can find $x_m \in E$ such that

$$\lim_{m\to\infty} I_m(x_m) < \inf_{x\in E} I(x),$$

where the limit on the left-hand side exists and may be $-\infty$. By condition (i), there exists a compact set $K \subset E$ such that $x_m \in K$ for all m, hence by going to a further subsequence if necessary, we may assume that $x_m \to x_*$ for some $x_* \in E$. Condition (ii) now tells us that

$$\lim_{m\to\infty} I_m(x_m) \geq I(x_*) \geq \inf_{x\in E} I(x),$$

which leads to a contradiction. ∎

Proof of Theorem 4.30 We set

$$\begin{aligned} M_T^\varepsilon(x) &:= \frac{1}{\lfloor T/\varepsilon\rfloor} \sum_{k=1}^{\lfloor T/\varepsilon\rfloor} 1_{\{(X_{\varepsilon(k-1)}, X_{\varepsilon k}) = (x,x)\}} \qquad (x \in S), \\ W_T^\varepsilon(x,y) &:= \frac{1}{\varepsilon\lfloor T/\varepsilon\rfloor} \sum_{k=1}^{\lfloor T/\varepsilon\rfloor} 1_{\{(X_{\varepsilon(k-1)}, X_{\varepsilon k}) = (x,y)\}} \qquad (x,y \in S,\ x \neq y), \end{aligned}$$

and we let $W_T^\varepsilon(x,x) := 0$ $(x \in S)$. By Proposition 4.33, the statements of the theorem will follow provided we prove the following three claims:

1. For each $\varepsilon > 0$, the laws $\mathbb{P}[(M_T^\varepsilon, W_T^\varepsilon) \in \cdot\,]$ satisfy a large deviation principle with speed T and good rate function I_ε.

2. The function I from Theorem 4.30 is a good rate function and the rate functions I_ε converge to I in the sense of (4.20) as $\varepsilon \downarrow 0$.

3. For each $T_m \to \infty$ and $\varepsilon_m \downarrow 0$, the random variables $(M_{T_m}^{\varepsilon_m}, W_{T_m}^{\varepsilon_m})$ and (M_{T_m}, W_{T_m}) are exponentially close with speed T_m.

Proof of Claim 1. For each $\varepsilon > 0$, let $(X_k^\varepsilon)_{k\geq 0}$ be the Markov chain given by

$$X_k^\varepsilon := X_{\varepsilon k} \qquad (k \geq 0),$$

and let $M_n^{(2)\,\varepsilon}$ be its empirical pair distributions. Then

$$\begin{aligned} M_T^\varepsilon(x) &= M_{\lfloor T/\varepsilon\rfloor}^{(2)\,\varepsilon}(x,x) && (x \in S),\\ W_T^\varepsilon(x,y) &= \varepsilon^{-1} M_{\lfloor T/\varepsilon\rfloor}^{(2)\,\varepsilon}(x,y) && (x,y \in S,\ x \neq y). \end{aligned}$$

For each $\varepsilon > 0$ and $\nu \in \mathcal{M}_1(S^2)$, let us define $\rho_\varepsilon \in [0,\infty)^S$ and $w_\varepsilon(\nu) \in [0,\infty)^{S^2}$ by

$$\begin{aligned} \rho_\varepsilon(\nu)(x) &:= \nu(x,x) && (x \in S),\\ w_\varepsilon(\nu)(x) &:= 1_{\{x\neq y\}}\varepsilon^{-1}\nu(x,y) && (x,y \in S). \end{aligned}$$

Then, by Theorem 4.2, for each $\varepsilon > 0$ the laws $\mathbb{P}[(M_T^\varepsilon, W_T^\varepsilon) \in \cdot\,]$ satisfy a large deviation principle on $[0,\infty)^S \times [0,\infty)^{S^2}$ with speed T and good rate function I_ε given by

$$I_\varepsilon(\rho_\varepsilon(\nu), w_\varepsilon(\nu)) := \varepsilon^{-1} H(\nu|\nu^1 \otimes P_\varepsilon) \qquad (\nu \in \mathcal{V}), \tag{4.22}$$

while $I_\varepsilon(\rho, w) := \infty$ if there exists no $\nu \in \mathcal{V}$ such that $(\rho, w) = (\rho_\varepsilon(\nu), w_\varepsilon(\nu))$. Note the overall factor ε^{-1} which is due to the fact that the speed T differs a factor ε^{-1} from the speed n of the embedded Markov chain.

