Large Deviation Theory

J.M. Swart

February 12, 2023

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. For this reason, the lecture notes start with a number of motivating examples which will be proved in the later sections. Also, the development of the abstract theory is at regular intervals interrupted in order to show how it can be applied to concrete examples. 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 $I(y_+) = -\log \mu(\{y_+\})$.*

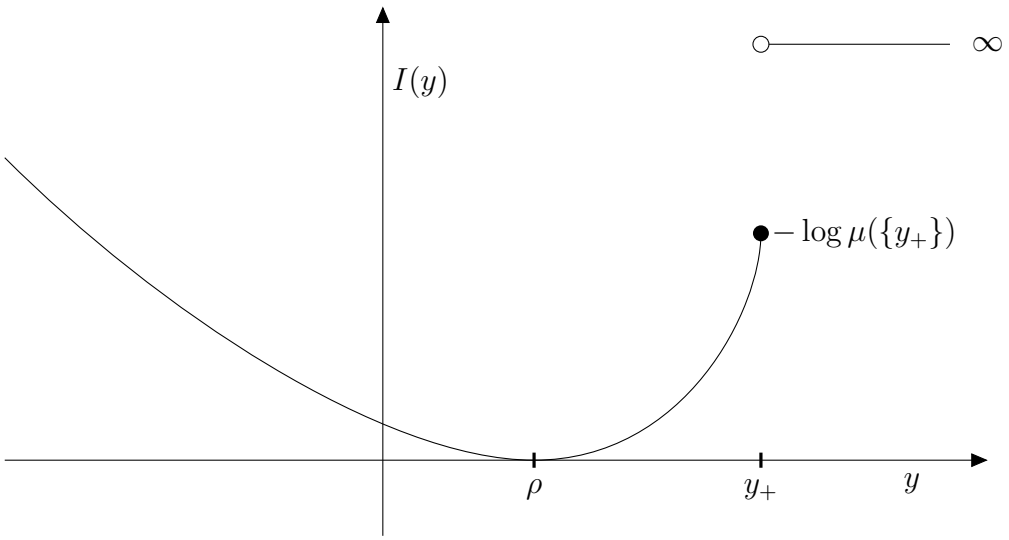

Figure 1: A typical example of a rate function.

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,

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

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

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

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

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). Moderate deviations are treated in [RS15, Section 11.2]. Some other 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 $0 \cdot \infty := 0$. 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 $S := \{1, \ldots, 6\}$, $\mu(x) := 1/6$ 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*

$$\begin{aligned} &\text{(i)} \quad f > 0 \text{ on } U, \\ &\text{(ii)} \quad f = 0 \text{ on } S\backslash U, \\ &\text{(iii)} \quad Gf(x) = -\lambda f(x) \qquad (x \in U). \end{aligned}$$

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 Chapter 2, we will already be able to prove Theorems 0.1 and 0.7. The proofs of other results, such as Theorems 0.4 and 0.8 will be postponed till later chapters when we have more abstract theory at our disposal.

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 $\text{int}(A) := \bigcup\{O : U \subset A,\ O \text{ open}\}$. Then clearly $\text{int}(A)$ 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*. A metrizable topology is uniquely characterized by its convergent sequences. Indeed, a subset A of a metrizable space is closed if and only if $x_n \in A$ and $x_n \to x$ imply $x \in A$, and once we know which sets are closed, we also know which sets are open, since they are the complements of closed sets.

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 $d(x_k, x_l) \le \varepsilon$ for all $k, l \ge n$. 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. The following proposition gives yet another characterization of weak convergence.

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

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

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

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 $\mathcal{O}_{\mathrm{up}} := \{[-\infty, a) : -\infty < a \le \infty\} \cup \{\emptyset, \overline{\mathbb{R}}\}$. Show that $\mathcal{O}_{\mathrm{up}}$ 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 $\mathcal{O}_{\rm up}$. The topology $\mathcal{O}_{\rm up}$ 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 $\Delta_n = \{a_0, \dots, a_{m(n)}\}$ with $a_0 < \cdots < a_{m(n)}$ and set

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

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'

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

We claim that

(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 $f + g \in \mathcal{G}_{\rm up}$ (resp. $f + g \in \mathcal{G}_{\rm low}$).

(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 statements (a) and (b) are easy. To prove (c), let $f_n \in \mathcal{G}_{\rm up}$, $f_n \downarrow f$. 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 $\mathcal{G}_{\rm low}$.

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 $f_k \in \mathcal{G}_{\rm up}$ by the fact that $\mu_n \Rightarrow \mu$ and therefore, by our claim (c) above, it follows that $1_C \in \mathcal{G}_{\rm up}$, 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 $\mathcal{G}_{\rm up}$, hence by Lemma 1.4 and claim (c), $\mathcal{U}_{b,+}(E) \subset \mathcal{G}_{\rm up}$. Similarly, condition (ii) implies that $\mathcal{L}_{b,+}(E) \subset \mathcal{G}_{\rm low}$. 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:

$$\text{A good rate function assumes its minimum on closed sets.} \tag{1.1}$$

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

$$\|f\|_{\infty,I} := \sup_{x\in E} e^{-I(x)} |f(x)| \qquad (f \in B_b(E)). \tag{1.2}$$

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

$$\lim_{n\to\infty} \|f\|_{s_n,\mu_n} = \|f\|_{\infty,I} \qquad \big(f \in \mathcal{C}_{b,+}(E)\big). \tag{1.3}$$

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

(ii) $\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(O) \geq -\inf_{x\in O} I(x) \qquad \forall O$ *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)' $\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) \leq -\inf_{x\in\overline{A}} I(x),$

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

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 $\overline{A} = \overline{\text{int}(A)}$, 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 (1.2). 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*

$$\limsup_{n\to\infty} \big(a_n^{s_n} + b_n^{s_n}\big)^{1/s_n} = \big(\limsup_{n\to\infty} a_n\big) \vee \big(\limsup_{n\to\infty} b_n\big). \tag{1.4}$$

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

$$\limsup_{n\to\infty} \frac{1}{s_n} \log(c_n + d_n) = \big(\limsup_{n\to\infty} \frac{1}{s_n} \log c_n\big) \vee \big(\limsup_{n\to\infty} \frac{1}{s_n} \log d_n\big). \tag{1.5}$$

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 (1.4).

We claim that (1.5) is just (1.4) in another guise. Indeed, setting $a_n := c_n^{1/s_n}$ and $b_n := d_n^{1/s_n}$ we see, using (1.4), 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 $f, g \in \mathcal{G}_{\rm up}$. Then, by (1.4),

$$\begin{aligned} &\limsup_{n\to\infty} \|f\vee g\|_{s_n,\mu_n} \\ &\quad= \limsup_{n\to\infty} \Big(\int_{\{x:f(x)\geq g(x)\}} f(x)^{s_n}\mu_n(\mathrm{d}x) + \int_{\{x:f(x)<g(x)\}} g(x)^{s_n}\mu_n(\mathrm{d}x)\Big)^{1/s_n} \\ &\quad\leq \limsup_{n\to\infty} \big(\|f\|_{s_n,\mu_n}^{s_n} + \|g\|_{s_n,\mu_n}^{s_n}\big)^{1/s_n} \leq \|f\|_{\infty,I} \vee \|g\|_{\infty,I} = \|f\vee g\|_{\infty,I}, \end{aligned} \tag{1.6}$$

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

To prove part (c), finally, assume that $f_k \in \mathcal{G}_{\rm up}$ 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 $f \in \mathcal{G}_{\rm up}$. The proof for $f_k \in \mathcal{G}_{\rm low}$ 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), $1_C \in \mathcal{G}_{\rm up}$ for each closed $C \subset E$ and $1_O \in \mathcal{G}_{\rm up}$ 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),

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

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 5.4 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 $y = 1$.

Exercise 1.12 (LDP for linear combination) Let E be a Polish space, let μ_n and ν_n be finite measures on E, and let s_n be positive constants converging to ∞. Assume that the μ_n satisfy the large deviation principle with speed s_n and good rate function I, and that the μ_n satisfy the large deviation principle with speed s_n and good rate function J. Let r, q be positive constants. Show that the measures $r\mu_n + q\nu_n$ satisfy the large deviation principle with speed s_n and good rate function $I \wedge J$. *Hint* Lemma 1.10.

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 1.13 (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 (1.3). 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 1.14 (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 (1.7) below.

Proposition 1.15 (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*

$$J(y) := \inf_{x \in \psi^{-1}(\{y\})} I(x) \qquad (y \in F), \tag{1.7}$$

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

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n \circ \psi^{-1}(C) &\leq - \inf_{x \in \psi^{-1}(C)} I(x) \\ &= - \inf_{y \in C} \inf_{x \in \psi^{-1}(\{y\})} I(x) = - \inf_{y \in C} J(y), \end{aligned} \tag{1.8}$$

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 (1.8)) 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. ∎

The following 'restriction principle' is sometimes also useful. We note that a subset F of a Polish space E is Polish in the induced topology if and only if F is a G_δ-subset of E; see Proposition 3.9 below.

Lemma 1.16 (Restriction principle) *Let E be a Polish space and let $F \subset E$ be a subset of E that is Polish in 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

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

[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!

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

Remark The implication (i)⇒(ii) of Lemma 1.16 alternatively follows by applying the contraction principle to the identity map $F \ni x \mapsto x \in E$.

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 1.17 (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 1.17 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 1.18 (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

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

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 1.19 (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*

$$\lim_{n\to\infty} \frac{1}{s_n} \log \mathbb{P}\big[d(X_n, Y_n) \geq \varepsilon\big] = -\infty \qquad (\varepsilon > 0). \tag{1.9}$$

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

Remark If (1.9) holds, then we say that the random variables X_n and Y_n are *exponentially close*. Note that condition (1.9) 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 1.19 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 (1.5) 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 (1.9),

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

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

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

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

Proposition 1.19 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 1.20 (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

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

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

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

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

Exercise 1.21 (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 $-\infty < \sup_{x\in E} F(x) < \infty$. Let $F_n : E \to \mathbb{R}$ be measurable and assume that $\lim_{n\to\infty} \|F_n - F\|_\infty = 0$. Show that the measures

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

satisfy the large deviation principle with speed s_n and good rate function $J := I - F$. Hint: combine Lemmas 1.18 and 1.20.

Chapter 2

Some first results

2.1 Relative entropy

In this chapter we prove some simple large deviation principles. The methods in the present chapter will largely be superceded by more powerful methods such as the Gärtner-Ellis theorem that will be proved in Chapter 5, but it is nice to see what can already be done with the more elementary theory developed so far. Our first aim is to prove the Boltzmann-Sanov theorem (Theorem 0.7). As a preparation, we study its rate function, the relative entropy. In particular, in the present section, we will prove Lemma 0.6.

Lemma 2.1 (Elementary properties) *Let S be a finite set and let $\mu \in \mathcal{M}_1(S)$ satisfy $\mu(x) > 0$ for all $x \in S$. Then $H(\nu|\mu) \geq 0$ for all $\nu \in \mathcal{M}_1(S)$ with equality if and only if $\nu = \mu$.*

Proof Let $\psi : [0, \infty] \to [0, \infty]$ be defined by

$$\psi(z) := \int_1^z \mathrm{d}y \int_1^y \mathrm{d}x \, \frac{1}{x} \qquad \big(x \in [0, \infty]\big). \tag{2.1}$$

It is easy to see that ψ is strictly convex, assumes its minimal value 0 in the point $z = 1$, and satisfies

$$\psi(0) = 1, \quad \psi(z) = z \log z - z + 1 \quad (0 < z < \infty), \quad \psi(\infty) = \infty.$$

We observe that for each $\nu \in \mathcal{M}_1(S)$,

$$\begin{aligned}\sum_{x\in S}\mu(x)\psi\Big(\frac{\nu(x)}{\mu(x)}\Big) &= \sum_{x\in S}\mu(x)\frac{\nu(x)}{\mu(x)}\log\Big(\frac{\nu(x)}{\mu(x)}\Big) - \sum_{x\in S}\mu(x)\frac{\nu(x)}{\mu(x)} + \sum_{x\in S}\mu(x) \\ &= H(\nu|\mu) - 1 + 1 = H(\nu|\mu).\end{aligned} \tag{2.2}$$

Since $\psi(z) \geq 0$ with equality if and only if $z = 1$, the claim of the lemma follows. ∎

Lemma 2.2 (Convexity of the relative entropy) *Let S be a finite set and let $\mu \in \mathcal{M}_1(S)$ satisfy $\mu(x) > 0$ for all $x \in S$. Let $p_1, \ldots, p_n$ be nonnegative constants summing up to one and let $\nu := \sum_{k=1}^n p_k\nu_k$ with $\nu_k \in \mathcal{M}_1(S)$. Then*

$$H(\nu|\mu) = \sum_{k=1}^n p_k H(\nu_k|\mu) - \sum_{k=1}^n p_k H(\nu_k|\nu). \tag{2.3}$$

In particular, the function $\mathcal{M}_1(S) \ni \nu \mapsto H\big(\nu|\mu\big)$ is strictly convex.

Proof This follows by writing

$$\begin{aligned}H\big(\nu|\mu\big) &= \sum_{x\in S}\sum_{k=1}^n p_k\nu_k(x)\log\Big(\frac{\nu(x)}{\mu(x)}\Big) \\ &= \sum_{k=1}^n p_k\sum_{x\in S}\nu_k(x)\Big[\log\Big(\frac{\nu_k(x)}{\mu(x)}\Big) - \log\Big(\frac{\nu_k(x)}{\nu(x)}\Big)\Big] \\ &= \sum_{k=1}^n p_k\Big[\sum_{x\in S}\nu_k(x)\log\Big(\frac{\nu_k(x)}{\mu(x)}\Big) - \sum_{x\in S}\nu_k(x)\log\Big(\frac{\nu_k(x)}{\nu(x)}\Big)\Big].\end{aligned}$$

∎

Proof of Lemma 0.6 The fact that $H(\nu|\mu) < \infty$ for all ν follows from formula (2.2). Now properties (i)–(iii) follow from Lemma 2.1. The fact that $\nu \mapsto H(\nu|\mu)$ is convex has been proved in Lemma 2.2. Since the function ψ is continuous on $[0, \infty)$ and infinitely differentiable on $(0, \infty)$, we see from formula (2.2) that the function $\nu \mapsto H(\nu|\mu)$ is continuous on $\mathcal{M}_1(S)$ and infinitely differentiable on the interior of $\mathcal{M}_1(S)$, i.e., on the set $\mathring{\mathcal{M}}_1(S) := \{\nu \in \mathcal{M}_1(S) : \nu(x) > 0\ \forall x \in S\}$. ∎

Exercise 2.3 (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

$$\mathcal{M}_1(S) \times \mathring{\mathcal{M}}_1(S) \ni (\nu, \mu) \mapsto H(\nu|\mu).$$

2.2 The Boltzmann-Sanov theorem

In this section, we prove the Boltzmann-Sanov theorem (Theorem 0.7 from the introduction). As in Section 0.3, S is a finite set and μ is a probability measure on S. We generalize a little bit and drop the condition from Section 0.3 that $\mu(x) > 0$ for all $x \in S$. Generalizing our earlier definition, we set

$$H(\nu|\mu) := \begin{cases} \displaystyle\sum_{x\in S} \nu(x) \log \frac{\nu(x)}{\mu(x)} = \sum_{x\in S} \mu(x)\psi\Big(\frac{\nu(x)}{\mu(x)}\Big) & \text{if } \nu \ll \mu, \\ \infty & \text{otherwise,} \end{cases}$$

where ψ is defined in (2.1) and the notation $\nu \ll \mu$ means that ν is absolutely continuous with respect to μ, i.e., $\nu(x) = 0$ whenever $\mu(x) = 0$. We let $(X_k)_{k\geq 1}$ be i.i.d. with common law μ and define the *empirical distributions* $(M_n)_{n\geq 1}$ as in (0.6), i.e.,

$$M_n := \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k},$$

where δ_x denotes the delta-measure at $x \in S$. We let

$$\rho_n := \mathbb{P}\big[M_n \in \cdot\,\big]$$

denote the law of M_n. Note that $M_n \in \mathcal{M}_1(S)$ and hence the law of M_n is a probability measure on $\mathcal{M}_1(S)$, i.e., $\rho_n \in \mathcal{M}_1(\mathcal{M}_1(S))$. We will prove the following theorem.

Theorem 2.4 (Boltzmann-Sanov) *The probability laws ρ_n satisfy the large deviation principle with speed n and good rate function $H(\,\cdot\,|\mu)$.*

Together with Lemma 0.6 this implies Theorem 0.7 from the introduction. Recall that in Theorem 0.7 we assumed that $\mu(x) > 0$ for all $x \in S$. Now Lemma 0.6 (iv) tells us that the rate function $\nu \mapsto H(\nu|\mu)$ is continuous, and so each closed set $C \subset \mathcal{M}_1(S)$ that is the closure of its interior is a continuity set for the rate function. Therefore, the large deviation principle implies (0.7), see Remark 1 below Proposition 1.7.

We let $\mathcal{M}_1^n(S)$ denote the space of all probability measures on S of the form

$$\frac{1}{n} \sum_{k=1}^{n} \delta_{x_k}$$

for some $x_1, \ldots, x_n \in S$. We let

$$\eta_n := \sum_{\nu \in \mathcal{M}_n^n(S)} \delta_\nu$$

denote the counting measure on $\mathcal{M}_1^n(S)$ and we define $\Phi_n : \mathcal{M}_1^n(S) \to \mathbb{R}$ by

$$\Phi_n(\nu) := \mathbb{P}\big[M_n = \nu\big] \qquad \big(\nu \in \mathcal{M}_1^n(S)\big).$$

We will derive Theorem 2.4 from the following two lemmas.

Lemma 2.5 (LDP for the reference measure) *Assume that $\mu(x) > 0$ for all $x \in S$. Then the measures η_n satisfy the large deviation principle with speed n and good rate function I defined as $I(\nu) := 0$ $(\nu \in \mathcal{M}_1(S))$.*

Lemma 2.6 (Convergence of the exponential density) *Assume that $\mu(x) > 0$ for all $x \in S$. Then the functions $F_n := n^{-1} \log \Phi_n$ satisfy*

$$\lim_{n\to\infty} \sup_{\nu \in \mathcal{M}_1^n(S)} \big|F_n(\nu) + H(\nu|\mu)\big| = 0.$$

We first show that these two lemmas imply Theorem 2.4 and then prove the lemmas.

Proof of Theorem 2.4 We first prove the claim under the additional assumption that that $\mu(x) > 0$ for all $x \in S$. Then by Lemma 0.6, the function $F(\nu) := -H(\nu|\mu)$ $(\nu \in \mathcal{M}_1(S))$ is continuous. Since $\mathcal{M}_1(S)$ is compact, F is moreover bounded from below and above. We have $\rho_n = e^{F_n}\eta_n$ $(n \geq 1)$, so by Lemmas 2.5 and 2.6 we can apply Exercise 1.21 to conclude that the ρ_n satisfy the large deviation principle with speed n and good rate function $I - F = 0 + H(\,\cdot\,|\mu)$.

To prove the general claim, we set $S' := \{x \in S : \mu(x) > 0\}$. By what we have already proved, the measures ρ_n, viewed as measures on $\mathcal{M}_1(S')$, satisfy the large deviation principle with speed n and good rate function $H(\,\cdot\,|\mu)$. We can naturally view $\mathcal{M}_1(S')$ as a closed subset of $\mathcal{M}_1(S)$. Since $H(\nu|\mu) = \infty$ for all $\nu \notin \mathcal{M}_1(S')$, the claim now follows from the restriction principle (Lemma 1.16). ∎

Before we prove Lemmas 2.5 and 2.6, we introduce more notation. We let $\mathcal{N}^n(S)$ be the set of all functions $\kappa : S \to \mathbb{N}$ such that $\sum_{x\in S} \kappa(x) = n$. Then

$$\mathcal{M}_1^n(S) = \big\{n^{-1}\kappa : \kappa \in \mathcal{N}_n\big\}.$$

Proof of Lemma 2.5 We check the conditions of Proposition 1.7. If $C \subset \mathcal{M}_1(S)$ is closed, then we can estimate $\rho_n(C) \leq \rho_n(\mathcal{M}_1(S)) = |\mathcal{N}^n(S)| \leq n^{|S|}$, which gives

$$\limsup_{n\to\infty} \frac{1}{n} \log \rho_n(C) \leq \limsup_{n\to\infty} \frac{|S|}{n} \log n = 0.$$

If $O \subset \mathcal{M}_1(S)$ is open and nonempty, then $O \cap \mathcal{M}_1^n(S) \neq \emptyset$ for n large enough and hence

$$\liminf_{n\to\infty} \frac{1}{n} \log \rho_n(C) \geq \liminf_{n\to\infty} \frac{1}{n} \log 1 = 0.$$

∎

Proof of Lemma 2.6 We have

$$\Phi_n(n^{-1}\kappa) = \mathbb{P}\big[M_n = n^{-1}\kappa\big] = n! \prod_{x\in S} \frac{1}{\kappa(x)!} \mu(x)^{\kappa(x)} \qquad \big(\kappa \in \mathcal{N}_n(S)\big),$$

Stirling's formula[1] says that $\log n! = n \log n - n + O(\log n)$, which gives

$$\begin{aligned}
&\log \Phi_n(n^{-1}\kappa) \\
&\quad = (n \log n - n) - \sum_{x\in S} \big(\kappa(x) \log \kappa(x) - \kappa(x)\big) + \sum_{x\in S} \kappa(x) \log \mu(x) + O(\log n) \\
&\quad = n \log n - \sum_{x\in S} \kappa(x) \log \kappa(x) + \sum_{x\in S} \kappa(x) \log \mu(x) + O(\log n),
\end{aligned}$$

where the error term can be estimated from above by $C \log n$, with a constant C that does not depend on $\kappa \in \mathcal{N}^n(S)$. Dividing by n, this gives

$$\begin{aligned}
\frac{1}{n} \log \Phi_n(n^{-1}\kappa) &= -\sum_{x\in S} \frac{\kappa(x)}{n} \log \Big(\frac{\kappa(x)}{n\mu(x)}\Big) + O(n^{-1} \log n) \\
&= -H\big(n^{-1}\kappa \,\big|\, \mu\big) + O(n^{-1} \log n).
\end{aligned}$$

∎

2.3 An LDP for pair empirical measures

Let S be a finite set, let μ be a probability law on S such that $\mu(x) > 0$ for all $x \in S$, and let $(X_k)_{k\geq 0}$ be i.i.d. with common law μ. Similar to what we did before,

[1] Recall that Stirling's formula says that $n! \sim \sqrt{2\pi n}(n/e)^n$.

we define the *pair empirical distributions* $(M_n^{(2)})_{n\geq 1}$ by

$$M_n^{(2)} := \frac{1}{n}\sum_{k=1}^{n} \delta_{(X_{k-1},X_k)},$$

where $\delta_{(x,y)}$ denotes the delta-measure at $(x,y) \in S^2$. We let

$$\rho_n^{(2)} := \mathbb{P}\big[M_n^{(2)} \in \cdot\,\big]$$

denote the law of $M_n^{(2)}$. Note that $\rho_n^{(2)} \in \mathcal{M}_1(\mathcal{M}_1(S^2))$. We will prove a large deviation principle for the measures $\rho_n^{(2)}$.