Proof of Claim 2. By Lemma 4.34, it suffices to prove, for any $\varepsilon_n \downarrow 0$, the following three statements.

(i) If $\rho_n \in [0,\infty)^S$ and $w_n \in [0,\infty)^{S^2}$ satisfy $w_n(x,y) \to \infty$ for some $x, y \in S$, then $I_{\varepsilon_n}(\rho_n, w_n) \to \infty$.

(ii) If $\rho_n \in [0,\infty)^S$ and $w_n \in [0,\infty)^{S^2}$ satisfy $(\rho_n, w_n) \to (\rho, w)$ for some $\rho \in [0,\infty)^S$ and $w \in [0,\infty)^{S^2}$, then $\liminf_{n\to\infty} I_{\varepsilon_n}(\rho_n, w_n) \geq I(\rho, w)$.

(iii) For each $\rho \in [0,\infty)^S$ and $w \in [0,\infty)^{S^2}$ there exist $\rho_n \in [0,\infty)^S$ and $w_n \in [0,\infty)^{S^2}$ such that $\limsup_{n\to\infty} I_{\varepsilon_n}(\rho_n, w_n) \leq I(\rho, w)$.

Obviously, it suffices to check conditions (i), (ii) for (ρ_n, w_n) such that $I_{\varepsilon_n}(\rho_n, w_n) < \infty$ and condition (iii) for (ρ, w) such that $I(\rho, w) < \infty$. Therefore, taking into account our definition of I_ε, Claim 2 will follow provided we prove the following three subclaims.

2.I. If $\nu_n \in \mathcal{V}$ satisfy $\varepsilon_n^{-1}\nu_n(x,y) \to \infty$ for some $x \neq y$, then

$$\varepsilon_n^{-1} H(\nu_n | \nu_n^1 \otimes P_{\varepsilon_n}) \underset{n\to\infty}{\longrightarrow} \infty.$$

2.II. If $\nu_n \in \mathcal{V}$ satisfy

$$\begin{aligned} \nu_n(x,x) &\underset{n\to\infty}{\longrightarrow} \rho(x) && (x \in S), \\ \varepsilon_n^{-1} 1_{\{x\neq y\}} \nu_n(x,y) &\underset{n\to\infty}{\longrightarrow} w(x,y) && (x,y \in S^2), \end{aligned} \tag{4.23}$$

for some $(\rho, w) \in [0,\infty)^S \times [0,\infty)^{S^2}$, then

$$\liminf_{n\to\infty} \varepsilon_n^{-1} H(\nu_n | \nu_n^1 \otimes P_{\varepsilon_n}) \geq I(\rho, w).$$

2.III. For each $(\rho, w) \in \mathcal{W}$, we can find $\nu_n \in \mathcal{V}$ satisfying (4.23) such that

$$\lim_{n\to\infty} \varepsilon_n^{-1} H(\nu_n | \nu_n^1 \otimes P_{\varepsilon_n}) = I(\rho, w).$$

We start by writing $H(\nu|\nu^1 \otimes P)$ in a suitable way. Let ψ be as defined in the theorem. We observe that if ν, μ are probability measures on a finite set S and $\mu(x) > 0$ for all $x \in S$, then

$$\begin{aligned} \sum_{x\in S} \mu(x)\psi\left(\frac{\nu(x)}{\mu(x)}\right) &= \sum_{x\in S} \mu(x)\big[1 - \frac{\nu(x)}{\mu(x)} + \frac{\nu(x)}{\mu(x)} \log\Big(\frac{\nu(x)}{\mu(x)}\Big)\big] \\ &= \sum_{x\in S} [\mu(x) - \nu(x)] + \sum_{x\in S} \nu(x) \log\left(\frac{\nu(x)}{\mu(x)}\right) = H(\nu|\mu), \end{aligned}$$

where we use the convention that $0 \log 0 := 0$. By Excercise 4.10, it follows that for any probability measure ρ on S and probability kernels P, Q on S such that $\rho \otimes Q \ll \rho \otimes P$,

$$\begin{aligned} H(\rho \otimes Q | \rho \otimes P) &= \sum_x \rho(x) H(Q_x | P_x) \\ &= \sum_x \rho(x) \sum_y P(x,y) \psi\Big(\frac{Q(x,y)}{P(x,y)}\Big) = \sum_{x,y} \rho(x) P(x,y) \psi\Big(\frac{\rho(x)Q(x,y)}{\rho(x)P(x,y)}\Big), \end{aligned}$$