We first define the appropriate rate function. For any $\nu \in \mathcal{M}_1(S^2)$, we let

$$\nu^-(x) := \sum_{y\in S} \nu(x,y) \quad \text{and} \quad \nu^+(y) := \sum_{x\in S} \nu(x,y)$$

denote the left and right marginals of ν, and we let

$$\mathcal{V}(S) := \big\{\nu \in \mathcal{M}_1(S^2) : \nu^- = \nu^+\big\}$$

denote the space of probability laws ν on S^2 whose left and right marginals agree. We define $I_\mu^{(2)} : \mathcal{M}_1(S^2) \to \mathbb{R}$ by

$$I_\mu^{(2)}(\nu) := \begin{cases} H(\nu|\nu^- \otimes \mu) & \text{if } \nu \in \mathcal{V}(S), \\ \infty & \text{otherwise,} \end{cases}$$

where $\nu^- \otimes \mu$ denotes the product measure of ν^- and μ. We will prove the following theorem.

Theorem 2.7 (LDP for pair empirical measure) *The probability laws $\rho_n^{(2)}$ satisfy the large deviation principle with speed n and good rate function $I_\mu^{(2)}$.*

Exercise 2.8 (Contraction principle) Use the contraction principle to derive the Boltzmann-Sanov theorem (Theorem 2.4) from Theorem 2.7.

Our proof of Theorem 2.7 will be similar to the proof of Theorem 2.4. We start by transforming the problem into an easier one, by defining

$$\tilde{M}_n^{(2)} := \frac{1}{n}\Big(\delta_{(X_{n-1},X_0)} + \sum_{k=1}^{n-1} \delta_{(X_{k-1},X_k)}\Big). \tag{2.4}$$

The advantage of defining pair empirical measures in such a way, with periodic boundary conditions, is that $\tilde{M}_n^{(2)} \in \mathcal{V}(S)$ a.s. The following exercise shows that instead of proving a large deviation principle for the laws of $M_n^{(2)}$ we can equivalently prove a large deviation principle for the laws of $\tilde{M}_n^{(2)}$.

Exercise 2.9 (Exponential closeness) Prove that the random variables $M_n^{(2)}$ and $\tilde{M}_n^{(2)}$ are exponentially close with speed n, in the sense of Proposition 1.19.

We let $\mathcal{V}_n(S)$ denote the space of all probability measures on S^2 of the form

$$\frac{1}{n}\Big(\delta_{(x_{n-1},x_0)} + \sum_{k=1}^{n-1} \delta_{(x_{k-1},x_k)}\Big)$$

for some $x_0, \ldots, x_{n-1} \in S$. We let

$$\eta_n^{(2)} := \sum_{\nu \in \mathcal{V}_n(S)} \delta_\nu$$

denote the counting measure on $\mathcal{V}_n(S)$ and we define $\Phi_n : \mathcal{V}_n(S) \to \mathbb{R}$ by

$$\Phi_n^{(2)}(\nu) := \mathbb{P}\big[\tilde{M}_n^{(2)} = \nu\big] \qquad \big(\nu \in \mathcal{V}_n(S)\big).$$

We will derive Theorem 2.7 from the following two lemmas. We first show how the lemmas imply Theorem 2.7 and then prove the lemmas.

Lemma 2.10 (LDP for the reference measure) *The measures $\eta_n^{(2)}$ satisfy the large deviation principle with speed n and good rate function $I : \mathcal{M}_1(S^2) \to \mathbb{R}$ defined as $I(\nu) := 0$ if $\nu \in \mathcal{V}(S)$ and $:= \infty$ otherwise.*

Lemma 2.11 (Convergence of the exponential density) *The function $I_\mu^{(2)} : \mathcal{V}(S) \to \mathbb{R}$ is continuous and the functions $F_n^{(2)} := n^{-1} \log \Phi_n^{(2)}$ satisfy*

$$\lim_{n\to\infty} \sup_{\nu \in \mathcal{V}_n(S)} \big|F_n^{(2)}(\nu) + I_\mu^{(2)}(\nu)\big| = 0.$$

Proof of Theorem 2.7 Let $\tilde{\rho}_n^{(2)}$ denote the law of $\tilde{M}_n^{(2)}$. By Exercise 2.9, it suffices to prove the claim with $\rho_n^{(2)}$ replaced by $\tilde{\rho}_n^{(2)}$. By the restriction principle (Lemma 1.16) it suffices to prove that the measures $\tilde{\rho}_n^{(2)}$, viewed as probability measures on $\mathcal{V}(S)$ (rather than the larger space $\mathcal{M}_1(S^2)$) satisfy the large deviation

principle with speed n and good rate function $I_\mu^{(2)}$. By Lemma 2.11, the function $I_\mu^{(2)} : \mathcal{V}(S) \to \mathbb{R}$ is continuous and uniformly approximated by the functions $-F_n^{(2)}$. Since $F_n^{(2)}$ is the exponential density of $\tilde{\rho}_n^{(2)}$ with respect to $\eta_n^{(2)}$, and since by Lemma 2.10 the latter satisfy the large deviation principle with the trivial rate function $I = 0$, the claim follows by applying Exercise 1.21 in exactly the same was as in the proof of Theorem 2.4. ∎

We still need to prove Lemmas 2.10 and 2.11. The proof of Lemma 2.10 depends on the following lemma, the proof of which we postpone till the end of this section.

Lemma 2.12 (Approximation with finite spaces) *For each $\nu \in \mathcal{V}(S)$, there exist $\nu_n \in \mathcal{V}_n(S)$ such that $\nu_n \to \nu$.*

Proof of Lemma 2.10 We observe that $\mathcal{V}(S)$ is a closed subset of $\mathcal{M}_1(S^2)$, and hence, since $\mathcal{M}_1(S^2)$ is compact, so is $\mathcal{V}(S)$. By the restriction principle (Lemma 1.16), we may alternatively show that the restricted measures $\eta_n^{(2)}\big|_{\mathcal{V}(S)}$ satisfy the LDP with speed n and trivial rate function $I(\nu) = 0$ for all $\nu \in \mathcal{V}(S)$.

We apply Proposition 1.7. For any closed set $C \subset \mathcal{V}(S)$, we can estimate

$$\eta_n^{(2)}(C) = \big|\mathcal{V}_n(S) \cap C\big| \leq \big|\mathcal{V}_n(S)\big| \leq n^{S^2}$$

which shows that

$$\limsup_{n\to\infty} \frac{1}{n} \log \eta_n^{(2)}(C) \leq \lim_{n\to\infty} \frac{1}{n} |S^2| \log n = 0.$$

On the other hand, if $O \subset \mathcal{V}(S)$ is open and nonempty, then by Lemma 2.12, $\eta_n^{(2)}(O) \geq 1$ for all n sufficiently large, and hence

$$\limsup_{n\to\infty} \frac{1}{n} \log \eta_n^{(2)}(O) \geq 0.$$

∎

Proof of Lemma 2.11 The proof will be similar to the proof of Lemma 2.6, but the combinatorical arguments are more involved. We closely follow [Hol00, Section II.2] who in turn bases himself on [Ell85, Section I.5].

Let us write $\mathcal{K}_n(S)$ for the set of functions $\kappa : S^2 \to \mathbb{N}$ of the form

$$\kappa = \sum_{k=1}^{n} \delta_{(x_{k-1}, x_k)}$$

for some $x_0, \dots, x_n \in S$ with $x_0 = x_n$. Then $\mathcal{V}_n(S) = \{n^{-1}\kappa : \kappa \in \mathcal{K}_n(S)\}$. Then $\mathcal{V}_n(S) = \{n^{-1}\kappa : \kappa \in \mathcal{K}_n(S)\}$. For $\kappa \in \mathcal{K}_n(S)$, let us write

$$\overline{\kappa}(x) := \sum_{y \in S} \kappa(x,y) = \sum_{x \in S} \kappa(x,y) \qquad (x \in S)$$

for the left and right marginals of κ, which are equal. We set

$$\mathcal{C}_n(\kappa) := \big\{(x_0, \dots, x_{n-1}) \in S^n : \delta_{(x_{n-1},x_0)} + \sum_{k=1}^{n-1} \delta_{(x_{k-1},x_k)} = \kappa\big\}.$$

Then

$$\Phi_n^{(2)}(\kappa/n) = \big|\mathcal{C}_n(\kappa)\big| \prod_{x \in S} \mu(x)^{\overline{\kappa}(x)} \qquad \big(\kappa \in \mathcal{K}_n(S)\big). \tag{2.5}$$

In order to estimate $\mathcal{C}_n(\kappa)$, we draw a directed graph with vertex set S such that for each $(x,y) \in S^2$, there are $\kappa(x,y)$ distinct directed edges from x to y. Let $\mathcal{W}_n(\kappa)$ denote the set of distinct walks in this graph that use each directed edge precisely once. Note that each such walk must end at its starting point. Then

$$|\mathcal{C}_n(\kappa)| = \frac{|\mathcal{W}_n(\kappa)|}{\prod_{(x,y) \in S^2} \kappa(x,y)!}, \tag{2.6}$$

where the denominator takes care of the fact that if a walk jumps multiple times from x to y, then in the set $\mathcal{W}_n(\kappa)$ we do care about which directed edge is used in which jump, while in $\mathcal{C}_n(\kappa)$ this information is disgarded. We claim that

$$\prod_{x : \overline{\kappa}(x) > 0} \big(\overline{\kappa}(x) - x\big)! \le \big|\mathcal{W}_n(\kappa)\big| \le |S| \prod_x \overline{\kappa}(x)!\,. \tag{2.7}$$

To prove the upper bound, it suffices to note that we can uniquely specify a walk that uses each directed edge precisely once by specifying the starting vertex of such a walk and then by specifying for each vertex in which order the outgoing directed edges should be used at each consecutive visit of the vertex.

To prove the lower bound, we use the fact that by the definition of $\mathcal{K}_n(S)$, there is at least one walk in our graph that uses each directed edge precisely once. Let S' be the set if all vertices visited by this walk and let z be the starting vertex. For each vertex in $S' \backslash \{z\}$, we mark the directed edge that is used the last time the walk leaves this vertex. We observe that:

From each vertex in $S' \backslash \{z\}$, there is a path of marked edges leading to z. (2.8)

We can now construct new walks, starting at the same vertex as our old walk, by changing the order in which the unmarked directed edges leaving a vertex are used, but keeping the marked directed edges as the last ones. The lower bound in (2.7) will follow if we show that each chosen order of the unmarked outgoing edges yields a walk that uses each directed edge precisely once. Since at ach vertex, there is an equal number of incoming and outgoing edges, if we just "follow the instructions" until we arrive at a vertex where all outgoing edges have already been used, then that vertex must be z, and at that point we have used all incoming edges at z. We observe that if we have used a marked outgoing edge at some vertex x, then we have used all incoming edges at x. Using this and (2.8), we see that all marked edges have been used and hence all edges have been used.

Combining (2.5), (2.6), and (2.7), we obtain the bounds

$$\frac{\prod_{x:\overline{\kappa}(x)>0}\big(\overline{\kappa}(x)-x\big)!}{\prod_{(x,y)\in S^2}\kappa(x,y)!}\prod_{x\in S}\mu(x)^{\overline{\kappa}(x)} \\ \le \Phi_n^{(2)}(\kappa/n) \le |S|\frac{\prod_{x\in S}\overline{\kappa}(x)!}{\prod_{(x,y)\in S^2}\kappa(x,y)!}\prod_{x\in S}\mu(x)^{\overline{\kappa}(x)}$$

($\kappa \in \mathcal{K}_n(S)$). We take logarithms and divide by n and as we did in the proof of Lemma 2.6, we use Stirling's formula which says that $\log n! = n\log n - n + O(\log n)$. This yields

$$\begin{aligned}\tfrac{1}{n}\log\Phi_n^{(2)}(\kappa/n) = {}& \frac{1}{n}\sum_{x\in S}\big[\overline{\kappa}(x)\log\overline{\kappa}(x)-\overline{\kappa}(x)\big]+\frac{1}{n}\sum_{y\in S}\overline{\kappa}(y)\log\mu(y)\\ & -\frac{1}{n}\sum_{(x,y)\in S^2}\big[\kappa(x,y)\log\kappa(x,y)-\kappa(x,y)\big]+O(n^{-1}\log n).\end{aligned}$$

Using the fact that $\sum_x \overline{\kappa}(x) = \sum_{x,y}\kappa(x,y)$, we can simplify this to

$$\begin{aligned}\frac{1}{n}\log\Phi_n^{(2)}(\kappa/n) &= \frac{1}{n}\sum_{x\in S}\overline{\kappa}(x)\log\overline{\kappa}(x)+\frac{1}{n}\sum_{y\in S}\overline{\kappa}(y)\log\mu(y)\\ &\quad -\frac{1}{n}\sum_{(x,y)\in S^2}\kappa(x,y)\log\kappa(x,y)+O(n^{-1}\log n)\\ &= \frac{1}{n}\sum_{(x,y)\in S^2}\kappa(x,y)\Big[\log\overline{\kappa}(x)+\log\mu(y)-\log\kappa(x,y)\Big]+O(n^{-1}\log n).\end{aligned}$$

Rewriting this in terms of $\nu = \kappa/n$ yields

$$\begin{aligned}\frac{1}{n}\log \Phi_n^{(2)}(\nu) &= -\sum_{(x,y)\in S^2} \nu(x,y)\log\Big(\frac{\nu(x,y)}{\nu^-(x)\mu(y)}\Big) + O(n^{-1}\log n)\\ &= -H(\nu^- \otimes \mu|\nu) + O(n^{-1}\log n),\end{aligned}$$

where the error term is of order $n^{-1}\log n$ as $n \to \infty$, uniformly in $\nu \in \mathcal{V}_n(S)$. ∎

We still have to provide the proof of Lemma 2.12. This will be done in a number of steps.

By definition, a *cycle* in S is an ordered collection $C = (x_0, \ldots, x_n)$ of elements of S such that $n \geq 1$, $x_0 = x_n$ and $x_1, \ldots, x_n$ are all different. Each cycle $C = (x_0, \ldots, x_n)$ defines a probability measure ν_C on S^2 through the formula

$$\nu_C(y_0, y_1) := \frac{1}{n}\sum_{k=1}^{n} 1_{\{(x_{k-1},x_k)=(y_0,y_1)\}} \qquad (y_0, y_1 \in S).$$

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 $x = py + (1-p)z$.

Lemma 2.13 (Extremal elements) *The set $\mathcal{V}(S)$ is a compact convex subset of $\mathcal{M}_1(S^2)$, and its extremal elements are the measures of the form ν_C where C is a cycle in S.*

Proof It is easy to see that $\nu_C \in \mathcal{V}(S)$ for each cycle C in S and that measures of the form ν_C are extremal. To complete the proof, we must show that each extremal element if $\mathcal{V}(S)$ is of this form ν_C.

We claim that for each $\nu \in \mathcal{V}(S)$ and $(y_0, y_1) \in S^2$ such that $\nu(y_0, y_1) > 0$, we can find a cycle $C \in \mathcal{C}(S^2)$ and a constant $p > 0$ such that $p\nu_C \leq \nu$. Indeed, since $\sum_{y_2} \nu(y_1, y_2) = \nu^-(y_1) = \nu^+(y_1) \geq \nu(y_0, y_1)$, there must be some $y_2 \in S$ such that $\nu(y_1, y_2) > 0$. Continuing in this way, by the finiteness of S, we must arrive back at y_0 at some point, hence there must exist some cycle $C = (y_0, \ldots, y_n)$ such that $\nu(y_{k-1}, y_k) > 0$ for all $k = 1, \ldots, n$. Setting $p := n \inf\{\nu(y_{k-1}, y_k) : k = 1, \ldots, n\}$, our claim follows.

Now let $\nu \in \mathcal{V}(S)$ be extremal. By what we have just proved, there exists a cycle C and constant $0 < p < 1$ such that $p\nu_C \leq \nu$. Then we can write $\nu = p\nu_C + (1-p)\mu$, where $\mu := (1-p)^{-1}(\nu - c\nu_C)$. Since ν is extremal, we have $\mu = \nu_C$ and hence $\nu = \nu_C$. ∎

Lemma 2.14 (Approximation of extremal elements) *For each cycle $C = (x_0, \ldots, x_n)$, there exist $\nu_n \in \mathcal{V}_n(S)$ such that $\nu_n \to \nu_C$.*

Proof Let $[k]$ denote the remainder of k after division by n and define $(z_k)_{k\geq 0}$ by $z_k := x_{[k]}$ $(k \geq 0)$. Then it is easy to see that

$$\nu_n(y_0, y_1) := \frac{1}{n}\sum_{k=1}^{n} 1_{\{(z_{k-1},z_k)=(y_0,y_1)\}} \qquad (y_0, y_1 \in S)$$

defines $\nu_n \in \mathcal{V}_n(S)$ such that $\nu_n \to \nu_C$. ∎

Proof of Lemma 2.12 For any $x = (x_0, \ldots, x_n) \in S^{n+1}$, let

$$M^{(2)}(x) := \frac{1}{n}\sum_{k=1}^{n} \delta_{(x_{k-1},x_k)}.$$

Let

$$\begin{aligned}\mathcal{A} &:= \big\{\nu \in \mathcal{V}(S) : \exists \nu_n \in \mathcal{V}_n(S) \text{ s.t. } \nu_n \to \nu\big\}\\ &= \big\{\nu \in \mathcal{V}(S) : \exists x^n \in S^{n+1} \text{ s.t. } M_n^{(2)}(x^n) \to \nu\big\}.\end{aligned}$$

We claim that if $\nu, \mu \in \mathcal{A}$ and $0 \leq p \leq 1$, then $p\nu + (1-p)\mu \in \mathcal{A}$. To see this, choose $x^n \in S^{n+1}$ such that $M_n^{(2)}(x^n) \to \nu$ and $y^n \in S^{n+1}$ such that $M_n^{(2)}(y^n) \to \mu$. Then it is easy to check that setting

$$z^n := (x_0^{\lfloor pn\rfloor}, \ldots x_{\lfloor pn\rfloor}^{\lfloor pn\rfloor}, y_1^{\lceil (1-p)n\rceil}, \ldots y_{\lceil (1-p)n\rceil}^{\lceil (1-p)n\rceil})$$

defines a sequence of sequences z^n such that $M_n^{(2)}(z^n) \to p\nu + (1-p)\mu$. This shows that $\mathcal{A}$ is a convex set.

We next claim that $\mathcal{A}$ is a closed subset of $\mathcal{V}(S)$. To see this, let d be any metric generating the topology on $\mathcal{V}(S)$, and let $d(\nu, \mathcal{V}_n(S)) := \inf\{d(\nu, \nu') : \nu' \in \mathcal{V}_n(S)\}$. Then

$$\mathcal{A} = \big\{\nu \in \mathcal{V}(S) : d(\nu, \mathcal{V}_n(S)) \underset{n\to\infty}{\longrightarrow} 0\big\}.$$

Assume that $\nu_m \in \mathcal{A}$ converge to a limit $\nu \in \mathcal{V}(S)$. Then for each $\varepsilon > 0$, we can choose m such that $d(\nu, \nu_m) \leq \varepsilon/2$, and we can choose N such that $d(\nu_m, \mathcal{V}_n(S)) \leq \varepsilon/2$ for all $n \geq N$. It follows that $d(\nu, \mathcal{V}_n(S)) \leq \varepsilon$ for all $n \geq N$. Since $\varepsilon > 0$ is arbitrary, it follows that $\nu \in \mathcal{A}$.

By Lemmas 2.13 and 2.14, $\mathcal{V}(S)$ is a compact convex set and $\mathcal{A}$ contains all extremal elements of $\mathcal{V}(S)$. It is well-known that a compact convex set is the closure of the convex hull of its extremal elements [Roc70, Corollary 18.5.1], so we conclude that $\mathcal{A} = \mathcal{V}(S)$. ∎

2.4 A LDP for Markov chains

In this section we extend Theorem 2.7, which holds for i.i.d. sequences, to Markov chains. We first introduce some general notation.

Let S be a finite set and let $\mathbb{R}^S$ be the space of functions $f : S \to \mathbb{R}$. We equip $\mathbb{R}^S$ with the standard inner product

$$\langle f, g\rangle := \sum_{x\in S} f(x)g(x) \qquad (f, g \in \mathbb{R}^S).$$

By definition, a *matrix indexed by* S is a function $A : S \times S \to \mathbb{R}$. We define the product of two matrices in the usual way, as

$$(AB)(x, z) := \sum_{y\in S} A(x, y)B(y, z).$$

We set $A^0(x, y) := 1(x, y)$ with $1(x, x) := 1$ and $1(x, y) := 0$ for all $x \neq y$ and we define $A^n := A^{n-1}A$ $(n \geq 1)$. If $f : S \to \mathbb{R}$ is a function, then de define functions Af and fA by

$$Af(x) := \sum_{y\in S} A(x, y)f(y) \quad (x \in S) \quad \text{and} \quad fA(y) := \sum_{x\in S} f(x)A(x, y) \quad (y \in S).$$

We let $A^\dagger(x, y) := A(y, x)$ $(x, y \in S)$ denote the *transpose* of A. Then $fA = A^\dagger f$ and

$$\langle f, Ag\rangle = \langle fA, g\rangle = \langle A^\dagger f, g\rangle \qquad (f, g \in \mathbb{R}^S).$$

We say that a function $f \in \mathbb{R}^S$ is *nonnegative* if $f(x) \geq 0$ for all $x \in S$ and we say that f is *positive* if $f(x) > 0$ for all $x \in S$. A matrix A is *nonnegative* if $A(x, y) \geq 0$ for all $x, y \in S$. A *probability law* on S is a nonnegative function μ on S such that $\sum_{x\in S} \mu(x) = 1$. A *probability kernel* on S is a nonnegative matrix P such that $\sum_{y\in S} P(x, y) = 1$ for all $x \in S$.