where the sum runs over all $x, y \in S$ such that $\rho(x)P(x,y) > 0$. In particular, if ν is a probability measure on S^2 and P is a probability kernel on S, then

$$H(\nu|\nu^1 \otimes P) = \begin{cases} \displaystyle\sum_{x,y\in S} \nu^1(x) P(x,y) \psi\Big(\frac{\nu(x,y)}{\nu^1(x)P(x,y)}\Big) & \text{if } \nu \ll \nu^1 \otimes P, \\ \infty & \text{otherwise,} \end{cases}$$

where we define $0\,\psi(a/b) := 0$, irrespective of the values of $a, b \geq 0$.

To prove Claim 2.I, now, we observe that if $\varepsilon_n^{-1}\nu_n(x,y) \to \infty$ for some $x \neq y$, then

$$\begin{aligned}\varepsilon_n^{-1}H(\nu_n|\nu_n^1 \otimes P_{\varepsilon_n}) &\geq \varepsilon_n^{-1}\nu_n^1(x)P_{\varepsilon_n}(x,y)\psi\Big(\frac{\nu_n(x,y)}{\nu_n^1(x)P_{\varepsilon_n}(x,y)}\Big)\\ &\geq \varepsilon_n^{-1}\nu_n(x,y)\Big(\log\Big(\frac{\nu_n(x,y)}{\nu_n^1(x)P_{\varepsilon_n}(x,y)}\Big)-1\Big),\end{aligned}$$

where

$$\frac{\nu_n(x,y)}{\nu^1(x)P_{\varepsilon_n}(x,y)} \geq \frac{\nu_n(x,y)}{P_{\varepsilon_n}(x,y)} = \frac{\nu_n(x,y)}{\varepsilon_n r(x,y)+O(\varepsilon_n^2)} \underset{n\to\infty}{\longrightarrow} \infty.$$

To prove Claim 2.II, we observe that if ν_n, ρ, w satisfy (4.23), then, as $n \to \infty$,

$$\left.\begin{aligned}\nu_n^1(x)P_{\varepsilon_n}(x,x) &= \rho(x)+O(\varepsilon_n),\\ \nu_n(x,x) &= \rho(x)+O(\varepsilon_n),\end{aligned}\right\} \quad (x \in S),$$

while

$$\left.\begin{aligned}\nu_n^1(x)P_{\varepsilon_n}(x,y) &= \varepsilon_n\rho(x)r(x,y)+O(\varepsilon_n^2),\\ \nu_n(x,y) &= \varepsilon_n w(x,y)+O(\varepsilon_n^2),\end{aligned}\right\} \quad (x,y \in S,\ x \neq y).$$

It follows that

$$\begin{aligned}\varepsilon_n^{-1}H(\nu_n|\nu_n^1 \otimes P_{\varepsilon_n}) &= \varepsilon_n^{-1}\sum_{x,y}\nu_n^1(x)P_{\varepsilon_n}(x,y)\psi\Big(\frac{\nu_n(x,y)}{\nu_n^1(x)P_{\varepsilon_n}(x,y)}\Big)\\ &= \varepsilon_n^{-1}\sum_{x}\big(\rho(x)+O(\varepsilon_n)\big)\psi\Big(\frac{\rho(x)+O(\varepsilon_n)}{\rho(x)+O(\varepsilon_n)}\Big)\\ &\quad+\sum_{x\neq y}\big(\rho(x)r(x,y)+O(\varepsilon_n)\big)\psi\Big(\frac{\varepsilon_n w(x,y)+O(\varepsilon_n^2)}{\varepsilon_n\rho(x)r(x,y)+O(\varepsilon_n^2)}\Big)\\ &\geq \sum_{x\neq y}\rho(x)r(x,y)\psi\Big(\frac{w(x,y)}{\rho(x)r(x,y)}\Big)+O(\varepsilon_n).\end{aligned} \tag{4.24}$$

To prove Claim 2.III, finally, we observe that for each $(\rho, w) \in \mathcal{W}$, we can find $\nu_n \in \mathcal{V}$ satisfying (4.23) such that moreover $\nu_n(x,x) = 0$ whenever $\rho(x) = 0$ and $\nu_n(x,y) = 0$ whenever $\rho(x)r(x,y) = 0$ for some $x \neq y$. It follows that $\nu_n^1(x) = 0$ whenever $\rho(x) = 0$, so for each x, y such that $\rho(x) = 0$, we have

$$\varepsilon_n^{-1}\nu_n^1(x)P_{\varepsilon_n}(x,y)\psi\Big(\frac{\nu_n(x,y)}{\nu_n^1(x)P_{\varepsilon_n}(x,y)}\Big) = 0,$$

while for $x \neq y$ such that $\rho(x) > 0$ but $r(x,y) = 0$, we have

$$\varepsilon_n^{-1}\nu_n^1(x)P_{\varepsilon_n}(x,y)\psi\Big(\frac{\nu_n(x,y)}{\nu_n^1(x)P_{\varepsilon_n}(x,y)}\Big) = O(\varepsilon_n)\psi(1).$$

Note also that if $\rho(x) > 0$, then

$$\psi\Big(\frac{\rho(x)+O(\varepsilon_n)}{\rho(x)+O(\varepsilon_n)}\Big) = \psi\big(1+O(\varepsilon_n)\big) = O(\varepsilon_n^2).$$

It follows that in (4.24), only the terms where $\rho(x)r(x,y) > 0$ contribute, and

$$\varepsilon_n^{-1}H(\nu_n|\nu_n^1 \otimes P_{\varepsilon_n}) = \sum_{x\neq y}\rho(x)r(x,y)\psi\Big(\frac{w(x,y)}{\rho(x)r(x,y)}\Big) + O(\varepsilon_n).$$

Proof of Claim 3. Set $\varepsilon\mathbb{N} := \{\varepsilon k : k \in \mathbb{N}\}$ and observe that $\varepsilon\lfloor T/\varepsilon\rfloor = \sup\{T' \in \varepsilon\mathbb{N} : T' \leq T\}$. It is not hard to show that for any $T_m \to \infty$ and $\varepsilon_m \downarrow 0$, the random variables

$$(M_{T_m}, W_{T_m}) \quad \text{and} \quad (M_{\varepsilon_m\lfloor T_m/\varepsilon_m\rfloor}, W_{\varepsilon_m\lfloor T_m/\varepsilon_m\rfloor}) \tag{4.25}$$

are exponentially close. Therefore, by Excercise 4.37 below and the fact that $(M_{T_m}^{\varepsilon_m}, W_{T_m}^{\varepsilon_m})$ are functions of $\varepsilon_m\lfloor T_m/\varepsilon_m\rfloor$ only, it suffices to prove the statement for times $T_m \in \varepsilon_m\mathbb{N}$.

Recall that $\Delta_T := \{t \in (0,T] : X_{t-} \neq X_t\}$ is the set of times, up to time T, when X makes a jump. For any $T \in \varepsilon\mathbb{N}$, let

$$J_i(\varepsilon,T) := \sum_{k=1}^{T/\varepsilon} 1\big\{\big|\Delta_T \cap (\varepsilon(k-1),\varepsilon k]\big| \geq i\big\} \qquad (i=1,2)$$

denote the number of time intervals of the form $(\varepsilon(k-1),\varepsilon k]$, up to time T, during which X makes at least i jumps. We observe that for any $T \in \varepsilon\mathbb{N}$,

$$\sum_{x\in S}\big|M_T^\varepsilon(x) - M_T(x)\big| \leq \frac{\varepsilon}{T}J_1(\varepsilon,T),$$
$$\sum_{x,y\in S}\big|W_T^\varepsilon(x,y) - W_T(x,y)\big| \leq \frac{1}{T}J_2(\varepsilon,T).$$

Thus, it suffices to show that for any $\delta > 0$, $\varepsilon_m \downarrow 0$ and $T_m \in \varepsilon_m\mathbb{N}$ such that $T_m \to \infty$

$$\lim_{m\to\infty}\frac{1}{T_m}\log\mathbb{P}\big[\varepsilon_m J_1(\varepsilon_m,T_m)/T_m \geq \delta\big] = -\infty,$$
$$\lim_{m\to\infty}\frac{1}{T_m}\log\mathbb{P}\big[J_2(\varepsilon_m,T_m)/T_m \geq \delta\big] = -\infty.$$