Let μ be a probability law and let P be a probability kernel. 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

$$\mathbb{P}[X_0 = x_0, \ldots, X_n = x_n] = \mu(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n)$$

$(n \geq 1,\ x_0, \ldots, x_n \in S)$. We observe that for any function $f \in \mathbb{R}^S$,

$$\mathbb{E}\big[f(X_n)\big] = \langle \mu P^n, f\rangle$$

and μP^n is the law of X_n.

For any nonnegative matrix A, we write $x \overset{A}{\rightsquigarrow} y$ if there exist $n \geq 0$ such that $A^n(x,y) > 0$ or equivalently, there exist $x = x_0, \ldots, x_n = y$ such that $A(x_{k-1}, x_k) > 0$ for each $k = 1, \ldots, n$. A nonnegative matrix A is called *irreducible* if $x \overset{A}{\rightsquigarrow} y$ for all $x, y \in S$. For a give nonnegative matrix A, the *period* of a state $x \in S$ is the greatest common divisor of the set $\{n \geq 1 : A^n(x,x) > 0\}$. If $x \overset{A}{\leftrightsquigarrow} y$, then x and y have the same period. If A is irreducible and all states have period one, then we say that A is *aperiodic*.

An *invariant law* of a probability kernel P is a probability law ν on S such that $\nu P = \nu$. Equivalently, ν is invariant if the Markov chain $(X_k)_{k\geq 0}$ with transition kernel P and initial law ν is *stationary*, i.e. $(X_k)_{k\geq 0}$ is equal in law to $(Y_k)_{k\geq 0}$ defined as $Y_k := X_{k+1}$ $(k \geq 0)$. Basic results of Markov chain theory tell us that an irreducible Markov chain with a finite state space S has a unique invariant law ν, which has the property that $\nu(x) > 0$ for all $x \in S$. If P is moreover aperiodic, then μP^n converges to ν as $n \to \infty$, for each initial law μ.

For any probability law π on S and probability kernel P, we let $\pi * P$ denote the probability law on S^2 defined as

$$\pi * P(x,y) := \pi(x)P(x,y) \qquad (x, y \in S).$$

The following exercise links this notation to the space $\mathcal{V}(S)$ defined in Section 2.3.

Exercise 2.15 (Stationary laws) Show that a probability measure $\nu \in \mathcal{M}_1(S^2)$ satisfies $\nu \in \mathcal{V}(S)$ if and only if there exist a probability law π and probability kernel P on S such that $\nu = \pi * P$ and π is an invariant law of the Markov chain with transition kernel P.

The following theorem generalizes Theorem 2.7 to Markov chains. The study of large deviations of Markov processes was initiated by Donsker and Varadhan [DV75a, DV75b, DV76]. For this reason, it is sometimes called *Donsker-Varadhan theory*.

Theorem 2.16 (LDP for Markov chains) *Let $X = (X_k)_{k\geq 0}$ be a Markov chain with finite state space S, transition kernel P, and arbitrary initial law. Assume that $\mathbb{P}[X_0 = x] > 0$ and $P(x,y) > 0$ for all $x, y \in S$. 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 $I_P^{(2)}$ given by*

$$I_P^{(2)}(\nu) := \begin{cases} H(\nu|\nu^- * P) & \text{if } \nu \in \mathcal{V}(S), \\ \infty & \text{otherwise.} \end{cases}$$

Proof Let μ be the law of X_0 and let $\hat{X} = (\hat{X}_k)_{k\geq 0}$ be an i.i.d. collection of random variables with common law μ. For any $x = (x_k)_{k\geq 0} \in S^{\mathbb{N}}$, let us write

$$M_n^{(2)}(x) := \sum_{k=1}^{n} \delta_{(x_{k-1},x_k)} \qquad (n \geq 1),$$

so that in particular, $M_n^{(2)}(X)$ and $M_n^{(2)}(\hat{X})$ are the pair empirical distributions of X and $\hat{X}$, respectively. We observe that

$$\begin{aligned}\mathbb{P}[X_0 = x_0, \ldots, X_n = x_n] &= \mu(x_0)\, e^{\sum_{k=1}^{n} \log P(x_{k-1},x_k)} \\ &= \mu(x_0)\, e^{n \sum_{(y_1,y_2)\in S^2} \log P(y_1,y_2) M_n^{(2)}(x)(y_1,y_2)},\end{aligned}$$

while

$$\begin{aligned}\mathbb{P}[\hat{X}_0 = x_0, \ldots, \hat{X}_n = x_n] &= \mu(x_0)\, e^{\sum_{k=1}^{n} \log \mu(x_k)} \\ &= \mu(x_0)\, e^{n \sum_{(y_1,y_2)\in S^2} \log \mu(y_2) M_n^{(2)}(x)(y_1,y_2)}.\end{aligned}$$

It follows that for any $x = (x_k)_{k\geq 0} \in S^{\mathbb{N}}$,

$$\begin{aligned}&\frac{\mathbb{P}[X_0 = x_0, \ldots, X_n = x_n]}{\mathbb{P}[\hat{X}_0 = x_0, \ldots, \hat{X}_n = x_n]} \\ &\qquad = e^{n \sum_{(y_1,y_2)\in S^2} (\log P(y_1,y_2) - \log \mu(y_2)) M_n^{(2)}(x)(y_1,y_2)},\end{aligned}$$

and as a consequence, for each $\nu \in \mathcal{M}_1(S^2)$,

$$\frac{\mathbb{P}[M_n^{(2)}(X) = \nu]}{\mathbb{P}[M_n^{(2)}(\hat{X}) = \nu]} = e^{n \sum_{(y_1,y_2)\in S^2} (\log P(y_1,y_2) - \log \mu(y_2)) \nu(y_1,y_2)}.$$

We define $F : \mathcal{M}_1(S^2) \to \mathbb{R}$ by

$$F(\nu) := \sum_{(y_1,y_2)\in S^2} (\log P(y_1,y_2) - \log \mu(y_2)) \nu(y_1,y_2).$$

Note that F is continuous since we are assuming that $\mu(y_2) > 0$ and $P(y_1,y_2) > 0$ for all $y_1, y_2 \in S$. Since $\mathcal{M}_1(S^2)$ is compact, it follows that F is also bounded. Let

$$\rho_n^{(2)} := \mathbb{P}\big[M_n^{(2)}(X) \in \cdot\,\big] \quad \text{and} \quad \hat{\rho}_n^{(2)} := \mathbb{P}\big[M_n^{(2)}(\hat{X}) \in \cdot\,\big]$$

denote the laws of $M_n^{(2)}(X)$ and $M_n^{(2)}(\hat{X})$. Then we have just shown that

$$\rho_n(\nu) = e^{nF(\nu)}\hat{\rho}_n(\nu) \qquad \big(\nu \in \mathcal{M}_1(S^2)\big).$$

By Theorem 2.7, the laws $\mathbb{P}[M_n^{(2)}(\hat{X}) \in \cdot\,]$ satisfy the large deviation principle with speed n and rate function $\hat{I}_\mu^{(2)}$ given by

$$I_\mu^{(2)}(\nu) = \begin{cases} H(\nu|\nu^- \otimes \mu) & \text{if } \nu \in \mathcal{V}(S), \\ \infty & \text{otherwise.} \end{cases}$$

Applying Lemma 1.17 to the function F, we find that the laws ρ_n satisfy the large deviation principle with speed n and rate function $I^{(2)} = I_\mu^{(2)} - F$. Since

$$\begin{aligned} &H(\nu|\nu^- \otimes \mu) - F(\nu) \\ &\quad = \sum_{(y_1,y_2)\in S^2} \nu(y_1,y_2)\big(\log \frac{\nu(y_1,y_2)}{\nu^-(y_1)\mu(y_2)} + \log \mu(y_2) - \log P(y_1,y_2)\big) \\ &\quad = \sum_{(y_1,y_2)\in S^2} \nu(y_1,y_2) \log \frac{\nu(y_1,y_2)}{\nu^-(y_1)P(y_1,y_2)} = H(\nu|\nu^- * P), \end{aligned}$$

this proves the theorem. ∎

Exercise 2.17 (Arbitrary initial laws) Generalize Theorem 2.16 by dropping the condition that $\mathbb{P}[X_0 = x] > 0$ for all $x \in S$. Hint: Let μ be any probability law on S such that $\mu(x) > 0$ for all $x \in S$ and let μ' be the law of X_0. Let $(\tilde{X}_k)_{k\geq 0}$ be independent random variables such that $\tilde{X}_0$ has law μ' and $\tilde{X}_k$ has law μ for all $k \geq 1$. Combine Proposition 1.19 with Theorem 2.7 to prove that the pair empirical measures $M_n^{(2)}(\tilde{X})$ satisfy an LDP and then follow the proof of Theorem 2.16 with $\hat{X}$ replaced by $\tilde{X}$.

Remark Our proof of Theorem 2.16 closely follows [Hol00, Thm IV.3]. In [Hol00, Comment (4) from Section IV.3], it is 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'. It is indeed possible to prove this more general statement by adapting the proof of Theorem 2.7 but the proof gets a lot more messy because periodic boundary conditions do not work well anymore in the more general setting. In Theorem **??** below, we will use the Gärtner-Ellis theorem to prove a large deviation principle for irreducible Markov chains.

2.5 Cramér's moment generating function

In this section we start preparing for the proof of Cramér's theorem (Theorem 0.1) by studying the moment generating function Z defined in (0.1) and its logarithm, the so-called free energy. For any probability measure μ on $\mathbb{R}$ which has at least a finite first, respectively second moment, we let

$$\begin{aligned}\langle\mu\rangle &:= \int \mu(\mathrm{d}x)x,\\ \mathrm{Var}(\mu) &:= \int \mu(\mathrm{d}x)x^2 - \Big(\int \mu(\mathrm{d}x)x(i)\Big)^2\end{aligned}$$

denote the *mean* and *variance* of μ.

Lemma 2.18 (Smoothness of the free energy) *Let μ be a probability measure on $\mathbb{R}$ and let Z be given by*

$$Z(\lambda) := \int e^{\lambda x}\mu(\mathrm{d}x) \qquad (\lambda \in \mathbb{R}^d). \tag{2.9}$$

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

$$\mu_\lambda(\mathrm{d}x) := \frac{1}{Z(\lambda)}e^{\lambda x}\mu(\mathrm{d}x) \qquad (\lambda \in \mathbb{R}). \tag{2.10}$$

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

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

Proof We claim that $\lambda \mapsto Z(\lambda)$ is infinitely differentiable and

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

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

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

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

$$\begin{aligned} &\text{(i)} \quad \tfrac{\partial}{\partial\lambda}\log Z(\lambda) = \tfrac{\partial}{\partial\lambda}\log\int e^{\lambda x}\mu(\mathrm{d}x) = \frac{\int xe^{\lambda x}\mu(\mathrm{d}x)}{\int e^{\lambda x}\mu(\mathrm{d}x)} = \langle\mu_\lambda\rangle, \\ &\text{(ii)} \quad \tfrac{\partial^2}{\partial\lambda^2}\log Z(\lambda) = \frac{Z(\lambda)\int x^2e^{\lambda x}\mu(\mathrm{d}x) - (\int xe^{\lambda x}\mu(\mathrm{d}x))^2}{Z(\lambda)^2} \\ &\qquad\qquad = \int x^2\mu_\lambda(\mathrm{d}x) - \Big(\int x\mu_\lambda(\mathrm{d}x)\Big)^2 = \mathrm{Var}(\mu_\lambda). \end{aligned} \tag{2.11}$$

■

Exercise 2.19 (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 2.18. 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 $y_- := \inf(\mathrm{support}(\mu))$, $y_+ := \sup(\mathrm{support}(\mu))$.

2.6 Cramér's theorem for simple variables

Let us say that a real random variable X is *simple* if there exists a finite subset $R \subset \mathbb{R}$ such that $X \in R$ a.s. In this section, we prove Cramér's theorem for simple random variables. This is of course much weaker than Theorem 0.1, but the proof is instructive and a good warm-up for the theory that will follow.

Theorem 2.20 (Simple version of Cramér's theorem) *Let $(X_k)_{k\geq1}$ be i.i.d. random variables taking values in a finite subset $R \subset \mathbb{R}$. Then the probability measures*

$$\eta_n := \mathbb{P}\big[\frac{1}{n}\sum_{k=1}^n X_k \in \cdot\,\big] \qquad (n \geq 1)$$

satisfy the large deviation principle with speed n and good rate function I given by

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

where $Z(\lambda) := \mathbb{E}[e^{\lambda X_1}]$ $(\lambda \in \mathbb{R})$.

To prepare for the proof of Theorem 2.20 we introduce some notation. Let S be a finite set. For functions $\phi : S \to \mathbb{R}$ and probability measures $\nu, \mu \in \mathcal{M}_1(S)$, we introduce the notation

$$\langle \nu, \phi \rangle := \sum_{x \in S} \nu(x)\phi(x) \quad \text{and} \quad \Gamma_\mu(\phi) := \log \sum_{x \in S} \mu(x) e^{\phi(x)}.$$

For each $\phi : S \to \mathbb{R}$ and $\mu \in \mathcal{M}_1(S)$, we define a probability law $\mu_\phi \in \mathcal{M}_1(S)$ on S by

$$\mu_\phi(x) := \frac{1}{Z(\phi)} \mu(x) e^{\phi(x)} \quad (x \in S) \quad \text{with} \quad Z(\phi) := \sum_{z \in S} \mu(z) e^{\phi(z)}. \tag{2.12}$$

Lemma 2.21 (Duality relation) *For each $\nu, \mu \in \mathcal{M}_1(S)$ and $\phi : S \to \mathbb{R}$, one has*

$$\langle \nu, \phi \rangle \leq H(\nu|\mu) + \Gamma_\mu(\phi), \tag{2.13}$$

with equality if and only if $\nu = \mu_\phi$.

Proof We trivially have a strict inequality if $H(\nu|\mu) = \infty$ so without loss of generality we may assume that $\nu \ll \mu$. Reducing the set S if necessary, we can also without loss of generality assume that $\mu(x) > 0$ for all $x \in S$. Let $S' := \{x \in S : \nu(x) > 0\}$. Since $r \mapsto \log(r)$ is a strictly concave function, Jensen's inequality gives

$$\begin{aligned}
\langle \nu, \phi \rangle - H(\nu|\mu) &= \sum_{x \in S} \nu(x) \Big(\log \big(e^{\phi(x)} \big) - \log \big(\frac{\nu(x)}{\mu(x)} \big) \Big) \\
&= \sum_{x \in S'} \nu(x) \log \Big(e^{\phi(x)} \frac{\mu(x)}{\nu(x)} \Big) \leq \log \Big(\sum_{x \in S'} \nu(x) e^{\phi(x)} \frac{\mu(x)}{\nu(x)} \Big) \\
&\leq \log \Big(\sum_{x \in S} \mu(x) e^{\phi(x)} \Big) = \Gamma_\mu(\phi).
\end{aligned}$$

This proves (2.13). 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 $e^{\phi(x)} \frac{\mu(x)}{\nu(x)}$ is constant on S'. Since the logarithm is a strictly increasing function and e^ϕ is strictly positive, the second inequality is an equality if and only if $\mu(x) = 0$ whenever $\nu(x) = 0$. Thus, we have equality in (5.7) if and only if

$$\nu(x) = \frac{1}{Z} e^{\phi(x)} \mu(x) \qquad (x \in S),$$

where Z is some constant. Since ν is a probability measure, we must have $Z = Z(\phi)$. ∎

Proof of Theorem 2.20 Let $(X_k)_{k\geq 1}$ be i.i.d. random variables with common law μ taking values in a finite set S, and let $\phi : S \to R \subset \mathbb{R}$ be a bijection. We will prove that the probability measures

$$\eta_n := \mathbb{P}\big[\frac{1}{n}\sum_{k=1}^{n} \phi(X_k) \in \cdot\,\big] \qquad (n \geq 1)$$

satisfy the large deviation principle with speed n and good rate function I given by

$$I(y) := \sup_{\lambda\in\mathbb{R}} \big[\lambda y - \Gamma_\mu(\lambda\phi)\big] \qquad (y \in \mathbb{R}). \tag{2.14}$$

In particular, applying this with $S = R \subset \mathbb{R}$ and ϕ the identity function, this then yields the statement of the theorem.

By Theorem 2.4, the laws ρ_n of the empirical distributions $M_n := n^{-1}\sum_{k=1}^{n} \delta_{X_k}$ satisfy the large deviation principle with speed n and good rate function $H(\,\cdot\,|\mu)$. The function $\psi : \mathcal{M}_1(S) \to \mathbb{R}$ defined as

$$\psi(\nu) := \langle \nu, \phi\rangle \qquad \big(\nu \in \mathcal{M}_1(S)\big)$$

is continuous, so by the contraction principle (Proposition 1.15), the measures $\eta_n = \rho_n \circ \psi^{-1}$ satisfy the large deviation principle with speed n and good rate function I' defined by

$$I'(y) := \inf\big\{H(\nu|\mu) : \nu \in \mathcal{M}_1(S),\ \langle \nu, \phi\rangle = y\big\} \qquad (y \in \mathbb{R}). \tag{2.15}$$

To complete the proof, we need to show that $I' = I$, the function defined in (2.14). To evaluate the infimum in (2.15), we use the method of Lagrange multipliers: we first try to find the minimum of the function $\nu \mapsto H(\nu|\mu) - \lambda\langle \nu, \phi\rangle$ for general $\lambda \in \mathbb{R}$, and then try to choose λ in such a way that the minimizer ν satisfies the constraint $\langle \nu, \phi\rangle = y$. Lemma 2.21 tells us that $H(\nu|\mu) - \langle \nu, \lambda\phi\rangle \geq -\Gamma_\mu(\lambda\phi)$, with equality if and only if $\nu = \mu_{\lambda\phi}$. In other words, for each $\lambda \in \mathbb{R}$, the function $\nu \mapsto H(\nu|\mu) - \lambda\langle \nu, \phi\rangle$ attains its minimal value in the unique point $\mu_{\lambda\phi}$, and the function value in this point is $-\Gamma_\mu(\lambda\phi)$.

Let $y_- := \min\{\phi(x) : \mu(x) > 0\}$ and $y_+ := \max\{\phi(x) : \mu(x) > 0\}$ be the minimal and maximal values that the random variables $\big(\phi(X_k)\big)_{k\geq 1}$ can obtain. By Lemma 2.18, the function $\lambda \mapsto \langle \mu_{\lambda\phi}, \phi\rangle$ is infinitely differentiable and strictly increasing. Using also Exercise 2.19, it follows that for each $y_- < y < y_+$, there

exists a unique $\lambda_\circ \in \mathbb{R}$ such that $\langle \mu_{\lambda_\circ \phi}, \phi\rangle = y$. The method of Lagrange multipliers then tells us that the function $\nu \mapsto H(\nu|\mu)$ attains its minimal value over the set $\{\nu \in \mathcal{M}_1(S) : \langle \nu, \phi\rangle = y\}$ in the unique point $\nu = \mu_{\lambda_\circ \phi}$, and in this point

$$H(\nu|\mu) = \lambda_\circ \langle \nu, \phi\rangle - \Gamma_\mu(\lambda_\circ \phi) = \lambda_\circ y - \Gamma_\mu(\lambda_\circ \phi).$$

In view of this, to prove that the functions in (2.14) and (2.15) satisfy $I(y) = I'(y)$, it suffices to show that the function $\lambda \mapsto \lambda y - \Gamma_\mu(\lambda\phi)$ attains its minimum in the point $\lambda_\circ$. Differentiating using Lemma 2.18 gives

$$\tfrac{\partial}{\partial \lambda}\big[\lambda y - \Gamma_\mu(\lambda\phi)\big] = y - \langle \mu_{\lambda\phi}, \phi\rangle.$$

Since $\lambda \mapsto \langle \mu_{\lambda\phi}, \phi\rangle$ is continuous and strictly increasing, the right-hand side of our formula is negative for $\lambda < \lambda_\circ$, zero for $\lambda = \lambda_\circ$, and positive for $\lambda > \lambda_\circ$. This completes the proof that $I(y) = I'(y)$ for $y_- < y < y_+$.

We next consider the case that $y = y_+$. By our assumption that ϕ is a bijection, there is a unique $x_+ \in S$ such that $\phi(x_+) = y_+$. Now the only $\nu \ll \mu$ such that $\langle \nu, \phi\rangle = y_+$ is given by $\nu = \delta_{x_+}$ and hence $I'(y_+) = H(\delta_{x+}|\mu) = -\log \mu(\{x_+\})$. By Lemma 2.18 and Exercise 2.19, the function $\lambda \mapsto \lambda y_+ - \Gamma_\mu(\lambda\phi)$ is strictly increasing, and hence

$$\begin{aligned} I(y_+) &= \lim_{\lambda\to\infty}\big[\lambda y_+ - \Gamma_\mu(\lambda\phi)\big] = -\lim_{\lambda\to\infty} \log\big(e^{-\lambda y_+}\Gamma_\mu(\lambda\phi)\big) \\ &= -\lim_{\lambda\to\infty} \log \sum_{x\in S} \mu(x)\, e^{\lambda(\phi(x) - \phi(x_+))} = -\log \mu(\{x_+\}). \end{aligned}$$

This proves that $I'(y_+) = I(y_+)$. By symmetry, also $I'(y_-) = I(y_-)$.

We finally treat the case that $y_+ > y$. In this case, there exist no $\nu \ll \mu$ such that $\langle \nu, \phi\rangle = y$ and hence we see from (2.15) that $I'(y) = \inf \emptyset = \infty$. By Lemma 2.18 and Exercise 2.19, $\frac{\partial}{\partial\lambda}\Gamma_\mu(\lambda\phi) \leq y_+$ for all $\lambda \in \mathbb{R}$, so we see from (2.14) that $I(y) = \lim_{\lambda\to\infty}[\lambda y_+ - \Gamma_\mu(\lambda\phi)] = \infty$. By the same argument $I(y) = \infty = I'(y)$ for $y < y_-$. ∎

2.7 Cramér's theorem

In the previous section, we gave a proof of Cramér's theorem that was based on the Boltzmann-Savov theorem and the contraction principle. A disadvantage of this approach is that we only obtained the result for simple random variables. In

the present section we give a direct proof of Cramér's theorem that does not have this disadvantage. Our proof makes use of Lemma 0.2. Although it is possible to prove Lemma 0.2 by elementary means, for convenience, we postpone the proof till Section 5.1 when we have the tools available to give a short and elegant proof.