We observe that $J_1(\varepsilon, T) \leq |\Delta_T|$, which can in turn be estimated from above by a Poisson distributed random variable N_{RT} with mean

$$T \sup_{x \in S} \sum_{y \in S} r(x, y) =: RT.$$

By Excercise 4.35 below, it follows that for any $0 < \varepsilon < \delta/R$,

$$\begin{aligned} &\limsup_{m \to \infty} \frac{1}{T_m} \log \mathbb{P}\big[\varepsilon_m J_1(\varepsilon_m, T_m)/T_m \geq \delta\big] \\ &\quad \leq \limsup_{m \to \infty} \frac{1}{T_m} \log \mathbb{P}\big[\varepsilon N_{RT_m}/T_m \geq \delta\big] \leq \psi(\delta/R\varepsilon) \underset{\varepsilon \to 0}{\longrightarrow} -\infty, \end{aligned}$$

where $\psi(z) := 1 - z + z \log z$. To also prove the statement for J_2, we observe that Δ_T can be estimated from above by a Poisson point process with intensity R, hence

$$\mathbb{P}\big[\big|\Delta_T \cap (\varepsilon(k-1), \varepsilon k]\big| \geq 2\big] \leq 1 - e^{-R\varepsilon} - R\varepsilon e^{-R\varepsilon}.$$

and $J_2(\varepsilon, T)$ can be estimated from above by a binomially distributed random variable with parameters $(n, p) = (T/\varepsilon, 1 - e^{-R\varepsilon} - R\varepsilon e^{-R\varepsilon})$. For small ε, this binomal distribution approximates a Poisson distribution. To turn this into a rigorous estimate, define λ_ε by

$$1 - e^{-\lambda_\varepsilon} := 1 - e^{-R\varepsilon} - R\varepsilon e^{-R\varepsilon}.$$

In other words, if M and N are Poisson distributed random variables with mean λ_ε and $R\varepsilon$, respectively, then this says that $\mathbb{P}[N \geq 1] = \mathbb{P}[M \geq 2]$. Since the right-hand side of this equation is of order $\frac{1}{2}R^2\varepsilon^2 + O(\varepsilon^3)$ as $\varepsilon \downarrow 0$, we see that

$$\lambda_\varepsilon = \tfrac{1}{2}R^2\varepsilon^2 + O(\varepsilon^3) \qquad \text{as } \varepsilon \downarrow 0.$$

Then $J_2(\varepsilon, T)$ can be estimated from above by a Poisson disributed random variable with mean $(T/\varepsilon)\lambda_\varepsilon = \frac{1}{2}R^2 T\varepsilon + O(\varepsilon^2)$. By the same argument as for J_1, we conclude that

$$\limsup_{m \to \infty} \frac{1}{T_m} \log \mathbb{P}\big[\varepsilon_m J_2(\varepsilon_m, T_m)/T_m \geq \delta\big] = -\infty.$$

∎

Exercise 4.35 (Large deviations for Poisson process) Let $N = (N_t)_{t \geq 0}$ be a Poisson process with intensity one, i.e., N has independent increments where

$N_t - N_s$ is Poisson distributed with mean $t - s$. Show that the laws $\mathbb{P}[N_T/T \in \cdot\,]$ satisfy the large deviation principle with speed T and good rate function

$$I(z) = \begin{cases} 1 - z + z\log z & \text{if } z \geq 0, \\ \infty & \text{otherwise.} \end{cases}$$

Hint: first consider the process at integer times and use that this is a sum of i.i.d. random variables. Then generalize to nontinteger times.

Exercise 4.36 (Rounded times) Prove that the random variables in (4.25) are exponentially close.

Exercise 4.37 (Triangle inequality for exponential closeness) Let $(X_n)_{n\geq 1}$, $(Y_n)_{n\geq 1}$ and $(Z_n)_{n\geq 1}$ be random variables taking values in a Polish space E and let d be a metric generating the topology on E. Let s_n be positive constants, converging to infinity, and assume that

$$\begin{aligned} \lim_{n\to\infty} \tfrac{1}{s_n} \log \mathbb{P}\big[d(X_n, Y_n) \geq \varepsilon\big] &= -\infty \qquad (\varepsilon > 0), \\ \lim_{n\to\infty} \tfrac{1}{s_n} \log \mathbb{P}\big[d(Y_n, Z_n) \geq \varepsilon\big] &= -\infty \qquad (\varepsilon > 0). \end{aligned}$$

Prove that

$$\lim_{n\to\infty} \frac{1}{s_n} \log \mathbb{P}\big[d(X_n, Z_n) \geq \varepsilon\big] = -\infty \qquad (\varepsilon > 0).$$

4.6 Excercises

Exercise 4.38 (Testing the fairness of a dice) Imagine that we want to test if a dice is fair, i.e., if all sides come up with equal probabilities. To test this hypothesis, we throw the dice n times. General statistical theory tells us that any test on the distribution with which each side comes up can be based on the relative freqencies $M_n(x)$ of the sides $x = 1, \ldots, 6$ in these n throws. Let μ_0 be the uniform distribution on $S := \{1, \ldots, 6\}$ and imagine that sides the dice come up according to some other, unknown distribution μ_1. We are looking for a test function $T : \mathcal{M}_1(S) \to \{0, 1\}$ such that if $T(M_n) = 1$, we reject the hypothesis that the dice is fair. Let $\mathbb{P}_\mu$ denote the distribution of M_n when in a single throw, the sides of the dice come up with law μ. Then

$$\alpha_n := \mathbb{P}_{\mu_0}[T(M_n) = 1] \quad \text{and} \quad \beta_n := \mathbb{P}_{\mu_1}[T(M_n) = 0]$$

are the probability that we incorrectly reject the hypothesis that the dice is fair and the probability that we do not reckognize the non-fairness of the dice, respectively. A good test minimalizes β_n when α_n is subject to a bound of the form $\alpha_n \leq \varepsilon$, with $\varepsilon > 0$ small and fixed. Consider a test of the form

$$T(M_n) := 1_{\{H(M_n|\mu_0) \geq \lambda\}},$$

where $\lambda > 0$ is fixed and small enough such that $\{\mu \in \mathcal{M}_1(S) : H(\mu|\mu_0) \geq \lambda\} \neq \emptyset$. Prove that

$$\lim_{n\to\infty} \frac{1}{n} \log \alpha_n = -\lambda,$$

and, for any $\mu_1 \neq \mu_0$,

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n = - \inf_{\mu:\, H(\mu|\mu_0)<\lambda} H(\mu|\mu_1).$$

Let $\tilde{T} : \mathcal{M}_1(S) \to \{0,1\}$ be any other test such that $\{\mu \in \mathcal{M}_1(S) : \tilde{T}(\mu) = 1\}$ is the closure of its interior and let $\tilde{\alpha}_n, \tilde{\beta}_n$ be the corresponding error probabilities. Assume that

$$\limsup_{n\to\infty} \frac{1}{n} \log \tilde{\alpha}_n \leq -\lambda.$$

Show that for any $\mu_1 \neq \mu_0$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \tilde{\beta}_n \geq - \inf_{\mu:\, H(\mu|\mu_0)<\lambda} H(\mu|\mu_0).$$

This shows that the test T is, in a sense, optimal.

Exercise 4.39 (Reducible Markov chains) Let $X = (X_k)_{k\geq 0}$ be a Markov chain with finite state space S and transition kernel P. Assume that $S = A \cup B \cup \{c\}$ where

(i) $\forall a, a' \in A\ \exists n \geq 0$ s.t. $P^n(a, a') > 0$,

(ii) $\forall b, b' \in B\ \exists n \geq 0$ s.t. $P^n(b, b') > 0$,

(iii) $\exists a \in A,\ b \in B$ s.t. $P(a, b) > 0$,

(iv) $\exists b \in B$ s.t. $P(b, c) > 0$,

(v) $P(a, c) = 0\ \forall a \in A$,

(vi) $P(b,a) = 0 \ \forall a \in A, \ b \in B$.

Assume that $X_0 \in A$ a.s. Give an expression for

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[X_n \neq c].$$

Hint: set $\tau_B := \inf\{k \geq 0 : X_k \in B\}$ and consider the process before and after τ_B.

Exercise 4.40 (Sampling without replacement) For each $n \geq 1$, consider an urn with n balls that have colors taken from some finite set S. Let $c_n(x)$ be the number of balls of color $x \in S$. Imagine that we draw m_n balls from the urn without replacement. We assume that the numbers $c_n(x)$ and m_n are deterministic (i.e., non-random), and that

$$\frac{1}{n} c_n(x) \underset{n\to\infty}{\longrightarrow} \mu(x) \quad (x \in S) \quad \text{and} \quad \frac{m_n}{n} \underset{n\to\infty}{\longrightarrow} \kappa,$$