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

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

If $y > \rho$, then, by Lemma 2.18, $\frac{\partial}{\partial\lambda}[\log Z(\lambda) - \lambda y]|_{\lambda=0} = \rho - y < 0$, so, by the convexity of the function $\lambda \mapsto [\log Z(\lambda) - \lambda y]$,

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

Together with our previous formula, this shows that

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

and hence, in particular,

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

To estimate the limit inferior from below, we distinguish three cases. If $y > y_+$, then $\mathbb{P}[T_n \ge y] = 0$ for all $n \ge 1$ while $I(y) = \infty$ by Lemma 0.2 (v), so (0.2) (i) is trivially fulfilled. If $y = y_+$, then $\mathbb{P}[T_n \ge y] = \mathbb{P}[X_1 = y_+]^n$ 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 differentiating using Lemma 2.18, we see that the function $\lambda \mapsto [y\lambda - \log Z(\lambda)]$ assumes its maximum in the point $\lambda_\circ$ that is uniquely characterized by the condition $\langle\mu_{\lambda_\circ}\rangle = y$. We observe that if $(\hat{X}_k)_{k\ge1}$ 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 $\lim_{n\to\infty}\mathbb{P}[\hat{T}_n \ge y] = \frac{1}{2}$ by the central limit theorem and therefore $\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}[\hat{T}_n \ge y] = 0$. 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

$$\begin{aligned}\mathbb{P}\big[T_n \geq y\big] &= \mathbb{P}\big[\sum_{k=1}^n (X_k - y) \geq 0\big] = \int \mu(\mathrm{d}x_1) \cdots \int \mu(\mathrm{d}x_n) 1_{\{\sum_{k=1}^n (x_k - y) \geq 0\}} \\ &= Z(\lambda_\circ)^n \int e^{-\lambda_\circ x_1} \mu_{\lambda_\circ}(\mathrm{d}x_1) \cdots \int e^{-\lambda_\circ x_n} \mu_{\lambda_\circ}(\mathrm{d}x_n) 1_{\{\sum_{k=1}^n (x_k - y) \geq 0\}} \\ &= Z(\lambda_\circ)^n e^{-n\lambda_\circ y} \int \mu_{\lambda_\circ}(\mathrm{d}x_1) \cdots \int \mu_{\lambda_\circ}(\mathrm{d}x_n) \\ &\qquad\qquad \times e^{-\lambda_\circ \sum_{k=1}^n (x_k - y)} 1_{\{\sum_{k=1}^n (x_k - y) \geq 0\}} \\ &= e^{-nI(y)} \mathbb{E}\big[e^{-n\lambda_\circ (\hat{T}_n - y)} 1_{\{\hat{T}_n - y \geq 0\}}\big]. \end{aligned} \tag{2.16}$$

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 (2.16) 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 (2.16) 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 5.4. 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].

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 2.18 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 $\mu_{\lambda_\circ}$.

2.8 Excercises

Exercise 2.22 (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 $x = 1, \dots, 6$ in these n throws. Let μ_0 be the uniform distribution on $S := \{1, \dots, 6\}$ 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 2.23 (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

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}[M_n = k_n] = -\kappa H(\nu_1|\mu) - (1-\kappa)H(\nu_2|\mu). \tag{2.17}$$

Sketch a proof, similar to the proof of Theorem 2.4, that the laws $\mathbb{P}[M_n \in \cdot\,]$ satisfy a large deviation principle with speed n and rate function I given by

$$I(\kappa) := \kappa H(\nu_1|\mu) + (1-\kappa)H(\nu_2|\mu).$$

Hint: use Stirling's formula to show that

$$\frac{1}{n} \log \binom{n}{m} \approx H(\frac{m}{n}),$$

where

$$H(z) := -z \log z - (1-z) \log(1-z).$$

Exercise 2.24 (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

$$H(\nu * Q|\mu * P) = H(\nu|\mu) + \sum_{x_1 \in S} \nu(x_1) H(Q_{x_1}|P_{x_1}),$$

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 $\nu = \nu^- * Q$, then

$$H(\nu|\nu^- * P) = \sum_{x_1 \in S} \nu^-(x_1) H(Q_{x_1}|P_{x_1}).$$

Exercise 2.25 (Minimizer of the rate function) Let P be irreducible. Show that the unique minimizer of the function $\mathcal{V}(S) \ni \nu \mapsto H(\nu|\nu^- * P)$ is given by $\nu = \mu * P$, where μ is the invariant law of P.

The following exercise asks you to prove that the function

$$\mathcal{M}_1(S^2) \ni \nu \mapsto H(\nu|\nu^- * P)$$

is convex.

Exercise 2.26 (Convexity of rate function) Let P be a probability kernel on S. Let $p_1, \dots, p_n$ be nonnegative constants summing up to one and let $\nu := \sum_{k=1}^n p_k \nu_k$ with $\nu_k \in \mathcal{V}(S)$. Let Q^k and Q be probability kernels on S such that $\nu = \nu^- * Q$ and $\nu_k = \nu_k^- * Q^k$ $(1 \le k \le n)$. Prove that

$$H(\nu|\nu^- * P) = \sum_{k=1}^n p_k H(\nu_k|\nu^- * P) - \sum_{k=1}^n p_k \sum_{x \in S} \nu_k^-(x) H(Q_x^k|Q_x),$$

where Q_x is the probability law on S defined as $Q_x(y) := Q(x,y)$ $(x, y \in S)$ and Q_x^k is defined similarly with Q replaced by Q^k.

Exercise 2.27 (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

$$[0,1] \ni p \mapsto H(\nu_p|\nu_p^- * P)$$

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 $\mathcal{M}_1(S^2) \ni \nu \mapsto H(\nu|\nu^- * P)$ is not strictly convex. Do you see a general pattern how to create such examples? Hint: You can use Excercise 2.24 or Excercise 2.26.

Exercise 2.28 (Strong law of large numbers) Let $X = (X_k)_{k\geq 0}$ be a Markov chain with finite state space S, transition kernel P, and arbitrary initial law. Assume that $\mathbb{P}[X_0 = x] > 0$ and $P(x, y) > 0$ for all $x, y \in S$. Let $(M_n^{(2)})_{n\geq 1}$ be the pair empirical distributions of X. Show that

$$M_n^{(2)} \underset{n\to\infty}{\longrightarrow} \pi * P \quad \text{a.s.}, \tag{2.18}$$

where π denotes the invariant law of P. Hint: use Theorem 2.16, Exercise 2.25, and Borel-Cantelli.

Exercise 2.29 (Approximation lemma) Use Exercises 2.15 and 2.28 to give an alternative proof of Lemma 2.12. Hint: first prove the claim under the additional assumption that $\nu(x, y) > 0$ for all $(x, y) \in S^2$.

Chapter 3

Exponential tightness

3.1 Tightness

In Sections 1.1 and 1.2, we have stressed the similarity between weak convergence of measures and large deviation principles. In 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 3.7), 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 3.1 (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 3.1 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 3.2 (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 3.2 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 3.3 (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 3.2 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 3.3. 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 3.3 we conclude that the original sequence (μ_n) converges weakly to μ. ∎

3.2 LDP's on compact spaces

Our aim is to prove an analogue of Lemma 3.2 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.[1] Below, $|f|$ denotes the absolute value of a function, i.e., the function $x \mapsto |f(x)|$.

Proposition 3.4 (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

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

(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

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

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 $B_\delta(y)^{\rm c}$, 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 $B_{2\delta}(y)^{\rm c}$. 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 $f \geq f_1 \geq f_2 \geq \cdots$ 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 34. The following proposition prepares for Theorem 3.7 below.

Proposition 3.5 (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

$$|\Lambda(f) - \Lambda(g)| \leq \|f-g\|_\infty \tag{3.1}$$

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 (3.1) that Λ can be uniquely extended to a function $\Lambda : \mathcal{C}(E) \to [0,\infty)$ such that (3.1) 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 3.4. 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 (1.6), 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 (1.4). 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 3.4, 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. ∎

3.3 Exponential tightness

We wish to generalize Proposition 3.5 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 3.6 (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), $\|1_{O^{\mathrm{c}}_{\delta,m}}\|_{\infty,I} \downarrow 0$. Thus, for each $\varepsilon, \delta > 0$ we can choose an $m \geq 1$ such that

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

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 3.5 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 3.7 (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 3.7 from Proposition 3.5 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 3.8 (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 3.9 (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.

Exercise 3.10 (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.,

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

Let $(\mu_n)_{n\geq 1}$ 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[2] of $\mathcal{M}_1(\overline{\mathbb{R}})$,

[2]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 3.8 proves that $\mathcal{M}_1(\mathbb{R})$ is not locally compact. On the other hand, since by Excercise 3.10, $\mathcal{M}_1(\mathbb{R})$ is Polish in the induced topology, we can conclude by Proposition 3.9 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 3.7 Let $\overline{E}$ be a metrizable compactification of E. By Proposition 3.5, 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

$$\limsup_{m\to\infty} \frac{1}{s_{n(m)}} \log \mu_{n(m)}(K^{\mathrm{c}}) \leq -b. \tag{3.2}$$

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 (3.2). 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 1.16) 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 3.7. As an appetizer, we conclude the present section by proving two simple lemmas. The argument used in the proof of Lemma 3.3 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 3.11 (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 1.13 and 1.17 to unbounded functions.

Lemma 3.12 (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*

$$\lim_{n\to\infty} \frac{1}{s_n} \log \int e^{s_n F} \mathrm{d}\mu_n = \sup_{x\in E} \big[F(x) - I(x)\big]. \tag{3.3}$$

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 3.11, 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 3.7, (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

$$\sup_{x\in E}\big[G(x) + F(x) - I(x)\big] = \lim_{n\to\infty} \frac{1}{s_n''} \int_E e^{s_n''(G(x)+F(x))}\mu_n''(\mathrm{d}x)$$
$$= \lim_{n\to\infty} \frac{1}{s_n''} \int_E e^{s_n'' G(x)}\nu_n''(\mathrm{d}x) = \sup_{x\in E}\big[G(x) - J(x)\big].$$

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 (3.3). ∎

3.4 Applications of exponential tightness

In this section, we look at some applications of Theorem 3.7. 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 3.6 and Theorem 3.7, we obtain the following analogue of Lemma 3.2. Note that by Lemma 3.6, the conditions (i) and (ii) below are clearly necessary for the measures μ_n to satisfy a large deviation principle.

Proposition 3.13 (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 3.7, 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 3.11, 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 3.13 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 3.13, 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 3.14 (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 3.4. ∎

Proposition 3.13 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

$$\inf_{x\in\overline{A}} I(x) \leq \inf_{x\in\mathrm{int}(A)} J(x) \quad \forall A \in \mathcal{A} \tag{3.4}$$

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 3.15 (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 (3.4) 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 3.16 (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) $\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) \leq -\inf_{x\in\overline{A}} I(x) \quad \forall A \in \mathcal{A}_{\mathrm{up}}$,

(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 3.6. To prove sufficiency, we use Lemma 3.11. By Theorem 3.7, 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 3.11, 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 3.16, instead of assuming that $\mathcal{A}_{\mathrm{low}}$ is rate function determining, it suffices to assume that

$$\forall \varepsilon > 0 \text{ and } z \in E \text{ s.t. } I(z) < \infty,\ \exists A \in \mathcal{A}_{\mathrm{low}} \text{ s.t. } z \in A \subset B_\varepsilon(z). \tag{3.5}$$

Indeed, the proof of Lemma 3.15 shows that if (3.4) holds with I and J interchanged, and we moreover have (3.5), 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 3.16, 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 3.15 this does not add anything new. However, if E is not locally compact, then $\{\text{int}(K) : K \text{ compact}\}$ 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 3.17 (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 1.15 below) may fail for measures satisfying a weak large deviation principle.

Proof of Lemma 3.17 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 (1.5), it follows that, for any closed $C \subset E$,

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

∎

3.5 Approximation of LDPs

In this section we prove two results (Proposition 3.19 and Theorem 3.21 below) that can be used to derive "difficult" large deviation principles by approximation with simpler large deviation principles.

Lemma 3.18 (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

$$\begin{aligned}&\limsup_{m\to\infty} \bigl| \|f_i\|_{s_{n'(m)},\mu_{m,n'(m)}} - \|f_i\|_{\infty,I} \bigr| \\ &\quad \leq \limsup_{m\to\infty} \bigl| \|f_i\|_{s_{n'(m)},\mu_{m,n'(m)}} - \|f_i\|_{\infty,I_m} \bigr| + \limsup_{m\to\infty} \bigl| \|f_i\|_{\infty,I_m} - \|f_i\|_{\infty,I} \bigr| = 0\end{aligned}$$

for all $i \geq 1$. By Lemma 3.14 (b), the functions $|f_i|$ are rate function determining, hence by Proposition 3.13, 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 1.16), they also satisfy the large deviation principle on E. ∎

Proposition 3.19 (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*

$$\lim_{m\to\infty} \|f\|_{\infty,I_m} = \|f\|_{\infty,I} \qquad \bigl(f \in \mathcal{C}_{b,+}(E)\bigr), \tag{3.6}$$

and that there exists a metric d generating the topology on E such that for each $n(m) \to \infty$,

$$\lim_{m\to\infty} \frac{1}{s_{n(m)}} \log \mathbb{P}[d(X_{n(m)}, X_{m,n(m)}) \geq \varepsilon] = -\infty \qquad (\varepsilon > 0), \tag{3.7}$$

i.e., $X_{n(m)}$ and $X_{m,n(m)}$ are exponentially close in the sense of (1.9). 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 3.13, 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 (3.6) and Lemma 3.18, 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 (3.7), the random variables $X_{n'(m)}$ and $X_{m,n'(m)}$ are exponentially close in the sense of Proposition 1.19, 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 (3.6).

Lemma 3.20 (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 (3.6).*

Proof Formula (3.6) 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. ∎

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 3.21 (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 $f \in \mathcal{C}_{b,+}(F^m)$,

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

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 3.14 (a), it follows that $\mathcal{D}$ is rate function determining.

Proposition 3.13 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

$$\|f \circ \vec{\psi}_m\|_{\infty,I} = \|f\|_{\infty,I_m} \qquad \big(m \geq 1,\ f \in \mathcal{C}_{b,+}(F^m)\big). \tag{3.8}$$

Set

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

which by the contraction principle (Proposition 1.15) 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 (3.8) 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 3.21 in terms of the rate functions $\vec{\psi}_m$.

Lemma 3.22 (Formula for high-level rate function) *In the set-up of Theorem 3.21,*

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

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 $\uparrow \infty$. ∎

3.6 Continuous time Markov chains

In the present section we give a first application of the abstract results proved in this chapter, by using Proposition 3.19 about approximation of LDP's to derive a large deviations result for continuous time Markov chains. We will see further applications in the coming chapters. In particular, in the proof of the Gärtner-Ellis theorem in Section 4.9 we will apply Theorem 3.7, and in the proof of Sanov's theorem in Section 5.4 we will apply Theorem 3.21 about projective limits.

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 2.16. 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 3.23 (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 3.24 (Properties of the rate function) Show that the function I from Theorem 3.23 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 $w(x, y) = \rho(x)r(x, y)$ $(x, y \in S)$.

Proof of Theorem 3.23 Our strategy is to derive Theorem 3.23 from Theorem 2.16 using approximation techniques from Section 3.5. We set

$$M_T^\varepsilon(x) := \frac{1}{\lfloor T/\varepsilon \rfloor} \sum_{k=1}^{\lfloor T/\varepsilon \rfloor} 1_{\{(X_{\varepsilon(k-1)}, X_{\varepsilon k}) = (x,x)\}} \qquad (x \in S),$$
$$W_T^\varepsilon(x, y) := \frac{1}{\varepsilon \lfloor T/\varepsilon \rfloor} \sum_{k=1}^{\lfloor T/\varepsilon \rfloor} 1_{\{(X_{\varepsilon(k-1)}, X_{\varepsilon k}) = (x,y)\}} \qquad (x, y \in S,\ x \neq y),$$

and we let $W_T^\varepsilon(x, x) := 0$ $(x \in S)$. By Proposition 3.19, 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 3.23 is a good rate function and the rate functions I_ε converge to I in the sense of (3.6) 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

$$X_k^\varepsilon := X_{\varepsilon k} \qquad (k \geq 0),$$

Let P_ε denote its transition kernel, and let $M_n^{(2)\,\varepsilon}$ be its empirical pair distributions. Then

$$M_T^\varepsilon(x) = M_{\lfloor T/\varepsilon \rfloor}^{(2)\,\varepsilon}(x, x) \qquad (x \in S),$$
$$W_T^\varepsilon(x, y) = \varepsilon^{-1} M_{\lfloor T/\varepsilon \rfloor}^{(2)\,\varepsilon}(x, y) \qquad (x, y \in S,\ x \neq y).$$

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

$$\begin{aligned} \rho_\varepsilon(\nu)(x) &:= \nu(x,x) && (x \in S),\\ w_\varepsilon(\nu)(x) &:= 1_{\{x\neq y\}}\varepsilon^{-1}\nu(x,y) && (x,y \in S). \end{aligned}$$

Using the formula $P_\varepsilon = \sum_{n=0}^\infty \frac{1}{n!} G^n \varepsilon^n$ and the fact that the transition rates are irreducible, it is easy to see that $P_\varepsilon(x,y) > 0$ for all $x, y \in S$. It follows that Theorem 2.16 is applicable (using also Exercise 2.17 to allow general initial laws). We conclude that 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

$$I_\varepsilon(\rho_\varepsilon(\nu), w_\varepsilon(\nu)) := \varepsilon^{-1} H(\nu|\nu^1 \otimes P_\varepsilon) \qquad (\nu \in \mathcal{V}), \tag{3.9}$$

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 3.20, 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

$$\begin{aligned} \nu_n(x,x) &\underset{n\to\infty}{\longrightarrow} \rho(x) && (x \in S), \\ \varepsilon_n^{-1} 1_{\{x\neq y\}} \nu_n(x,y) &\underset{n\to\infty}{\longrightarrow} w(x,y) && (x,y \in S^2), \end{aligned} \tag{3.10}$$

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 (3.10) 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

$$\begin{aligned} \sum_{x\in S} \mu(x) \psi\left(\frac{\nu(x)}{\mu(x)}\right) &= \sum_{x\in S} \mu(x) \big[1 - \frac{\nu(x)}{\mu(x)} + \frac{\nu(x)}{\mu(x)} \log\Big(\frac{\nu(x)}{\mu(x)}\Big)\big] \\ &= \sum_{x\in S} [\mu(x) - \nu(x)] + \sum_{x\in S} \nu(x) \log\Big(\frac{\nu(x)}{\mu(x)}\Big) = H(\nu|\mu), \end{aligned}$$

where we use the convention that $0 \log 0 := 0$. By Excercise 2.24, 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 (3.10), 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

$$\begin{aligned}\varepsilon_n^{-1} H(\nu_n|\nu_n^1 \otimes P_{\varepsilon_n}) &= \varepsilon_n^{-1}\sum_{x,y}\nu_n^1(x)P_{\varepsilon_n}(x,y)\psi\Big(\frac{\nu_n(x,y)}{\nu_n^1(x)P_{\varepsilon_n}(x,y)}\Big)\\ &= \varepsilon_n^{-1}\sum_x \big(\rho(x) + O(\varepsilon_n)\big)\psi\Big(\frac{\rho(x)+O(\varepsilon_n)}{\rho(x)+O(\varepsilon_n)}\Big)\\ &\quad + \sum_{x\neq y}\big(\rho(x)r(x,y) + O(\varepsilon_n)\big)\psi\Big(\frac{\varepsilon_n w(x,y) + O(\varepsilon_n^2)}{\varepsilon_n\rho(x)r(x,y) + O(\varepsilon_n^2)}\Big)\\ &\geq \sum_{x\neq y}\rho(x)r(x,y)\psi\Big(\frac{w(x,y)}{\rho(x)r(x,y)}\Big) + O(\varepsilon_n).\end{aligned} \tag{3.11}$$

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 (3.10) 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 (3.11), 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

$$(M_{T_m}, W_{T_m}) \quad \text{and} \quad (M_{\varepsilon_m\lfloor T_m/\varepsilon_m\rfloor}, W_{\varepsilon_m\lfloor T_m/\varepsilon_m\rfloor}) \tag{3.12}$$

are exponentially close. Therefore, by Excercise 3.27 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 $J_1(\varepsilon, T) \le |\Delta_T|$, 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 3.25 below, it follows that for any $0 < \varepsilon < \delta/R$,

$$\begin{aligned} &\limsup_{m \to \infty} \frac{1}{T_m} \log \mathbb{P}\big[\varepsilon_m J_1(\varepsilon_m, T_m)/T_m \ge \delta\big] \\ &\quad \le \limsup_{m \to \infty} \frac{1}{T_m} \log \mathbb{P}\big[\varepsilon N_{RT_m}/T_m \ge \delta\big] \le \psi(\delta/R\varepsilon) \underset{\varepsilon \to 0}{\longrightarrow} -\infty, \end{aligned}$$

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

$$\mathbb{P}\big[\big|\Delta_T \cap (\varepsilon(k-1), \varepsilon k]\big| \ge 2\big] \le 1 - e^{-R\varepsilon} - R\varepsilon e^{-R\varepsilon}.$$

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 $\mathbb{P}[N \ge 1] = \mathbb{P}[M \ge 2]$. 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 $(T/\varepsilon)\lambda_\varepsilon = \frac{1}{2}R^2T\varepsilon + O(\varepsilon^2)$. By the same argument as for J_1, we conclude that

$$\limsup_{m \to \infty} \frac{1}{T_m} \log \mathbb{P}\big[\varepsilon_m J_2(\varepsilon_m, T_m)/T_m \ge \delta\big] = -\infty.$$

∎

Exercise 3.25 (Large deviations for Poisson process) Let $N = (N_t)_{t \ge 0}$ 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 3.26 (Rounded times) Prove that the random variables in (3.12) are exponentially close.

Exercise 3.27 (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

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

Chapter 4

Convex analysis

4.1 Dual linear spaces

Large deviations theory is based on two pillars. The first pillar consists of the abstract theory of large deviation principles. We have covered this pillar in Chapters 1 and 3. The second pillar consists of convex analysis, and in particular the theory of the Legendre transform. This is what the present chapter is devoted to. At the end of the chapter, in Section 4.9, we will prove the Gärtner-Ellis theorem, which, as we will see in the final chapters of these lecture notes, is a very powerful tool for proving large deviation principles.

Let V be a finite dimensional real linear space. By definition, a *linear form* on V is a linear function $l : V \to \mathbb{R}$. The space V^* of all linear forms on V naturally has the structure of a real linear space. We call V^* the *dual* of V. It is well-known that V and V^* have the same dimension. Moreover, there is a natural isomorphism between the dual $(V^*)^*$ of V^* and the original space V. Indeed, each $x \in V$ defines a linear form $L_x : V^* \to \mathbb{R}$ by the formula $L_x(l) := l(x)$ $(l \in V^*)$, and each linear form on V^* is of this form. As a result, $x \mapsto L_x$ is a natural linear bijection from V to $(V^*)^*$ and we can (and will) for most purposes identify $(V^*)^*$ with V. Since V and V^* have the same dimension, there are also plenty of linear bijections from V to V^*. In general, however, there is no natural way to choose one particular linear bijection between these spaces, which is why we have to distinguish them.

To have notation that treats a space and its dual in a more symmetric way, we also say that two finite dimensional real linear spaces V and W are *dual to each*

other if there is defined a function

$$V \times W \ni (x, y) \mapsto \langle x, y \rangle \in \mathbb{R}$$

such that:

(i) $V^* = \{\langle \cdot\, , y \rangle : y \in W\}$,

(ii) $W^* = \{\langle x, \cdot \rangle : x \in V\}$,

where $\langle x, \cdot \rangle$ denotes the function $V \ni x \mapsto \langle x, y \rangle$ and likewise $\langle \cdot\, , y \rangle$ denotes the function $W \ni y \mapsto \langle x, y \rangle$. To make our notation even more symmetric, we sometimes write $\langle y, x \rangle$ instead of $\langle x, y \rangle$.

Let V be a finite dimensional real linear space, let V^* be its dual, and let $V^* \times V \ni (l, x) \mapsto \langle l, x \rangle \in \mathbb{R}$ be the function

$$\langle l, x \rangle := l(x) \qquad (x \in V,\ l \in V^*).$$

Then clearly V and V^* are dual to each other with respect to the function $\langle \cdot\, , \cdot \rangle$. We will often denote elements of V by $x, y, z, \ldots$ and elements of V^* by x^*, y^*, z^*. Here, the asterix just serves to remind us what space a vector belongs to. Thus, in using this notation, we regard x^* as a single symbol, and not (!) as a function of another vector called x.

If $\{e(1), \ldots, e(d)\}$ is a basis for V, then setting

$$\langle e^*(i), e(j) \rangle := \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \end{cases}$$

defines a basis $\{e^*(1), \ldots, e^*(d)\}$ of the dual space V^*. We call $\{e^*(1), \ldots, e^*(d)\}$ the *dual basis*. We can uniquely write elements $x \in V$ and $x^* \in V^*$ in terms of the basis and dual basis as

$$x = \sum_{i=1}^{d} x_i e(i) \quad \text{and} \quad x^* = \sum_{i=1}^{d} x_i^* e^*(i),$$

where $x_1, \ldots, x_d$ and $x_1^*, \ldots, x_d^*$ are real constants that are called the *coordinates* of x and x^* with respect to the bases $\{e^*(1), \ldots, e^*(d)\}$ and $\{e^*(1), \ldots, e^*(d)\}$. It follows immediately from our definition of the dual basis that

$$\langle x^*, x \rangle = \sum_{i=1}^{d} x_i^* x_i \qquad (x \in V,\ x^* \in V^*).$$

In other words, in terms of a basis and its dual basis, $\langle x^*, x\rangle$ takes the form of the usual inner product on $\mathbb{R}^d$.

For any linear subspace $W \subset V$, we define $W^\perp \subset V^*$ by

$$W^\perp := \big\{x^* \in V^* : \langle x^*, x\rangle = 0 \ \forall x \in W\big\}.$$

It is easy to see that $(W^\perp)^\perp = W$, i.e.,

$$W := \big\{x \in V : \langle x^*, x\rangle = 0 \ \forall x^* \in W^\perp\big\}. \tag{4.1}$$

4.2 Convex sets

Throughout this section, V is a finite dimensional real linear space and V^* is its dual. By definition, a set $C \subset V$ 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 $A \subset V$ 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 $C \subset V$ 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 $A \subset V$ is *affine* if $(1-p)x + py \in A$ for all $x, y \in A$ and $p \in \mathbb{R}$. The *affine hull* of a set $A \subset V$ 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 $A \subset V$ is of the form $A = \{x + y : y \in F\}$ where F is a linear subspace of V. 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 $C \subset V$ is the interior of C when viewed as a subset of its affine hull. Each nonempty convex set $C \subset V$ has a nonempty relative interior[1] and each closed convex set $C \subset V$ is the closure of its relative interior.

[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 $\{x\}$.

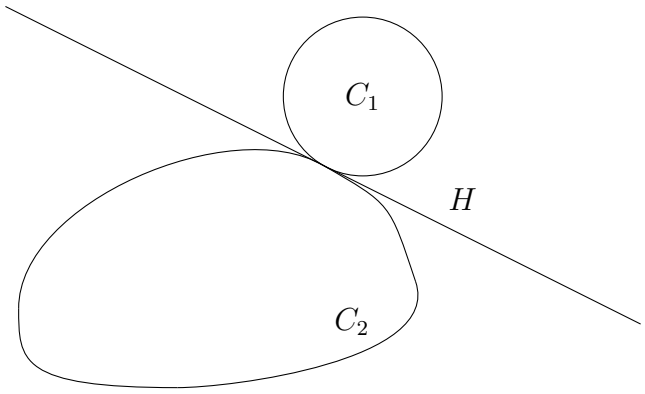

Figure 4.1: A hyperplane H separating the convex sets C_1 and C_2.

Recall that V^* is the dual space of V. Each $x^* \in V^* \backslash \{0\}$ and $c^* \in \mathbb{R}$ define two closed *half-spaces* in V by

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

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:

$$\overline{C}(A) = \bigcap \big\{H^{\leq}_{x^*,c^*} : x^* \in V^* \backslash \{0\},\ c^* \in \mathbb{R},\ A \subset H^{\leq}_{x^*,c^*}\big\}. \tag{4.2}$$

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 (4.2) is the following separation theorem, which we cite from [Roc70, Theorem 11.3]. See Figure 4.1 for an illustration.

Theorem 4.1 (Separating hyperplane *Let $C_1, C_2 \subset V$ be convex sets with relative interiors* $\mathrm{ri}(C_i)$ *$(i = 1, 2)$. Assume that* $\mathrm{ri}(C_1) \cap \mathrm{ri}(C_2) = \emptyset$*. Then there exists an $x^* \in V^* \backslash \{0\}$ and $c^* \in \mathbb{R}$ such that $C_1 \in H^{\leq}_{x^*,c^*}$ and $C_2 \in H^{\leq}_{x^*,c^*}$.*

The following lemma is a simple consequence of Theorem 4.1.

Lemma 4.2 (Supporting hyperplane) *Let $C \subset V$ 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 $x^* \in V^* \backslash \{0\}$ and $c^* \in \mathbb{R}$ such that*

$$C \subset H^{\leq}_{x^*,c^*} \quad \textit{and} \quad x \in H^{\geq}_{x^*,c^*}. \tag{4.3}$$

Proof Apply Theorem 4.1 to the convex sets C and $\{x\}$, using the fact that the relative interior of C is $\text{int}(C)$ and the relative interior of $\{x\}$ is $\{x\}$, and these are disjoint. ∎

If (4.3) holds, then we say that H_{x^*} is a *supporting hyperplane* at x.

4.3 Convex functions

We continue to assume that V is a finite dimensional real linear space and V^* is its dual. For any function $f : V \to (-\infty, \infty]$, we call

$$\mathcal{D}_f := \big\{x \in V : f(x) < \infty\big\} \quad \text{and} \quad \mathcal{U}_f := \text{int}(\mathcal{D}_f).$$

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

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

the *epigraph* of f.

Recall that a function $f : V \to (-\infty, \infty]$ is *convex* if $f(px_1 + (1-p)x_2) \le pf(x_1) + (1-p)f(x_2)$ for all $0 \le p \le 1$ and $x_1, x_2 \in V$. We say that a function f is *strictly convex* on a 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 $x \ne y$. We let $\text{Conv}(V)$ denote the space of functions $f : V \to (-\infty, \infty]$ 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 $f : V \to (-\infty, \infty]$ satisfies $f \in \text{Conv}(V)$ if and only if the epigraph $\mathcal{E}(f)$ is a nonempty, closed, and convex subset of $\mathbb{R}^{d+1}$.

Exercise 4.3 (Epigraph of a lower semi-continuous function) Show that a function $f : V \to (-\infty, \infty]$ is lower semi-continuous if and only if its epigraph $\mathcal{E}(f)$ is a closed subset of $\mathbb{R}^{d+1}$.

Exercise 4.4 (Epigraph of a convex function) Show that a function $f : V \to (-\infty, \infty]$ 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 V. For a proof of the following well-known fact we refer to [Roc70, Thm 10.2].

Lemma 4.5 (Continuity of convex functions) *If $f \in \mathrm{Conv}(V)$, then its restriction to $\mathcal{D}_f$ is a continuous function.*

A function $f : V \to \mathbb{R}$ is *affine* if f and $-f$ are both convex, i.e., if

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

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

$$f(x) = \langle x^*, x\rangle - c^* \qquad (x \in V)$$

for some $x^* \in V^*$ and $c^* \in \mathbb{R}$.

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

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

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

Sometimes, to know a function, it suffices to know only its convex hull.

Lemma 4.7 (Function determined by its convex hull) *Assume that $f : V \to (-\infty, \infty]$ 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 $h \in \mathrm{Conv}(V)$ if there exists a supporting affine function $y \mapsto h(x) + \langle x^*, y - x\rangle$ at x such that

$$h(x) + \langle x^*, y - x\rangle < h(y) \quad \forall y \in V\backslash\{x\}. \tag{4.4}$$

We claim that $f(x) = \overline{f}(x)$ for each exposed point x of $\overline{f}$. To see this, let x^* be as in (4.4) 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 $\overline{f} \le f$ it follows that

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

It is not hard to see that the closed convex sets

$$C_n := \big\{y \in V : f(y) < \overline{f}(x) + \varepsilon_n + \langle x^*, y - x\rangle\big\}$$

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

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

Since $\overline{f} \le f$, 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

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

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

4.4 The Legendre transform

We continue to assume that V is a finite dimensional real linear space and V^* is its dual. The *Legendre transform*[2] of a function $f : V \to (-\infty, \infty]$ is defined as

$$f^*(x^*) := \sup_{x\in V} \big[\langle x^*, x\rangle - f(x)\big] \qquad (x^* \in V^*).$$

This definition is demonstrated in Figure 4.2.

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

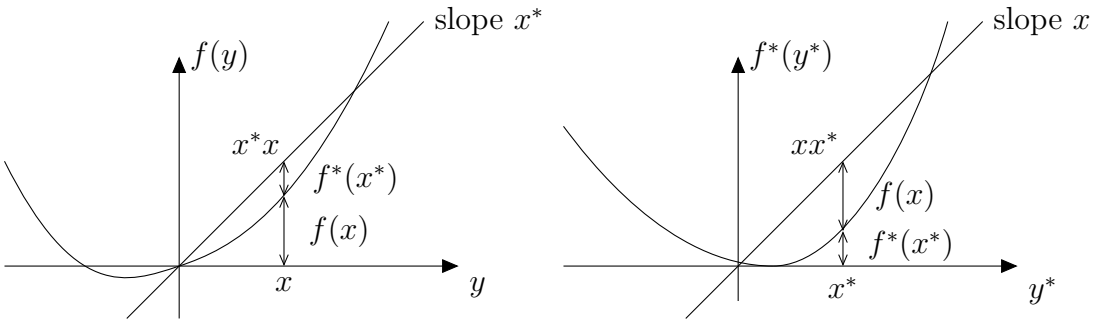

Figure 4.2: The Legendre transform.

Exercise 4.8 For $a^* \in V^*$, let l_{a^*} denote the linear function $l_{a^*}(x) := \langle a^*, x\rangle$. For any function $f : V \to [-\infty, \infty]$ and $a \in V$, define $T_a f(x) := f(x-a)$ $(x \in V)$. Let $l_a : V^* \to \mathbb{R}$ and $T_{a^*} : V^* \to V^*$ be defined similarly. 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 4.9 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 $\mathrm{Conv}(V)$ into $\mathrm{Conv}(V^*)$ and that $(f^*)^* = f$ for each $f \in \mathrm{Conv}(V)$. Below, $\overline{f}$ denotes the convex hull of f.

Lemma 4.10 (Legendre transform) *Assume that $f : V \to (-\infty, \infty]$ is not identically ∞. Then $f^* \in \mathrm{Conv}(V^*)$. One has*

$$\text{(i) } f^*(x^*) = \sup_{(x,c)\in\mathcal{E}(f)} \big[\langle x^*, x\rangle - c\big], \qquad \text{(ii) } \overline{f}(x) = \sup_{(x^*,c^*)\in\mathcal{E}(f^*)} \big[\langle x, x^*\rangle - c^*\big], \tag{4.5}$$

and

$$\begin{aligned} &\text{(i)} \quad \mathcal{E}(\overline{f}) = \big\{(x,c) : \langle x^*, x\rangle - c \leq f^*(x^*)\ \forall x^* \in V^*\big\}, \\ &\text{(ii)} \quad \mathcal{E}(f^*) = \big\{(x^*,c^*) : \langle x, x^*\rangle - c^* \leq f(x)\ \forall x \in V\big\}. \end{aligned} \tag{4.6}$$

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 $f^* \in \mathrm{Conv}(V^*)$.

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

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

which proves (4.5) (i). We next observe that

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

which proves (4.6) (ii). This in turn implies

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

which proves (4.5) (ii). We postpone the proof of (4.6) (i) and first prove the remaining statements.

Since $\langle x^*, x\rangle - c^* \leq f(x)\ \forall x \in V$ if and only if $\langle x^*, x\rangle - c^* \leq \overline{f}(x)\ \forall x \in V$, formula (4.6) (ii) shows that $\mathcal{E}(f^*) = \mathcal{E}((\overline{f})^*)$ and hence $f^* = (\overline{f})^*$.

If $f \in \mathrm{Conv}(V)$ or equivalently $f = \overline{f}$, then (4.5) 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 $f : V \to (-\infty, \infty]$ is not identically ∞, then we can apply what we have just proved to $\overline{f}$ to conclude that $f^{**} = ((\overline{f})^*)^* = \overline{f}$. Formula (4.6) (i) now follows by applying (4.6) (ii) to f^*. ∎

4.5 The essential part of a convex function

We continue to assume that V is a finite dimensional real linear space and V^* is its dual. It often happens that a function $f \in \mathrm{Conv}(V)$ is infinite everywhere except on a lower dimensional affine subspace of V. Also, it often happens that f behaves as an affine function in certain directions. In the present section, we will show how in such cases we can separate the subspaces of V in which f behaves

trivially and reduce f to its essential part, which is a convex function on a lower dimensional space.

We say that a function $f \in \mathrm{Conv}(V)$ is *flat* in the direction $y \in V$ if there exists a constant $r \in \mathbb{R}$ such that

$$f(x + \lambda y) = f(x) + r\lambda \qquad (x \in V,\ \lambda \in \mathbb{R}),$$

and we define the *space of flat directions* of f as

$$\mathcal{F}_f := \big\{ y \in V : f \text{ is flat in the direction } y \big\}.$$

We call the linear form L_f of the following lemma the *affine slope* of f.

Lemma 4.11 (Subspace of flat directions) *For each $f \in \mathrm{Conv}(V)$, the set $\mathcal{F}_f$ is a linear subspace of V. Moreover, there exists a linear form $L_f : \mathcal{F}_f \to \mathbb{R}$ such that*

$$f(x + y) = f(x) + L_f(y) \qquad (x \in V,\ y \in \mathcal{F}_f).$$

Proof For each $y \in \mathcal{F}_f$, let $L_f(y) \in \mathbb{R}$ denote the constant such that

$$f(x + \lambda y) = f(x) + \lambda L_f(y) \qquad (x \in V,\ \lambda \in \mathbb{R}).$$

Then for each $y_1, y_2 \in \mathcal{F}_f$ and $a_1, a_2 \in \mathbb{R}$, one has

$$\begin{aligned} f\big(x + \lambda(a_1 y_1 + a_2 y_2)\big) &= f(x + \lambda a_1 y_1 + \lambda a_2 y_2) \\ &= f(x + \lambda a_1 y_1) + \lambda a_2 L_f(y_2) = f(x) + \lambda(a_1 L_f(y_1) + a_2 L_f(y_2)) \end{aligned}$$

for each $x \in V$ and $\lambda \in \mathbb{R}$. This proves that $a_1 y_1 + a_2 y_2 \in \mathcal{F}_f$ and

$$L_f(a_1 y_1 + a_2 y_2) = a_1 L_f(y_1) + a_2 L_f(y_2).$$

∎

Let f^* denote the Legendre transform of a convex function $f \in \mathrm{Conv}(V)$. The following lemma relates the space of flat directions $\mathcal{F}_{f^*}$ and the affine slope L_{f^*} of f^* to the affine hull of the domain of f.

Lemma 4.12 (Affine hull of the domain) *For each $f \in \mathrm{Conv}(V)$, the affine hull $\mathcal{A}_f$ of $\mathcal{D}_f$ is given by*

$$\mathcal{A}_f = \big\{ x \in V : \langle x^*, x \rangle = L_{f^*}(x^*)\ \forall x^* \in \mathcal{F}_{f^*} \big\}.$$

Proof Assume that $\langle y^*, x\rangle \neq L_{f^*}(y^*)$ for some $x \in V$ and $y^* \in \mathcal{F}_{f^*}$. By Lemma 4.10, f is the Legendre transform of f^*, so

$$f(x) = \sup_{x^* \in V^*} \big[\langle x^*, x\rangle - f^*(x^*)\big].$$

It follows that for each $x^* \in V^*$ and $\lambda \in \mathbb{R}$

$$\begin{aligned} f(x) &\geq \langle x^* + \lambda y^*, x\rangle - f^*(x^* + \lambda y^*) \\ &= \langle x^*, x\rangle - f^*(x^*) + \lambda\big[\langle y^*, x\rangle - L_{f^*}(y^*)\big]. \end{aligned}$$

By assumption, the term in square brackets is nonzero, so since λ is arbitrary we conclude that $f(x) = \infty$. This implies that

$$\mathcal{D}_f \subset \big\{x \in V : \langle x^*, x\rangle = L_{f^*}(x^*)\ \forall x^* \in \mathcal{F}_{f^*}\big\},$$

and hence also $\mathcal{A}_f$ must be contained in the right-hand side of this equation.

To prove the opposite inclusion, let $\tilde{\mathcal{F}}$ be the set of all $x^* \in V^*$ for which there exists a real constant $L(x^*)$ such that

$$\langle x^*, x\rangle = L(x^*) \text{ for all } x \in \mathcal{A}_f.$$

Since $\mathcal{A}_f$ is an affine subspace of V,

$$\mathcal{A}_f = \big\{x \in V : \langle x^*, x\rangle = L(x^*)\ \forall x^* \in \tilde{\mathcal{F}}\big\},$$

so to complete the proof, it suffices to show that $\tilde{\mathcal{F}} \subset \mathcal{F}_{f^*}$ and $L(x^*) = L_{f^*}(x^*)$ for all $x^* \in \tilde{\mathcal{F}}$. Assume that $y^* \in \tilde{\mathcal{F}}$. Then for each $x^* \in V^*$ and $\lambda \in \mathbb{R}$, one has

$$\begin{aligned} f^*(x^* + \lambda y^*) &= \sup_{x \in V} \big[\langle x^* + \lambda y^*, x\rangle - f(x)\big] = \sup_{x \in \mathcal{A}_f} \big[\langle x^*, x\rangle + \lambda\langle y^*, x\rangle - f(x)\big] \\ &= \sup_{x \in \mathcal{A}_f} \big[\langle x^*, x\rangle - f(x)\big] + \lambda L(y^*) = f^*(x^*) + \lambda L(y^*), \end{aligned}$$

which proves that $y^* \in \mathcal{F}_{f^*}$ and $L(y^*) = L_{f^*}(y^*)$. ∎

For any $f \in \mathrm{Conv}(V)$, we set

$$\mathcal{L}_f := \big\{x \in V : \langle x^*, x\rangle = 0\ \forall x^* \in \mathcal{F}_{f^*}\big\}. \tag{4.7}$$

We call $\mathcal{L}_f$ the space of *nontrivial* directions of f. In view of (4.1), (4.7) implies that

$$\mathcal{F}_{f^*} = \big\{x^* \in V^* : \langle x^*, x\rangle = 0\ \forall x \in \mathcal{L}_f\big\}. \tag{4.8}$$

Note that as a result of Lemma 4.12, for any fixed $x_\circ \in \mathcal{A}_f$, one has $\mathcal{A}_f = \{x_\circ + x : x \in \mathcal{L}_f\}$. Equivalently, $\mathcal{L}_f$ is the linear span of all vectors of the form $y - x$ with $x, y \in \mathcal{D}_f$.

Lemma 4.13 (Nonempty interior) *For a function $f \in \mathrm{Conv}(V)$, the following statements are equivalent: (i) $\mathcal{U}_f \neq \emptyset$, (ii) $\mathcal{L}_f = V$, (iii) $\mathcal{F}_{f^*} = \{0\}$.*

Proof The interior $\mathcal{U}_f$ of the domain $\mathcal{D}_f$ is nonempty if and only if the affine hull $\mathcal{A}_f$ of $\mathcal{D}_f$ is the whole space V. As a result of Lemma 4.12, for any fixed $x_\circ \in \mathcal{A}_f$, one has $\mathcal{A}_f = \{x_\circ + x : x \in \mathcal{L}_f\}$, so $\mathcal{A}_f = V$ if and only if $\mathcal{L}_f = V$. This proves the equivalence of (i) and (ii). The implication (ii)⇒(iii) follows from (4.8) and the converse (iii)⇒(ii) follows from (4.7). ∎

We define

$$\begin{aligned}\mathrm{Conv}^+(V) &:= \big\{f \in \mathrm{Conv}(V) : \mathcal{U}_f \neq \emptyset \text{ and } \mathcal{U}_{f^*} \neq \emptyset\big\}\\ &= \big\{f \in \mathrm{Conv}(V) : \mathcal{L}_f = V,\ \mathcal{F}_f = \{0\}\big\},\end{aligned}$$

where the equivalence of both definitions follows from Lemma 4.13. Note that the first definition shows that $f \in \mathrm{Conv}^+(V)$ implies $f^* \in \mathrm{Conv}^+(V^*)$. Our aim will be to show that each $f \in \mathrm{Conv}(V)$ can be decomposed in a nonessential part and an essential part, which is a convex function g on a lower dimensional space W that satisfies $g \in \mathrm{Conv}^+(W)$.