where μ is a probability measure on S and $0 < \kappa < 1$. Let $M_n(x)$ be the (random) number of balls of color x that we have drawn. Let $k_n(x)$ satisfy

$$\frac{k_n(x)}{m_n} \underset{n\to\infty}{\longrightarrow} \nu_1(x) \quad \text{and} \quad \frac{c_n(x) - k_n(x)}{n - m_n} \underset{n\to\infty}{\longrightarrow} \nu_2(x) \qquad (x \in S),$$

where ν_1, ν_2 are probability measures on S such that $\nu_i(x) > 0$ for all $x \in S$, $i = 1, 2$. Prove that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[M_n = k_n] = -\kappa H(\nu_1|\mu) - (1-\kappa) H(\nu_2|\mu). \tag{4.26}$$

Sketch a proof, similar to the arguments following (4.9), that the laws $\mathbb{P}[M_n \in \cdot\,]$ satisfy a large deviation principle with speed n and rate function given by the right-hand side of (4.26). Hint: use Stirling's formula to show that

$$\frac{1}{n} \log \binom{n}{m} \approx H\big(\frac{m}{n}\big),$$

where

$$H(z) := -z \log z - (1-z) \log(1-z).$$

Exercise 4.41 (Conditioned Markov chain) Let S be a finite set and let P be a probability kernel on S. Let

$$U \subset \{(x,y) \in S^2 : P(x,y) > 0\}$$

be irreducible, let $\overline{U}$ be as in (4.3), and let A be the restriction of P to $\overline{U}$, i.e., A is the linear operator on $\mathbb{R}^{\overline{U}}$ whose matrix is given by $A(x,y) := P(x,y)$ $(x,y \in \overline{U})$. Let α, η and h denote its Perron-Frobenius eigenvalue and associated left and right eigenvectors, respectively, normalized such that $\sum_{x\in\overline{U}} h(x)\eta(x) = 1$, and let A_h be the irreducible probability kernel on $\overline{U}$ defined as in (4.17).

Fix $x_0 \in \overline{U}$, let $X = (X_k)_{k\geq 0}$ be the Markov chain in S with transition kernel P started in $X_0 = x_0$, and let $X^h = (X^h_k)_{k\geq 0}$ be the Markov chain in $\overline{U}$ with transition kernel A_h started in $X^h_0 = x_0$. Show that

$$\begin{aligned}&\mathbb{P}\big[X_1 = x_1, \ldots, X_n = x_n \,\big|\, (X_{k-1}, X_k) \in U \ \forall k = 1, \ldots, n\big]\\ &\quad \mathbb{E}\big[h^{-1}(X^h_n)\big]^{-1}\mathbb{E}\big[1_{\{X^h_1 = x_1, \ldots, X^h_n = x_n\}} h^{-1}(X^h_n)\big],\end{aligned}$$

where h^{-1} denotes the function $h^{-1}(x) = 1/h(x)$. Assuming moreover that A_h is aperiodic, prove that

$$\begin{aligned}&\mathbb{P}\big[X_1 = x_1, \ldots, X_m = x_m \,\big|\, (X_{k-1}, X_k) \in U \ \forall k = 1, \ldots, n\big]\\ &\quad \underset{n\to\infty}{\longrightarrow} \mathbb{P}\big[X^h_1 = x_1, \ldots, X^h_m = x_m\big]\end{aligned}$$

for each fixed $m \geq 1$ and $x_1, \ldots, x_m \in \overline{U}$. Hint:

$$\begin{aligned}&\mathbb{P}\big[X_1 = x_1, \ldots, X_m = x_m \,\big|\, (X_{k-1}, X_k) \in U \ \forall k = 1, \ldots, n\big]\\ &\quad (A^n_h h^{-1})(x_0)^{-1}\mathbb{E}\big[1_{\{X^h_1 = x_1, \ldots, X^h_m = x_m\}}(A^{n-m}_h h^{-1})(X^h_m)\big].\end{aligned}$$

Bibliography

[Aco02] A. de Acosta. Moderate deviations and associated Laplace transformations for sums of independent random vectors. *Trans. Am. Math. Soc.* 329(1), 357–375, 2002.

[Bil99] P. Billingsley. *Convergence of Probability Measures. 2nd ed.* Wiley, New York, 1999.

[Bou58] N. Bourbaki. *Éléments de Mathématique. VIII. Part. 1: Les Structures Fondamentales de l'Analyse. Livre III: Topologie Générale. Chap. 9: Utilisation des Nombres Réels en Topologie Générale. 2iéme éd.* Actualités Scientifiques et Industrielles 1045. Hermann & Cie, Paris, 1958.