Lemma 4.14 (Nontrivial and flat directions) *For any $f \in \mathrm{Conv}(V)$ one has $\mathcal{F}_f \subset \mathcal{L}_f$.*

Proof Let $z \in \mathcal{F}_f$. Since $f \in \mathrm{Conv}(V)$, there exists an $x \in V$ such that $f(x) < \infty$. Now $f(x+z) = f(x) + L_f(z) < \infty$, showing that $x \in \mathcal{D}_f$ and $x + z \in \mathcal{D}_f$. Since $\mathcal{L}_f$ is the linear span of all vectors of the form $y - x$ with $x, y \in \mathcal{D}_f$, this proves that $z = (x+z) - x \in \mathcal{L}_f$. ∎

Lemma 4.15 (A natural choice of bases) *For any $f \in \mathrm{Conv}(V)$, it is possible to choose a basis $\{e(1), \ldots, e(d)\}$ of V such that $e(1), \ldots, e(d_1)$ span $\mathcal{F}_f$ and $e(1), \ldots, e(d_2)$ span $\mathcal{L}_f$, for some $0 \leq d_1 \leq d_2 \leq d$. For any such basis, let $\{e^*(1), \ldots, e^*(d)\}$ be the corresponding dual basis of V^*. Then $e^*(d_1+1), \ldots, e^*(d)$ span $\mathcal{L}_{f^*}$ and $e^*(d_2+1), \ldots, e^*(d)$ span $\mathcal{F}_{f^*}$.*

Proof The first statement is immediate from Lemma 4.14. To prove the statement about the dual basis, for each $x \in V$ and $x^* \in V^*$, we let $x_1, \ldots, x_d$ and $x_1^*, \ldots, x_d^*$ denote the coordinates of x and x^* with respect to the basis $\{e(1), \ldots, e(d)\}$ and the dual basis $\{e^*(1), \ldots, e^*(d)\}$, respectively, i.e., these are the real numbers uniquely defined by the relations

$$x = \sum_{i=1}^{d} x_i e(i) \quad \text{and} \quad x^* = \sum_{i=1}^{d} x_i^* e^*(i).$$

Then $\langle x^*, x\rangle = \sum_{i=1}^d x_i^* x_i$ ($x \in V,\ x^* \in V^*$). By (4.7) applied to $\mathcal{L}_{f^*}$,

$$x^* \in \mathcal{L}_{f^*} \quad \Leftrightarrow \quad \langle x^*, x\rangle = 0\ \forall x \in \mathcal{F}_f \quad \Leftrightarrow \quad x_i^* = 0\ \forall i \in \{1, \dots, d_1\},$$

which shows that $e^*(d_1+1), \dots, e^*(d)$ span $\mathcal{L}_{f^*}$. Similarly, by (4.8)

$$x^* \in \mathcal{F}_{f^*} \quad \Leftrightarrow \quad \langle x^*, x\rangle = 0\ \forall x \in \mathcal{L}_f \quad \Leftrightarrow \quad x_i^* = 0\ \forall i \in \{1, \dots, d_2\},$$

which shows that $e^*(d_2+1), \dots, e^*(d)$ span $\mathcal{F}_{f^*}$. ∎

We recall that if V is a linear space and $W \subset V$ is a linear subspace, then setting $x \sim y$ if and only if $x - y \in W$ defines an equivalence relation on V. Let $\underline{x} := \{y \in V : y \sim x\}$ denote the equivalence class containing x. Then the *quotient space* $V/W := \{\underline{x} : x \in V\}$ naturally has the structure of a linear space, with $\lambda\underline{x} := \underline{\lambda x}$ and $\underline{x} + \underline{y} := \underline{x+y}$ ($x, y \in V,\ \lambda \in \mathbb{R}$).

Let $f \in \mathrm{Conv}(V)$ and let $f^* \in \mathrm{Conv}(V^*)$ denote its Legendre transform. By Lemma 4.14, we have $\mathcal{F}_f \subset \mathcal{L}_f$ and $\mathcal{F}_{f^*} \subset \mathcal{L}_{f^*}$. Our aim will be to define functions $g \in \mathrm{Conv}(\mathcal{L}_f/\mathcal{F}_f)$ and $g^* \in \mathrm{Conv}(\mathcal{L}_{f^*}/\mathcal{F}_{f^*})$ that are each other's Legendre transforms and that represent the "essential" parts of the functions f and f^*, after we neglect the nontrivial and flat directions. We first show that the spaces $\mathcal{L}_f/\mathcal{F}_f$ and $\mathcal{L}_{f^*}/\mathcal{F}_{f^*}$ are naturally dual to each other.

Lemma 4.16 (Duality of quotient spaces) *For any $f \in \mathrm{Conv}_1(V)$, setting*

$$\langle \underline{x}^*, \underline{x}\rangle := \langle x^*, x\rangle \qquad (x \in \mathcal{L}_f,\ x^* \in \mathcal{L}_{f^*}) \tag{4.9}$$

unambiguously defines a function $\langle\,\cdot\,,\,\cdot\,\rangle$ such that $\mathcal{L}_f/\mathcal{F}_f$ and $\mathcal{L}_{f^*}/\mathcal{F}_{f^*}$ are dual to each other with respect to this function.

Proof We need to show that $\langle x^* + y^*, x + x\rangle = \langle x^*, x\rangle$ for all $x^* \in \mathcal{L}_{f^*}$, $y^* \in \mathcal{F}_{f^*}$, $x \in \mathcal{L}_f$, and $y \in \mathcal{F}_f$. Since $\mathcal{F}_f \subset \mathcal{L}_f$ and $\mathcal{F}_{f^*} \subset \mathcal{L}_{f^*}$, it suffices to observe that by (4.7), $\langle x^*, y\rangle = 0$ for all $x^* \in \mathcal{L}_{f^*}$ and $y \in \mathcal{F}_f$, and $\langle y^*, x\rangle = 0$ for all $y^* \in \mathcal{F}_{f^*}$ and $x \in \mathcal{L}_f$. ∎

The following lemma gives the anticipated decomposition of a convex function in is essential and inessential parts.

Lemma 4.17 (Reduction to the essential part) *Let $f \in \mathrm{Conv}(V)$ and let $x_\circ \in \mathcal{A}_f$ and $x_\circ^* \in \mathcal{A}_{f^*}$ satisfy $\langle x_\circ^*, x_\circ\rangle = 0$. Then setting*

$$g(\underline{x}) := f(x_\circ + x) - \langle x_\circ^*, x\rangle \quad (x \in \mathcal{L}_f),$$
$$g^*(\underline{x}^*) := f^*(x_\circ^* + x^*) - \langle x^*, x_\circ\rangle \quad (x^* \in \mathcal{L}_{f^*}),$$

unambiguously defines functions $g \in \mathrm{Conv}^+(\mathcal{L}_f/\mathcal{F}_f)$ *and* $g^* \in \mathrm{Conv}^+(\mathcal{L}_{f^*}/\mathcal{F}_{f^*})$. *Moreover,* g^* *is the Legendre transform of* g *when we view* $\mathcal{L}_{f^*}/\mathcal{F}_{f^*}$ *as the dual of* $\mathcal{L}_f/\mathcal{F}_f$ *in the sense of Lemma 4.16.*

Proof We claim that

$$f(x+y) = f(x) + \langle x_\circ^*, y\rangle \qquad (x \in V,\ y \in \mathcal{F}_f).$$

Indeed, we have

$$f(x+y) = f(x) + L_f(y) \qquad (x \in V,\ y \in \mathcal{F}_f),$$

where $L_f(y) = \langle x^*, y\rangle$ for all $x^* \in \mathcal{A}_f$ and $y \in \mathcal{F}_f$ by Lemma 4.12. Applying this to $x_\circ^* \in \mathcal{A}_f$, the claim follows. It follows that $g(\underline{x+y}) = g(\underline{x})$ for all $x \in V$ and $y \in \mathcal{F}_f$, so the definition of $g(\underline{x})$ does not depend on the choice of the representative x of the equivalence class $\underline{x}$. By the same argument, g^* is also well-defined.

Let h denote the Legendre transform of g. Then for each $x^* \in \mathcal{L}_{f^*}$,

$$\begin{aligned} h(\underline{x}^*) &= \sup_{\underline{x}\in\mathcal{L}_f/\mathcal{F}_f} \big[\langle \underline{x}^*, \underline{x}\rangle - g(\underline{x})\big] = \sup_{x\in\mathcal{L}_f} \big[\langle x^*, x\rangle - f(x_\circ + x) + \langle x_\circ^*, x\rangle\big] \\ &= \sup_{x\in\mathcal{L}_f} \big[\langle x_\circ^* + x^*, x\rangle - f(x_\circ + x)\big] \overset{!}{=} \sup_{x\in V} \big[\langle x_\circ^* + x^*, x\rangle - f(x_\circ + x)\big] \\ &= \sup_{y\in V} \big[\langle x_\circ^* + x^*, y - x_\circ\rangle - f(y)\big] = f^*(x_\circ^* + x^*) - \langle x_\circ^* + x^*, x_\circ\rangle, \end{aligned}$$

where in the equality marked with ! we have used that $\mathcal{A}_f = \{x_\circ + x : x \in \mathcal{L}_f\}$ and hence $f(x_\circ + x) = \infty$ for all $x \in V\backslash\mathcal{L}_f$. Using moreover the assumption that $\langle x_\circ^*, x_\circ^*\rangle = 0$, we see that $h(\underline{x}) = g^*(\underline{x})$.

We claim that $\mathcal{F}_g = \{0\}$. Indeed, if $\underline{y} \in \mathcal{F}_g$, then there exists a constant $r \in \mathbb{R}$ such that $g(\underline{x}+\lambda\underline{y}) = g(\underline{x}) + r\lambda$ for all $\lambda \in \mathbb{R}$. It follows that for any representative y of $\underline{y}$, we must have $y \in \mathcal{F}_f$ and hence $\underline{y} = 0$. By symmetry also $\mathcal{F}_{g^*} = \{0\}$ and hence $g \in \mathrm{Conv}^+(\mathcal{L}_f/\mathcal{F}_f)$. ∎

The following lemma shows that it is always possible to choose $x_\circ$ and $x_\circ^*$ as in Lemma 4.17.

Lemma 4.18 (Orthogonal reference points) *For each* $f \in \mathrm{Conv}(V)$, *it is possible to choose* $x_\circ \in \mathcal{A}_f$ *and* $x_\circ^* \in \mathcal{A}_{f^*}$ *such that* $\langle x_\circ^*, x_\circ\rangle = 0$.

Proof If $0 \in \mathcal{A}_f$, then we can choose $x_\circ := 0$ and $x_\circ^*$ arbitrary. If the opposite case we choose $x_\circ \in \mathcal{A}_f$ arbitrary. Since $0 \notin \mathcal{A}_f$, we have $\mathcal{A}_f \cap \mathcal{L}_f = \emptyset$. By

Lemma 4.14, it follows that $x_\circ \notin \mathcal{F}_f$ and hence by (4.8) we can choose $z^* \in \mathcal{L}_{f^*}$ such that $\langle x_\circ, z^* \rangle \neq 0$. It follows that we can set $x_\circ^* := y^* + \lambda z^*$ where $y^* \in \mathcal{A}_{f^*}$ is arbitrary and $\lambda \in \mathbb{R}$ is chosen such that $\lambda \langle x_\circ, z^* \rangle = -\langle x_\circ, y^* \rangle$. ∎

The following lemma is a a reformulation of Lemma 4.17 in terms of the bases $\{e(1), \ldots, e(d)\}$ and $\{e^*(1), \ldots, e^*(d)\}$ from Lemma 4.15. Below, we equip $\mathbb{R}^{d_2 - d_1}$ with the standard inner product, making it dual to itself.

Lemma 4.19 (Essential part of a convex function) *Let $f \in \mathrm{Conv}(V)$ and let $\{e(1), \ldots, e(d)\}$ and $\{e^*(1), \ldots, e^*(d)\}$ be bases of V and V^* as in Lemma 4.15. Then there exist real constants $a_1^*, \ldots, a_{d_1}^*$ and $a_{d_2+1}, \ldots, a_d$ and a function $g \in \mathrm{Conv}^+(\mathbb{R}^{d_2 - d_1})$ such that*

$$f(x_1, \ldots, x_d) = \begin{cases} \displaystyle\sum_{i=1}^{d_1} a_i^* x_i + g(x_{d_1+1}, \ldots, x_{d_2}) & \text{if } (x_{d_2+1}, \ldots, x_d) = (a_{d_2+1}, \ldots, a_d) \\ \infty & \text{otherwise.} \end{cases}$$

Moreover, the Legendre transform of f is given by

$$f^*(x_1^*, \ldots, x_d^*) = \begin{cases} g^*(x_{d_1+1}^*, \ldots, x_{d_2}^*) + \displaystyle\sum_{i=d_2+1}^{d} x_i^* a_i & \text{if } (x_1^*, \ldots, x_{d_1}^*) = (a_1^*, \ldots, a_{d_1}^*) \\ \infty & \text{otherwise,} \end{cases}$$

where g^* is the Legendre transform of g.

Proof By Lemma 4.12, $x \in \mathcal{A}_f$ if and only if $\langle y^*, x \rangle = L_{f^*}(y^*)$ for all $y^* \in \mathcal{F}_{f^*}$. It follows that there exist real constants $a_{d_2+1}, \ldots, a_d$ such that $x \in \mathcal{A}_f$ if and only if $x_i = a_i$ for all $i \in \{d_2 + 1, \ldots, d\}$. Similarly, there exist $a_1^*, \ldots, a_{d_1}^*$ such that $x^* \in \mathcal{A}_{f^*}$ if and only if $x_i^* = a_i^*$ for all $i \in \{1, \ldots, d_1\}$. Setting

$$x_\circ := (0, \ldots, 0, a_{d_2+1}, \ldots, a_d) \quad \text{and} \quad x_\circ^* := (a_1^*, \ldots, a_{d_1}^*, 0, \ldots, 0),$$

now defines $x_\circ \in \mathcal{A}_f$ and $x_\circ^* \in \mathcal{A}_{f^*}$ such that $\langle x_\circ^*, x_\circ \rangle = 0$. The claim now follows from Lemma 4.17. ∎

In the following sections, we will often need the assumption that a function $f \in \mathrm{Conv}(V)$ satisfies $\mathcal{U}_f \neq \emptyset$. Often, such an assumption can be made more or less without loss of generality, if we replace f by its essential part. The following exercise demonstrates this.

Exercise 4.20 (Function determined by its convex hull) Let $f : V \to (-\infty, \infty]$ be lower semi-continuous, let $\overline{f}$ be its convex hull, and let g be the essential part of $\overline{f}$ as in Lemma 4.19. Assume that $\mathcal{F}_{\overline{f}} = \{0\}$ and g is strictly convex on $\mathcal{U}_g$. Show that $f = \overline{f}$. Hint: combine Lemmas 4.7 and 4.19.

4.6 The generalized gradient

In this and the following sections, we will be interested in the derivatives of convex functions. We continue to assume that V is a finite dimensional real linear space and V^* is its dual.

Recall from Section 4.2 the definition of a supporting hyperplane. Assume that $f \in \mathrm{Conv}(V)$ 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 $(x^*, a^*) \in V^* \times \mathbb{R}$ such that

$$\langle x^*, y - x\rangle + a^*\big(z - f(x)\big) \leq 0 \quad \forall y \in \mathcal{D}_f \text{ and } z \geq f(y). \tag{4.10}$$

Note that this implies $a^* \leq 0$, since otherwise (4.10) 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

$$Hf := \big\{(x, x^*, a^*) : x \in \mathcal{D}_f,\ (x^*, a^*) \in Hf(x)\big\} \tag{4.11}$$

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

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

and we define $H'f$ and $H''f$ as in (4.11) 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 4.21 (Supporting hyperplanes) *Assume that $f \in \mathrm{Conv}(V)$ and that $\mathcal{U}_f \neq \emptyset$. Then,*

(a) *$Hf(x)$ is a closed convex cone in $V^* \times \mathbb{R}$ for each $x \in \mathcal{D}_f$,*

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

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

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

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

Proof Part (a) is immediate from (4.10). 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 4.2 to conclude that for each $x \in \mathcal{D}_f$, there exist $(x^*, a^*) \in (V^* \times \mathbb{R})\backslash\{(0,0)\}$ 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 (4.10). This proves part (b).

To prove part (c), we use part (b) and observe that by (4.10), $(x^*, 0) \in H'f(x)$ 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 $(x_n, x_n^*, a_n^*) \in H''f$ converge to a limit (x, x^*, a^*) in $V \times V^* \times (\mathbb{R}\backslash\{0\})$. Then, taking the limit in (4.10), we see that $(x_n, x_n^*, a_n^*) \in Hf$. Since $a_n^* < 0$, formula (4.10) moreover implies that $f(x) < \infty$ and hence $x \in \mathcal{D}_f$, so we see that $(x_n, x_n^*, a_n^*) \in H''f$.

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 $\mathcal{D}_f \times V^* \times (\mathbb{R}\backslash\{0\})$. 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. ∎

We say that an affine function $y \mapsto x^*y - c^*$ is *supporting* at a point $x \in \mathcal{D}_f$ if

$$\langle x^*, x\rangle - c^* = f(x) \quad \text{and} \quad \langle x^*, y\rangle - c^* \le f(y) \quad (y \in V).$$

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

$$\begin{aligned} Df(x) &:= \big\{x^* \in V^* : f(x) + \langle x^*, y - x\rangle \le f(y)\ \forall y \in V\big\}, \\ Df &:= \big\{(x, x^*) : x \in \mathcal{D}_f,\ x^* \in Df(x)\big\}. \end{aligned} \tag{4.12}$$

Df is the collection of all slopes of supporting affine functions at x. We say that a function $f : V \to \mathbb{R}$ is differentiable at $x \in V$ if there exists a $\partial f(x) \in V^*$ such that

$$\langle \partial f(x), y\rangle = \lim_{\varepsilon \to 0} \varepsilon^{-1}\big[f(x + \varepsilon y) - f(x)\big] \qquad (y \in V). \tag{4.13}$$

Note that the right-hand side of this equation is the *directional derivative* of f at x in the direction y. We call $\partial f(x)$ the *gradient* of f at x. If $f \in \mathrm{Conv}(V)$ is differentiable at $x \in \mathcal{D}_f$, 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 4.21.

Proposition 4.22 (Generalized gradient) *Assume that $f \in \mathrm{Conv}(V)$ 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 $V \times V^*$,*

(d) *Df is a connected subset of $V \times V^*$.*

(e) *$\{(x, x^*) \in Df : x \in K\}$ is a compact subset of $V \times V^*$ for each compact $K \subset \mathcal{U}_f$.*

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 4.21 (b) and (c) imply that for each $x \in \mathcal{U}_f$, there exists an $x^* \in V$ such that $(x^*, -1) \in H''f(x)$. Then (4.10) 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 4.21 (b) and (c). Part (b) is immediate from the definition of $Df(x)$ in (4.12). Parts (c) and (d) follow from Proposition 4.21 (d) and (e) and our earlier observation that each $(x^*, a^*) \in H''f(x)$ can be normalized so that $a^* = -1$.

To prove part (e), let $K \subset \mathcal{U}_f$ be compact. Then $\{(x, x^*) \in Df : x \in K\}$ is closed by part (c), so it suffices to show that it is moreover bounded. Assume, to the contrary, that there exist $(x_n, x_n^*) \in Df$ with $x_n \in K$ and $|x_n^*| \to \infty$. Since K is compact, by going to a subsequence, we may assume that $x_n \to x \in \mathcal{U}_f$. Let $c_n^* := f(x_n) - \langle x_n^*, x_n\rangle$. Then $\langle x_n^*, x_n\rangle - c_n^* = f(x_n)$ and $\langle x_n^*, y\rangle - c_n^* \leq f(y)$ for all $y \in V$. Equivalently, this says that $H_{(x_n^*, -1), c_n^*}$ is a supporting hyperplane for $\mathcal{E}(f)$ at the point $(x_n, f(x_n))$. Let $\varepsilon_n := |x_n^*|^{-1}$. Then $H_{(\varepsilon_n x_n^*, -\varepsilon_n), \varepsilon_n c_n^*}$ is the same supporting hyperplane. By going to a subsequence, we can assume that $\varepsilon_n x_n^* \to x^*$ where $|x^*| = 1$. Then $\varepsilon_n c_n^* \to -\langle z^*, x\rangle =: x^*$ and $H_{(x^*, 0), c^*}$ is a vertical supporting hyperplane for $\mathcal{E}(f)$ at the point $(x, f(x))$. By Proposition 4.21 (c), this contradicts the fact that $x \in \mathcal{U}_f$. ∎

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 4.23 (Uniqueness of the slope) *Let $f \in \mathrm{Conv}(V)$. 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 (4.13)

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 $g : \mathcal{U}_f \to V$ such that

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

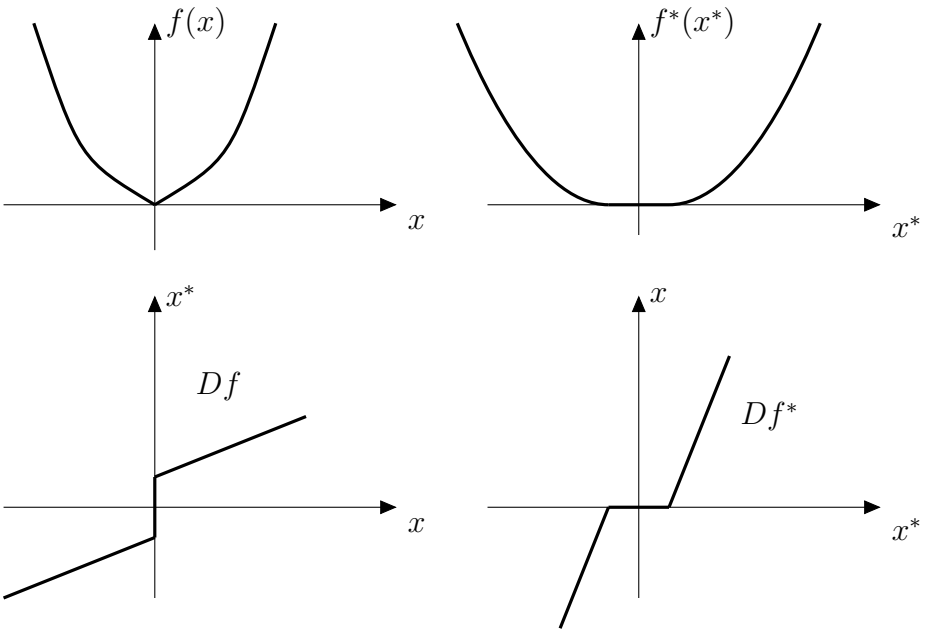

Figure 4.3: Legendre transform of a non-smooth function.

Now Proposition 4.22 (c) says that the graph of g is a closed subset of $\mathcal{U}_f \times V$. Using moreover Proposition 4.22 (e) and the closed graph theorem, we can deduce that g is continuous. The technical part of the proof is showing that g is indeed the gradient of 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 4.2. See also Figure 4.3 for an illustration of the Legendre transform of a non-smooth function.

Lemma 4.24 (Slope of the Legendre transform) *For any $f \in \mathrm{Conv}(V)$ and $x, x^* \in V^*$, one has*

$$\langle x^*, x\rangle \leq f(x) + f^*(x^*) \tag{4.14}$$

Moreover,

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

Proof The inequality (4.14) follows immediately from the definition $f^*(x^*) = \sup_{x\in V}\big[\langle x^*, x\rangle - f(x)\big]$. Assume that $(x, x^*) \in Df$. Then there exists a $c^* \in \mathbb{R}$

such that $\langle x^*, x\rangle - c^* = f(x)$ and $\langle x^*, y\rangle - c^* \leq f(y)$ for all $y \in V$. By (4.6) (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 $\langle x^*, y\rangle - c^* + \varepsilon \leq f(y)$ for all $y \in V$, which again by (4.6) (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 (4.6) (ii) implies that $x^*y - f^*(x^*) \leq f(y)$ for all $y \in V$. 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 $(x^*, x) \in Df^*$. ∎

Exercise 4.25 (Nonempty generalized gradient) Show that $Df \neq \emptyset$ for all $f \in \mathrm{Conv}(V)$. Hint: combine Lemma 4.19 and Proposition 4.22.

4.7 Extensions of convex functions

We continue to assume that V is a finite dimensional real linear space and V^* is its dual. For any $g \in \mathrm{Conv}(V)$ and closed convex set $D \subset V$, setting $f(x) := g(x)$ for $x \in D$ and $:= \infty$ otherwise defines a function $f \in \mathrm{Conv}(V)$. 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 4.26 (Restriction of a convex function) *Let $f, g \in \mathrm{Conv}(V)$ 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 $f \in \mathrm{Conv}(V)$ and assume that $\mathcal{U}_f \neq \emptyset$. By definition, we say that a f is *on natural domain* if it satisfies the equivalent conditions (i) and (ii) of the following lemma. If f satisfies condition (iii), then we say that f is *on maximal domain*.[3]

Lemma 4.27 (Convex functions on natural domain) *Let $f \in \mathrm{Conv}(V)$ 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 $V \times V^*$,*

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

(iii) *$f = g$ for all $g \in \mathrm{Conv}(V)$ that extend f.*

Proof The implication (ii)⇒(i) follows from the fact that by Proposition 4.22 (c), Df is a closed subset of $V \times V^*$. Assume, conversely, that $D_\circ f$ is a closed subset of $V \times V^*$. Then $D_\circ f$ is a closed subset of Df. On the other hand, since $D_\circ f = Df \cap (\mathcal{U}_f \times V)$, it is also open as a subset of Df. By Proposition 4.22 (d), Df is connected so, also using the fact that $D_\circ f \neq \emptyset$, by Proposition 4.22 (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 $g \in \mathrm{Conv}(V)$ 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 4.22 (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 In dimension $d = 1$, it is easy to check that the conditions (i)–(iii) of Lemma 4.27 are in fact all equivalent. It is tempting to conjecture that the same is true in higher dimensions, but this seems to be false. Here is a sketch of a counterexample. We define an open square by $W := \{x \in \mathbb{R}^2 : 0 < x_i < 1 \ \forall i = 1, 2\}$ and let $\overline{W}$ denote its closure and $\partial W := \overline{W} \backslash W$ its boundary. We also write

$$I_1 := \{x \in \mathbb{R}^2 : x_1 \in \{0, 1\},\ 0 < x_2 < 1\},\ I_2 := \{x \in \mathbb{R}^2 : 0 < x_1 < 1,\ x_2 \in \{0, 1\}\}.$$

[3] These definitions are not standard, and there does not seem to exist established terminology for this.

It should be possible to construct a function $h \in \mathrm{Conv}(\mathbb{R}^2)$ with the following properties:

(i) $\mathcal{D}_h = \overline{W}$,

(ii) $h(x) = 1$ for all $x \in \partial W$,

(iii) h is continuously differentiable on W,

(iv) $|\frac{\partial}{\partial x_i} h(x(n))| \to \infty$ if $W \ni x(n) \to x \in I_i$ $(i = 1, 2)$.

Now let l be the linear function $l(x) := (x_1 + x_2)/2$. Note that $l(x) = 1 = h(x)$ in the point $x = (1, 1)$ but $l(x) < 1 = h(x)$ for $x \in I_1 \cup I_2$. Then the convex function $f := h \vee l$ does not satisfy condition (ii) of Lemma 4.27, since there exists some $(x, x^*) \in Df$ with $x = (1, 1) \notin \mathcal{U}_f$, but it satisfies condition (iii) of Lemma 4.27 since the condition on the derivatives of h makes it impossible to extend f outside the square $\overline{W}$.

4.8 Well-behaved convex functions

We continue to assume that V is a finite dimensional real linear space and V^* is its dual. We say that a function $f \in \mathrm{Conv}(V)$ is *well-behaved*[4] if there exists a homeomorphism $f' : \mathcal{U}_f \to \mathcal{U}_{f^*}$ such that $Df = \big\{(x, f'(x)) : x \in \mathcal{U}_f\big\}$.

Lemma 4.28 (Well-behaved functions) *If $f \in \mathrm{Conv}(V)$ is well-behaved, then $\mathcal{U}_f \neq \emptyset$ and f^* is also well-behaved. Moreover, f is continuously differentiable on $\mathcal{U}_f$, the gradient $\partial f : \mathcal{U}_f \to \mathcal{U}_{f^*}$ is a homeomorphism, and $(\partial f)^{-1} = \partial f^*$.*

Proof By Exercise 4.20, $Df \neq \emptyset$ so the assumption that $Df = \big\{(x, f'(x)) : x \in \mathcal{U}_f\big\}$ implies that $\mathcal{U}_f \neq \emptyset$. Let f'^{-1} denote the inverse of f'. Then f'^{-1} is a homeomorohism from $\mathcal{U}_{f^*} \to \mathcal{U}_f$ and by Lemma 4.24, $Df^* = \big\{(x^*, x) : (x, x^*) \in Df\big\} = \big\{(x^*, f'{-}1(x^*)) : x^* \in \mathcal{U}_{f^*}\big\}$, which shows that f^* is well-behaved. Since $Df = \big\{(x, f'(x)) : x \in \mathcal{U}_f\big\}$, Lemma 4.23 implies that f is continuously differentiable on $\mathcal{U}_f$ and $Df(x) = \{\partial f(x)\}$ for all $x \in \mathcal{U}_f$. This shows that $\partial f = f'$, the

[4]This definition is not standard, and there does not seem to exist established terminology for this.

homeomorphism in the definition of a well-behaved convex function. Lemma 4.24 now implies that $(\partial f)^{-1} = \partial f^*$. ∎

Lemma 4.28 shows that if $f \in \mathrm{Conv}(V)$ is well-behaved, then $f \in \mathrm{Conv}^+(V)$. For each $1 \leq n \leq \infty$, we set

$$\mathrm{Conv}_n^+(V) := \big\{ f \in \mathrm{Conv}(V) : f \text{ is well-behaved and } n \text{ times continuously differentiable on } \mathcal{U}_f \big\}.$$

In view of Lemma 4.24, $\mathrm{Conv}_n^+(V) \subset \mathrm{Conv}^+(V)$, and $\mathrm{Conv}_1^+(V)$ is simply the space of well-behaved convex functions on V. We also set

$$\mathrm{Conv}_n(V) := \big\{ f \in \mathrm{Conv}(V) : \text{the essential part } g \text{ of } f \text{ satisfies } g \in \mathrm{Conv}_n^+(\mathcal{L}_f/\mathcal{F}_f) \big\}.$$

We recall that the essential part of a convex function is defined in Lemma 4.17. It is easy to see that the definition of $\mathrm{Conv}_n(V)$ does not depend on the choice of the reference points $x_\circ$ and $x_\circ^*$. We call elements of $\mathrm{Conv}_1(V)$ *essentially well-behaved* convex functions.

Lemma 4.29 (Well-behaved Legendre transform) *If $f \in \mathrm{Conv}_n^+(V)$, then $f^* \in \mathrm{Conv}_n^+(V^*)$, and if $f \in \mathrm{Conv}_n(V)$, then $f^* \in \mathrm{Conv}_n(V^*)$.*

Proof In Lemma 4.28 we have already seen that $f \in \mathrm{Conv}_1^+(V)$ implies $f^* \in \mathrm{Conv}_1^+(V^*)$. If f is n times continuously differentiable on $\mathcal{U}_f$, then the homeomorphism ∂f is $n-1$ times continuously differentiable, hence its inverse $(\partial f)^{-1} = \partial f^*$ has the same properties and as a result f^* is n times continuously differentiable on $\mathcal{U}_{f^*}$. This proves that $f \in \mathrm{Conv}_n^+(V)$ implies $f^* \in \mathrm{Conv}_n^+(V^*)$. The final claim follows from the fact that by Lemma 4.17, if g is the essential part of f, then its Legendre transform g^* is the essential part of f^*. ∎

For a function $f : \mathbb{R}^d \to \mathbb{R}$ that is twice continuously differentiable in an open neighborhood of a point x, we let $\partial^2 f(x)$, defined as

$$\partial_{ij}^2 f(x) := \tfrac{\partial^2}{\partial x_i \partial x_j} f(x),$$

denote the matrix of its second derivatives. In the abstract setting, we note that each linear map $L : V \to V^*$ has a unique adjoint map $L^\dagger : V \to V^*$ such that $\langle Lx, y\rangle = \langle L^\dagger y, x\rangle$ $(x, y \in V)$. We say that L is *self-adjoint* if $L = L^\dagger$. Now if $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable in an open neighborhood of a point x, then $\partial^2 f(x)$ is the unique self-adjoint linear map from V to V^* such that

$$f(x+y) = f(x) + \langle \partial f(x), y\rangle + \tfrac{1}{2}\langle \partial^2 f(x) y, y\rangle + o(|y|^2) \qquad \text{as } y \to 0.$$

Proposition 4.30 (Conditions for well-behavedness) *Let $f \in \mathrm{Conv}(V)$. Then f is well-behaved if and only if it satisfies the following conditions.*

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

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

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

(iv) *f is on natural domain.*

Condition (iv) is equivalent to

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

If f is twice continuously differentiable on $\mathcal{U}_f$, then condition (iii) is equivalent to

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

Proof We first prove the necessity of conditions (i)–(iv). The necessity of (i) and (ii) has already been proved in Lemma 4.28. To prove the necessity of (iii), we observe that if f would not be stricly convex on $\mathcal{U}_f$, then there would be $x, y \in \mathcal{U}_f$ with $x \neq y$ such that $\partial f(x) = \partial f(y)$, contradicting the fact proved in Lemma 4.28 that ∂f is a bijection implies that f is strictly convex. Finally, the definition of a well-behaved convex function immediately implies condition (ii) of Lemma 4.27, showing that f is on maxial domain and hence (iv) is necessary.

We next show the sufficiency of conditions (i)–(iv). Using Lemma 4.27 and the assumptions that f is on natural 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 on $\mathcal{U}_f$, no two points in $\mathcal{U}_f$ can have the same slope, so ∂f is a bijection from $\mathcal{U}_f$ to its image

$$\mathcal{V} := \big\{\partial f(x) : x \in \mathcal{U}_f\big\}$$

It now follows from Lemma 4.24 that

$$\mathcal{V} = \{x^* \in V^* : Df^*(x^*) \neq \emptyset\} \quad \text{and} \quad Df^*(x^*) = \{(\partial f)^{-1}(x^*)\} \quad (x^* \in \mathcal{V}).$$

In particular, by Proposition 4.22 (a), it follows that $\mathcal{U}_{f^*} \subset \mathcal{V}$. We claim that this is in fact an equality. Indeed, if $Df^*(x^*) \neq \emptyset$ for some $x^* \in \mathcal{D}_{f^*} \backslash \mathcal{U}_{f^*}$, then there is a supporting hyperplane for $\mathcal{E}(f^*)$ at the point $(x^*, f(x^*))$ that is not vertical. But using the convexity of $\mathcal{U}_{f^*}$, it is easy to see that for each $x^* \in \mathcal{D}_{f^*} \backslash \mathcal{U}_{f^*}$, there must also be a vertical supporting hyperplane for $\mathcal{E}(f^*)$ at the point $(x^*, f(x^*))$. But then, by Proposition 4.21 (a), all convex combinations of the vertical and non-vertical supporting hyperplanes are also supporting hyperplanes. This shows that if $Df^*(x^*) \neq \emptyset$ for some $x^* \in \mathcal{D}_{f^*} \backslash \mathcal{U}_{f^*}$, then $Df^*(x^*)$ is never a singleton. This contradicts what we have just proved, allowing us to conclude that $\mathcal{U}_{f^*} = \mathcal{V}$. We can now invoke Lemma 4.23 to conclude that f^* is continuously differentiable on $\mathcal{U}_{f^*}$ and $(\partial f)^{-1} = \partial f^*$ on $\mathcal{U}_{f^*}$. In particular, ∂f is a continuous function having a continuous inverse, i.e., a homeomorphism This concludes the proof of the sufficiency of conditions (i)–(iv).

To see that (iv) and (iv)' are equivalent, we observe that if (iv)' holds, then $\{(x, x^*) \in Df : x \in \mathcal{U}_f\}$ is a closed subset of Df and hence, by Proposition 4.22 (c), also of $V \times V^*$, so f is on natural domain by condition (i) of Lemma 4.27. On the other hand, if (iv)' fails, then by going to a subsequence we can find $\mathcal{U}_f \ni x_n \to x \in V \backslash \mathcal{U}_f$ such that $\partial f(x_n)$ converges to a finite limit $x^* \in V^*$. Since Df is closed by Proposition 4.22 (c), it follows that $(x, x^*) \in Df$ which contradicts condition (ii) of Lemma 4.27.

If f is twice continuously differentiable in an open neighborhood of x, then

$$\tfrac{\partial^2}{\partial \varepsilon^2} f(x + \varepsilon y)\big|_{\varepsilon=0} = \sum_{i=1}^{d} y_i \tfrac{\partial}{\partial x_i} \Big(\sum_{j=1}^{d} y_j \tfrac{\partial}{\partial x_j} f(x) \Big) = \sum_{i,j=1}^{d} y_i \partial_{ij}^2 f(x) y_j.$$

It is easy to see that f is strictly convex on $\mathcal{U}_f$ if and only if the left-hand side of this equation is strictly positive for all $x \in \mathcal{U}_f$ and $y \in V \backslash \{0\}$. ∎

The following lemma says that for well-behaved convex functions, the supremum occurring in the definition of the Legendre transform is assumed in a unique point.

Lemma 4.31 (Unique maximizer) *Assume that $f \in \mathrm{Conv}_1^+(V)$. Then 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 note that f is strictly convex on $\mathcal{U}_f$ by Proposition 4.30 (iii). By Lemma 4.28, $\partial f : \mathcal{U}_f \to \mathcal{U}_{f^*}$ is a bijection and $(\partial f)^{-1} = \partial f^*$, so for each $x^* \in \mathcal{U}_{f^*}$ there exists a unique $x \in \mathcal{U}_f$ such that $\partial f(x) = x^*$, which is given by $x =$

$\partial f^*(x^*)$. It follows that the strictly concave function $y \mapsto \langle x^*, y\rangle - f(y)$ assumes its maximum in the unique point x. ∎

In Proposition 4.30, we have seen that well-behaved convex functions are on natural domain. The following lemma generalizes this to some functions that are only essentially well-behaved.

Lemma 4.32 (Natural domain) *Assume that $f \in \mathrm{Conv}_1(V)$ and $\mathcal{U}_f \neq \emptyset$. Then f is on natural domain.*

Proof We use coordinates with respect to bases of V and V^* as in Lemma 4.15. Then $d_2 = d$ by the assumption that $\mathcal{U}_f \neq \emptyset$, so by Lemma 4.19 f is of the form

$$f(x_1, \ldots, x_d) = \sum_{i=1}^{d_1} a_i^* x_i + g(x_{d_1+1}, \ldots, x_d)$$

for some constants $a_1^*, \ldots, a_{d_1}^*$ and function $g \in \mathrm{Conv}_1^+(\mathbb{R}^{d-d_1})$, and hence, setting $x := (x_1, \ldots, x_d)$ and $x' := (x_{d_1+1}, \ldots, x_d)$, we have

$$\partial f(x) = \big(a_1^*, \ldots, a_{d_1}^*, \partial_{d_1+1} g(x'), \ldots, \partial_d g(x')\big) \qquad (x \in \mathcal{U}_f).$$

By Lemma 4.27 (i), to check that f is on natural domain, it suffices to show that $|\partial f(x(n))| \to \infty$ whenever $\mathcal{U}_f \ni x(n) \to x \in V \backslash \mathcal{U}_f$. Here $\mathcal{U}_f = \mathbb{R}^{d_1} \times \mathcal{U}_g$ and hence $\mathcal{U}_f \ni x(n) \to x \in V\backslash\mathcal{U}_f$ implies $\mathcal{U}_g \ni x'(n) \to x' \in \mathbb{R}^{d_1}\backslash\mathcal{U}_g$. Since $g \in \mathrm{Conv}_1^+(\mathbb{R}^{d-d_1})$, by Proposition 4.30 (iv)', this implies that $|\partial g(x'(n))| \to \infty$ and hence $|\partial f(x(n))| \to \infty$. ∎

4.9 The Gärtner-Ellis theorem

We can finally start reaping the benefits of our study of convex functions. Below is a version of the Gärtner-Ellis theorem. If S is a finite set, then we equip the space $\mathbb{R}^S$ of real functions on S with the standard inner product $\langle x, y\rangle := \sum_{i\in S} x_i y_i$. Recall that $\mathrm{Conv}_1(\mathbb{R}^S)$ is the class of essentially well-behaved convex functions defined in the previous section.

Theorem 4.33 (Gärtner-Ellis) *Let S be a finite set, let μ_n be finite measures on $\mathbb{R}^S$ and let s_n be positive constants such that $s_n \to \infty$. Assume that for each $\lambda \in \mathbb{R}^S$, the limit*

$$\Gamma(\lambda) := \lim_{n\to\infty} \frac{1}{s_n} \log \int_{\mathbb{R}^S} e^{s_n\langle \lambda, x\rangle}\, \mu_n(\mathrm{d}x) \tag{4.15}$$

exists in $[0,\infty]$ *and that* $\Gamma \in \mathrm{Conv}_1(\mathbb{R}^S)$ *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*

$$I(x) := \sup_{\lambda \in \mathbb{R}^S} \big[\langle \lambda, x\rangle - \Gamma(\lambda)\big] \qquad (x \in \mathbb{R}^S).$$

Remark The theorem above is a bit weaker than the usual Gärtner-Ellis theorem as stated, e.g., in [Hol00, Thm V.6], but for our purposes the version above will suffice. Our proof of Theorem 4.33 differs quite significantly from the more traditional proofs which check the large deviations lower and upper bounds of Proposition 1.7.

Remark In the context of the Gärtner-Ellis theorem. we will call the function Γ from (4.15) the *free energy*. In Theorem 5.4 below, where we apply the Gärtner-Ellis theorem to prove a multi-dimensional version of Cramér's theorem, we will see that in the context of i.i.d. random variables, $\Gamma = \log Z$, so our present use of the term "free energy" is a generalization of our use of this term in Section 0.1.

Proof of Theorem 4.33 We start by proving exponential tightness. For each $0 \neq \lambda \in \mathbb{R}^S$ and $c > 0$, let $H_{\lambda,c}$ denote the half-space

$$H_{\lambda,c} := \big\{x \in \mathbb{R}^S : \langle \lambda, x\rangle > c\big\}.$$

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

$$K_{\lambda_1,\ldots,\lambda_n,c} := \mathbb{R}^S \setminus \bigcup_{k=1}^n H_{\lambda_k,c}$$

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 $\lambda' \in \mathbb{R}^S$ 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 3.11. By Theorem 3.7, 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 deviation principle with speed s_n'' and some good rate function J. By Lemma 3.11, 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 3.12), we see that for each $\lambda \in \mathcal{U}_\Gamma$

$$\Gamma(\lambda) = \lim_{n\to\infty} \frac{1}{s_n''} \log \int e^{s_n''\langle \lambda, x\rangle} \mu_n''(\mathrm{d}x) = \sup_{x\in\mathbb{R}^S} \big[\langle \lambda, x\rangle - J(x)\big]. \tag{4.16}$$

Let $g(\lambda)$ be defined by the right-hand side of (4.16). Then $g \in \mathrm{Conv}(\mathbb{R}^S)$ by Lemma 4.10. Lemma 4.26 tells us that Γ is a restriction of g. By assumption $\Gamma \in \mathrm{Conv}_1(\mathbb{R}^S)$. Moreover $\mathcal{U}_\Gamma \neq \emptyset$ so $\mathcal{L}_\Gamma = \mathbb{R}^S$ by Lemma 4.13. This means that we can apply Lemma 4.32 to conclude that Γ is on natural domain. Using Lemma 4.27, we conclude that $\Gamma = g$, so (4.16) holds for all $\lambda \in \mathbb{R}^S$.

Taking the Legendre transform on both sides of (4.16), applying Lemma 4.10, we see that $I = \overline{J}$, where $\overline{J}$ denotes the convex hull of J. Since $\Gamma \in \mathrm{Conv}_1(\mathbb{R}^S)$, Lemma 4.29 tells us that $I \in \mathrm{Conv}_1(\mathbb{R}^S)$. Since $\mathcal{L}_\Gamma = \mathbb{R}^S$, Lemma 4.13 tells us that $\mathcal{F}_I = \{0\}$, so we can apply Exercise 4.20 to conclude that $I = J$. ∎

In applications of Theorem 4.33, we need a practical way to verify that the function Γ defined in (4.15) satisfies $\Gamma \in \mathrm{Conv}_1(\mathbb{R}^S)$ and $0 \in \mathcal{U}_\Gamma$. For this aim, the following proposition will turn out to be handy.