[Bry90] W. Bryc. Large deviations by the asymptotic value method. Pages 447–472 in: *Diffusion Processes and Related Problems in Analysis* Vol. 1 (ed. M. Pinsky), Birkhäuser, Boston, 1990.

[Cho69] G. Choquet. *Lectures on Analysis. Volume I. Integration and Topological Vector Spaces.* Benjamin, London, 1969.

[Cra38] H. Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles* 736, 5—23, 1938.

[Csi06] I. Csiszár. A simple proof of Sanov's theorem. *Bull. Braz. Math. Soc. (N.S.)* 37(4), 453–459, 2006.

[DB81] C.M. Deo and G.J. Babu. Probabilities of moderate deviations in Banach spaces. *Proc. Am. Math. Soc.* 83(2), 392–397, 1981.

[DE97] P. Dupuis and R.S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations.* Wiley Series in Probability and Statistics. Wiley, Chichester, 1997.

[DS89] J.-D. Deuschel and D.W. Stroock. *Large Deviations.* Academic Press, Boston, 1989.

[Dud02] R.M. Dudley. *Real Analysis and Probability.* Reprint of the 1989 edition. Camebridge University Press, Camebridge, 2002.

[DZ93] A. Dembo and O. Zeitouni. *Large deviations techniques and applications.* Jones and Bartlett Publishers, Boston, 1993.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications 2nd edition.* Applications of Mathematics 38. Springer, New York, 1998.

[EL03] P. Eichelsbacher and M. Löwe. Moderate deviations for i.i.d. random variables. *ESAIM, Probab. Stat.* 7, 209–218, 2003.

[Ell85] R.S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics.* Grundlehren der Mathematischen Wissenschaften 271. Springer, New York, 1985.

[Eng89] R. Engelking. *General Topology.* Heldermann, Berlin, 1989.

[EK86] S.N. Ethier and T.G. Kurtz. *Markov Processes; Characterization and Convergence.* John Wiley & Sons, New York, 1986.

[Gan00] F.R. Gantmacher. *The Theory of Matrices, Vol. 2.* AMS, Providence RI, 2000.

[Hol00] F. den Hollander. *Large Deviations.* Fields Institute Monographs 14. AMS, Providence, 2000.

[Kel75] J.L. Kelley. *General Topology.* Reprint of the 1955 edition printed by Van Nostrand. Springer, New York, 1975.

[Led92] M. Ledoux. Sur les déviations modérés des sommes de variables aléatoires vectorielles indépendantes de même loi. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 28(2), 267–280, 1992.

[OV91] G.L. O'Brien and W. Verwaat. Capacities, large deviations and loglog laws. Page 43–83 in: *Stable Processes and Related Topics* Progress in Probability 25, Birkhäuser, Boston, 1991.

[Oxt80] J.C. Oxtoby. *Measure and Category. Second Edition.* Springer, New York, 1980.

[Puk91] A.A. Pukhalski. On functional principle of large deviations. Pages 198–218 in: *New Trends in Probability and Statistics* (eds. V. Sazonov and T. Shervashidze) VSP-Mokslas, 1991.

[Puh01] A. Puhalskii. *Large Deviations and Idempotent Probability.* Monographs and Surveys in Pure and Applied Mathematics 119. Chapman & Hall, Boca Raton, 2001.

[RS15] F. Rassoul-Agha and Timo Seppäläinen. *A Course on Large Deviations with an Introduction to Gibbs Measures.* Graduate studies in Mathemematics 162, AMS, 2015.

[Roc70] R.T. Rockafellar. *Convex Analysis.* Princeton, New Jersey, 1970.

[San61] I.N. Sanov. On the probability of large deviations of random variables. *Mat. Sb.* 42 (in Russian). English translation in: *Selected Translations in Mathematical Statistics and Probability I*, 213–244, 1961.

[Sen73] E. Seneta. *Non-Negative Matrices: An Introduction to Theory and Applications.* George Allen & Unwin, London, 1973.

[Ste87] J. Štěpán. *Teorie Pravěpodobnosti.* Academia, Prague, 1987.

[Var66] S.R.S. Varadhan. Asymptotic probabilities and differential equations. *Comm. Pure Appl. Math.* 19, 261–286, 1966.

Index