Proposition 4.34 (Dual convex functions) *Let S be a finite set, let $\mathcal{F} \subset \mathbb{R}^S$ be a linear subspace, and let $L : \mathcal{F} \to \mathbb{R}$ be a linear form. Let*

$$\mathcal{A} := \big\{x \in \mathbb{R}^S : \langle \psi, x\rangle = L(\psi)\ \forall \psi \in \mathcal{F}\big\}.$$

Let $\Gamma : \mathbb{R}^S \to \mathbb{R}$ and $I : \mathcal{A} \to (-\infty, \infty]$ be functions, let $\mathcal{D}_I := \{x \in \mathcal{A} : I(x) < \infty\}$, and let $\mathcal{U}_I$ denote the interior of $\mathcal{D}_I$, viewed as a subset of $\mathcal{A}$. Assume that

(i) $\Gamma(\phi+\psi)=\Gamma(\phi)+L(\psi) \qquad (\phi\in\mathbb{R}^S,\ \psi\in\mathcal{F})$.

Assume moreover that there exists a function $\chi:\mathbb{R}^S\to\mathcal{U}_I$ *such that*

(ii) χ *is surjective,*

(iii) $\chi(\phi)=\chi(\psi')$ *if and only if* $\phi-\phi'\in\mathcal{F}$,

(iv) $\langle\phi,x\rangle\le\Gamma(\phi)+I(x)$ *for all* $\phi\in\mathbb{R}^E$ *and* $x\in\mathcal{A}$, *with equality if and only if* $x=\chi(\phi)$.

Then $\Gamma\in\mathrm{Conv}_1(\mathbb{R}^S)$ *and its Legendre transform is the function* $\overline{I}:\mathbb{R}^S\to(-\infty,\infty]$ *defined as* $\overline{I}(x):=I(x)$ *if* $x\in\mathcal{A}$ *and* $:=\infty$ *otherwise.*

Proof We fix an arbitrary $x_\circ\in\mathcal{A}$ and define

$$\begin{aligned} V&:=\big\{x\in\mathbb{R}^S:\langle\psi,x\rangle=0\ \forall\psi\in\mathcal{F}\big\},\\ V^*&:=\mathbb{R}^S/\mathcal{F}. \end{aligned}$$

Let $\underline{\phi}:=\{\phi'\in\mathbb{R}^S:\phi'-\phi\in\mathcal{F}\}$ denote the equivalence class $\underline{\phi}\in\mathbb{R}^S/\mathcal{F}$ containing $\phi\in\mathbb{R}^S$. The spaces V and V^* are dual with respect to the function $\langle\,\cdot\,,\,\cdot\,\rangle$ defined as

$$\langle\underline{\phi},x\rangle:=\langle\phi,x\rangle \qquad (\phi\in\mathbb{R}^S,\ x\in V),$$

where by the definition of V, the right-hand side does not depend on the choice of the representative $\phi\in\mathbb{R}^S$. We define $J:V\to(-\infty,\infty]$ and $\Lambda:V^*\to\mathbb{R}$ by

$$\begin{aligned} J(x)&:=I(x_\circ+x) && (x\in V),\\ \Lambda(\underline{\phi})&:=\Gamma(\phi)-\langle\phi,x_\circ\rangle && (\phi\in\mathbb{R}^S), \end{aligned}$$

where in the second formula, by our assumption (i), the right-hand side does not depend on the choice of the representative ϕ. We set $\mathcal{D}_J:=\{x\in V:J(x)<\infty\}$ and let $\mathcal{U}_J$ denote the interior of $\mathcal{D}_J$. By our assumption (iii), we can unambiguously define a function $\eta:V^*\to\mathcal{U}_J$ by

$$\eta(\underline{\phi}):=\chi(\phi)-x_\circ \qquad (\phi\in\mathbb{R}^S).$$

By assumption (ii), η is a bijection. Our assumption (iv) says that for any $\phi\in\mathbb{R}^S$ and $x\in V$, one has

$$\langle\phi,x_\circ+x\rangle\le\Gamma(\phi)+I(x_\circ+x)=\Lambda(\underline{\phi})+\langle\phi,x_\circ\rangle+J(x)$$

with equality if and only if $x_\circ+x=\chi(\phi)$. Equivalently, this says that

(iv)' $\langle \underline{\phi}, x\rangle \leq \Lambda(\underline{\phi}) + J(x)$ for all $\underline{\phi} \in V^*$ and $x \in V$, with equality if and only if $x = \eta(\phi)$.

Since $\eta : V^* \to \mathcal{U}_J$ is a bijection, we have

$$J(x) \leq \Lambda(\underline{\phi}) - \langle \underline{\phi}, x\rangle \qquad (x \in V,\ \underline{\phi} \in V^*),$$

with equality if and only if $x \in \mathcal{U}_J$ and $\underline{\phi} = \eta^{-1}(x)$. It follows that

$$J(x) = \sup_{\underline{\phi} \in V^*} \big[\Lambda(\underline{\phi}) - \langle \underline{\phi}, x\rangle\big],$$

which shows that $J \in \mathrm{Conv}(V)$ and J is the Legendre transform of Λ. In the same way, we see that $\Lambda \in \mathrm{Conv}(V)$ and Λ is the Legendre transform of J. The condition (iv)' moreover shows that the generalized gradients of J and Λ are given by

$$DJ = \big\{\big(x, \eta^{-1}(x)\big) : x \in \mathcal{U}_J\big\},$$
$$D\Lambda = \big\{\big(\underline{\phi}, \eta(\underline{\phi})\big) : \underline{\phi} \in V^*\big\}.$$

Applying Lemma 4.23, we see that J is continuously differentiable on $\mathcal{U}_J$ and Λ is continuously differentiable on V^*. Moreover, $\partial J(x) = \eta^{-1}(x)$ and $\partial\Lambda(\underline{\phi}) = \eta(\underline{\phi})$. In particular, this shows that η and η^{-1} are continuous, so $\partial\Lambda : V^* \to \mathcal{U}_J$ is a homeomorphism. This proves that $\Lambda \in \mathrm{Conv}_1^+(V^*)$.

It now follows easily that $\Gamma \in \mathrm{Conv}_1(\mathbb{R}^S)$. Moreover, $\mathcal{F} = \mathcal{F}_\Gamma$, the space of flat directions of Γ, and $L : \mathcal{F} \to \mathbb{R}$ is the affine slope of Γ. Letting $\overline{I}$ denote the Legendre transform of Γ, we see from Lemma 4.12 that $\mathcal{A} = \mathcal{A}_{\overline{I}}$, the affine hull of the domain of $\overline{I}$. Using this, it is easy to see that $\overline{I}(x) = I(x)$ if $x \in \mathcal{A}$ and $= \infty$ otherwise. ∎

Chapter 5

Large deviations of i.i.d. random variables

5.1 The multi-dimensional Cramér's theorem

In this chapter we will use the Gärtner-Ellis theorem to prove a number of large deviations results for i.i.d. random variables. Our first aim is to prove a multi-dimensional version of Cramér's theorem. We start by studying the rate function. In the present section, we will also give the proof of Lemma 0.2, which was still outstanding. We first need a multidimensional version of Lemma 2.18.

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.

Lemma 5.1 (Smoothness of the free energy) *Let μ be a probability measure on $\mathbb{R}^d$. Assume that the function*

$$Z(\lambda) := \int e^{\langle\lambda,\, x\rangle}\mu(\mathrm{d}x) \qquad (\lambda\in\mathbb{R}^d). \tag{5.1}$$

satisfies $Z(\lambda) < \infty$ *for all* $\lambda \in \mathbb{R}^d$. *For* $\lambda \in \mathbb{R}$, *let* μ_λ *denote the tilted law*

$$\mu_\lambda(\mathrm{d}x) := \frac{1}{Z(\lambda)} e^{\langle \lambda, x\rangle} \mu(\mathrm{d}x) \qquad (\lambda \in \mathbb{R}^d). \tag{5.2}$$

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

$$\left.\begin{array}{ll} \text{(i)} & \frac{\partial}{\partial \lambda(i)} \log Z(\lambda) = \langle \mu_\lambda \rangle(i), \\ \text{(ii)} & \frac{\partial^2}{\partial \lambda(i) \partial \lambda(j)} \log Z(\lambda) = \mathrm{Cov}_{ij}(\mu_\lambda) \end{array}\right\} \quad (\lambda \in \mathbb{R}^d,\ i,j = 1,\ldots,d). \tag{5.3}$$

Proof The proof is basically the same as in the one-dimensional case (see Lemma 2.18). ∎

Recall that $\mathrm{Conv}_\infty(\mathbb{R}^d)$ is the class of infinitely differentiable, essentially well-behaved convex functions on $\mathbb{R}^d$, defined in Section 4.8.

Lemma 5.2 (The free energy is essentially well-behaved) *Let* μ *be a probability measure on* $\mathbb{R}^d$ *and assume that* Z, *given by (5.1), satisfies* $Z(\lambda) < \infty$ *for all* $\lambda \in \mathbb{R}^d$. *Then* $\log Z \in \mathrm{Conv}_\infty(\mathbb{R}^d)$. *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 first consider the case that $\langle y, \mathrm{Cov}(\mu) y\rangle > 0$ for all $y \in \mathbb{R}^d \backslash \{0\}$. We will check that $\Gamma := \log Z$ satisfies conditions (i), (ii), (iii)', and (iv)' of Proposition 4.30. By assumption $\log Z(\lambda) < \infty$ for all $\lambda \in \mathbb{R}^d$, so $\mathcal{U}_\Gamma = \mathbb{R}^d$ which implies that Γ satisfies conditions (i) and (iv)'. By Lemma 5.1, Γ is infinitely differentiable, so condition (ii) is also satisfied. Lemma 5.1 moreover tells us that $\partial^2 \Gamma(\lambda) = \mathrm{Cov}(\mu_\lambda)$, the covariance matrix of μ_λ defined in (5.2). Therefore, to check condition (iii)', it suffices to show that $\langle y, \mathrm{Cov}(\mu_\lambda) y\rangle > 0$ for all $y \in \mathbb{R}^d \backslash \{0\}$ and $\lambda \in \mathbb{R}^d$. Imagine that $\langle y, \mathrm{Cov}(\mu_\lambda) y\rangle = 0$ for some $y \in \mathbb{R}^d \backslash \{0\}$ and $\lambda \in \mathbb{R}^d$. Let X^λ denote a random variable with law μ^λ and let X^0 denote a random variable with law $\mu^0 = \mu$. Then

$$\mathrm{Var}(\langle X^\lambda, y\rangle) = \sum_{i,j} y_i \mathrm{Cov}(X_i^\lambda, X_j^\lambda) y_j = \langle y, \mathrm{Cov}(\mu_\lambda), y\rangle = 0.$$

Since μ_λ has a density with respect to μ, this implies that $0 = \mathrm{Var}(\langle X^0, y\rangle) = \langle y, \mathrm{Cov}(\mu), y\rangle$, which contradicts our assumption. This completes the proof that $\log Z \in \mathrm{Conv}_\infty^+(\mathbb{R}^d)$ if $\langle y, \mathrm{Cov}(\mu) y\rangle > 0$ for all $y \in \mathbb{R}^d \backslash \{0\}$.

In general, since $\mathrm{Cov}(\mu)$ is a symmetric real matrix, we can choose an orthonormal basis with respect to which it is diagonal. In view of this, we can without loss

of generality assume that $\mathrm{Cov}(\mu)$ is a diagonal matrix and that there exists a $0 \le d' \le d$ such that $\mathrm{Cov}_{ii}(\mu) > 0$ for $1 \le i \le d'$ and $\mathrm{Cov}_{ii}(\mu) = 0$ for $d' < i \le d$. There then exist real constants $a_{d'+1}, \ldots, a_d$ such that a random variable X with law μ satisfies $X_i = a_i$ a.s. for $d' < i \le d$. Let $X' := (X_1, \ldots, X_{d'})$ and let

$$Z'(\lambda_1, \ldots, \lambda_{d'}) := \mathbb{E}\big[e^{\sum_{i=1}^{d'} \lambda_i X'_i}\big] \qquad (\lambda \in \mathbb{R}^{d'})$$

denote its moment generating function. Then, for any $\lambda \in \mathbb{R}^d$,

$$\begin{aligned}\log Z(\lambda_1, \ldots, \lambda_d) &= \log \mathbb{E}\big[e^{\langle \lambda, X\rangle}\big] = \log \mathbb{E}\big[e^{\sum_{i=1}^{d'} \lambda_i X_i}\big] + \sum_{i=d'+1}^{d} \lambda_i a_i \\ &= \log Z'(\lambda_1, \ldots, \lambda_{d'}) + \sum_{i=d'+1}^{d} \lambda_i a_i.\end{aligned}$$

By what we have already proved, $\log Z' \in \mathrm{Conv}^+_\infty(\mathbb{R}^{d'})$, so by Lemma 4.19, we conclude that $\log Z \in \mathrm{Conv}_\infty(\mathbb{R}^d)$. ∎

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. See Figure 5.1 for an illustration.

Lemma 5.3 (Properties of the rate function) *Let μ be a probability measure on $\mathbb{R}^d$. Assume that the moment generating function Z defined in (5.1) is finite for all $\lambda \in \mathbb{R}^d$ and that*

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

For $\lambda \in \mathbb{R}^d$, define μ_λ as in (5.2) 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* $y \ne \langle \mu\rangle$.

(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* $\mathrm{support}(\mu)$.

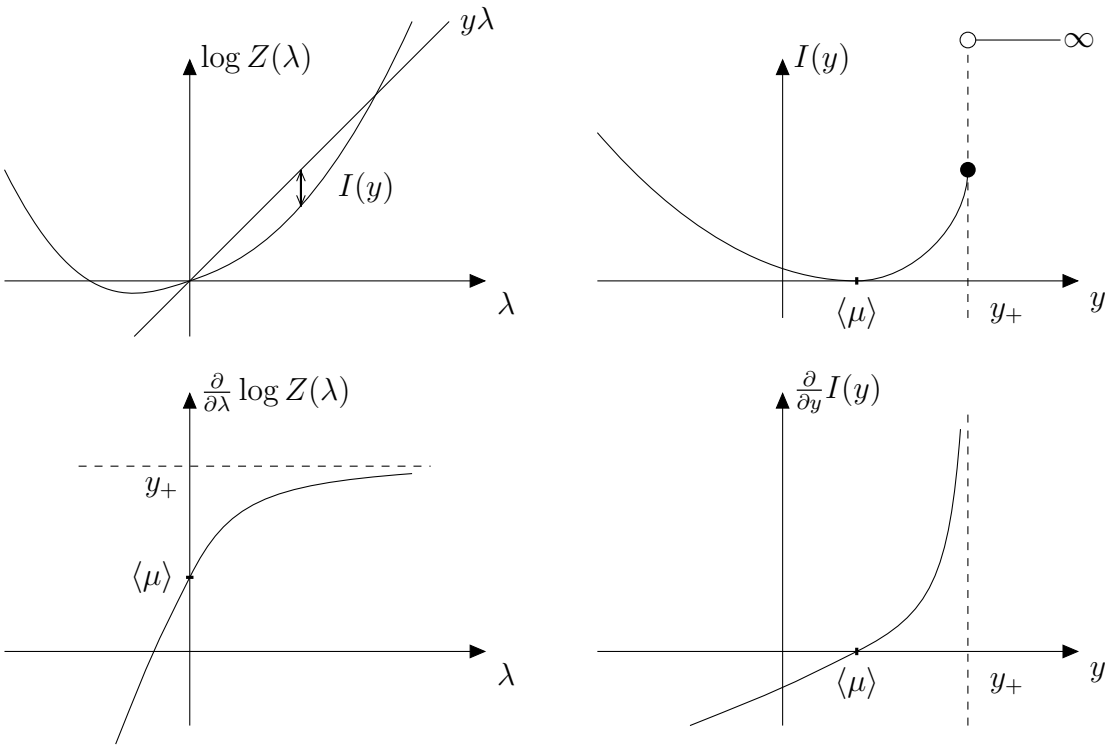

Figure 5.1: Definition of the rate function in Cramér's theorem. The functions below are derivatives of the functions above, and inverses of each other.

(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 In Lemma 5.2 it has been shown that $\log Z \in \mathrm{Conv}^+_\infty(\mathbb{R}^d)$, so Lemma 4.29 implies that $I \in \mathrm{Conv}^+_\infty(\mathbb{R}^d)$, proving (i).

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 4.10 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 5.1, $\partial \log Z(0) = \langle\mu\rangle =: \rho$, 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 $I \in \mathrm{Conv}^+_\infty(\mathbb{R})$, by Proposition 4.30 (iii), I is strictly convex on $\mathcal{U}_I$, 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, proving (iii).

Since $\log Z \in \mathrm{Conv}^+_\infty(\mathbb{R}^d)$, the map $\partial \log Z$ is a homeomorphism from $\mathcal{U}_{\log Z} = \mathbb{R}^d$ to $\mathcal{U}_I$, so property (iv) follows from Lemma 5.1 which tells us that $\partial \log Z(\lambda) = \langle \mu_\lambda \rangle$.

We next prove (v). Since $\mathrm{support}(\mu_\lambda) = \mathrm{support}(\mu)$ for all $\lambda \in \mathbb{R}^d$, it is easy to see that if H is an open half-space such that $H \cap \mathrm{support}(\mu) = \emptyset$, then $\langle \mu_\lambda \rangle \notin H$. Since by (4.2), the complement of $\overline{\mathrm{C}}(\mathrm{support}(\mu))$ is the union of all open half-spaces that do not intersect $\mathrm{support}(\mu)$, this proves the inclusion $\mathcal{U}_I \subset \overline{\mathrm{C}}(\mathrm{support}(\mu))$.

On the other hand, if $H = \{y \in \mathbb{R}^d : \langle \lambda, y \rangle > c\}$ is an open half-space such that $H \cap \mathrm{support}(\mu) \neq \emptyset$, then, in the same way as in Exercise 2.19, one can check that there exists some $r > 0$ large enough such that $\langle \mu_{r\lambda} \rangle \in H$. This proves that $\overline{\mathrm{C}}(\mathcal{U}_I) \supset \overline{\mathrm{C}}(\mathrm{support}(\mu))$. 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 $\overline{\mathcal{U}_I} \supset \overline{\mathrm{C}}(\mathrm{support}(\mu))$, completing our proof.

Since $\log Z \in \mathrm{Conv}^+_\infty(\mathbb{R}^d)$, Lemma 4.31 tells us that 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 Lemma 4.28, 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 5.1, this says that $\langle \mu_{\lambda_\circ} \rangle = y_\circ$, proving (vi). ∎

We also provide the proof of Lemma 0.2 from the introduction, which in the one-dimensional case gives a bit more detail than Lemma 5.3.

Proof of Lemma 0.2 Properties (i), (ii), (vi) follow from Lemma 5.3 (i), properties (iii) and (iv) follow from Lemma 5.3 (ii), and property (v) follows from Lemma 5.3 (vi).

By Lemma 5.3 (i), $I \in \mathrm{Conv}^+_\infty(\mathbb{R})$, so $\partial I : \mathcal{U}_I \to \mathcal{U}_{\log Z}$ is a bijection. Since $\mathcal{U}_I = (y_-, y_+)$ by property (v), and $\mathcal{U}_{\log Z} = \mathbb{R}$, this implies property (viii). The fact that $I'' > 0$ on $\mathcal{U}_I$ follows from Proposition 4.30 (iii)'.

We recall that if f is smooth and strictly increasing and $f(x) = y$, then $\frac{\partial}{\partial x} f(x) = 1/(\frac{\partial}{\partial y} f^{-1}(y))$. Applying this to ∂I and $\partial \log Z$, which are each other's inverses by Lemma 4.28, using the fact that $\partial \log Z(0) = \rho$, and Lemma 2.18, we see 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. ∎

As an application of the Gärtner-Ellis theorem, we can give a quick proof of a multi-dimensional version of Cramér's theorem.

Theorem 5.4 (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 (5.1) 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 good 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 $\Gamma(\lambda) = \log Z(\lambda)$. Indeed, by the independence of the random variables $(X_k)_{k\geq 1}$, we have

$$\frac{1}{n} \log \mathbb{E}\big[e^{\sum_{k=1}^{n} \langle \lambda, X_k\rangle}\big] = \log Z(\lambda)$$

for each n, so the right-hand side of (4.15) is constant as a function of n. It has been proved in Lemma 5.2 that $\log Z \in \mathrm{Conv}_1(\mathbb{R}^d)$, so Theorem 4.33 is applicable and the claim follows. ∎

Remark We recall that elementary properties of the rate function I are listed in Lemma 5.3.

Remark Our proof of Theorem 5.4 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.

5.2 Moderate deviations

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 5.5 (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 good 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 5.4, $\log \int e^{\langle \lambda, x\rangle} \mu_n(\mathrm{d}x) = n \log Z(\lambda)$. 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 2.18 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 (4.15) 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 4.9, the Legendre transform of Γ is the function I defined above. ∎

5.3 Relative entropy

In the final section of this chapter, we will prove Sanov's theorem, which generalizes the Boltzmann-Sanov theorem to Polish spaces. To prepare for this, in the present section, we study the rate function, which is the relative entropy.

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 property of the relative entropy is well-known, and easy to prove.

Lemma 5.6 (Unique minimizer) *For $\mu, \nu \in \mathcal{M}_1(E)$, one has $H(\nu|\mu) \geq 0$ with equality if and only if $\nu = \mu$.*

Proof Define $\Phi : [0, \infty) \to \mathbb{R}$ by

$$\Phi(z) := \int_1^z \mathrm{d}y \int_1^y \mathrm{d}x \, \frac{1}{x} = \begin{cases} z \log z - z + 1 & (z > 0), \\ 1 & (z = 0). \end{cases} \tag{5.4}$$

Then Φ is continuous at 0 and

$$\Phi'(z) = \log z \quad \text{and} \quad \Phi''(z) = \frac{1}{z} \qquad (z > 0).$$

We observe that

$$\int \mathrm{d}\mu \, \Phi\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) = \int \mathrm{d}\nu \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) - \int \mathrm{d}\nu + 1 = H(\nu|\mu).$$

Since $\Phi(1) = 0$ and $\Phi(z) > 0$ for all $z \neq 1$, we see that

$$H(\nu|\mu) = \int \mathrm{d}\mu \, \Phi\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) \geq 0,$$

with equality if and only if $\mathrm{d}\nu/\mathrm{d}\mu = 1$ a.s. w.r.t. μ. ∎

Our next aim is to prove a variational formula for the relative entropy that can be interpreted as a sort of infinite dimensional Legendre transform. In particular, when E is finite, it shows that the functions $H(\,\cdot\,|\mu)$ and Γ_μ are each other's Legendre transforms. As before, 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. For each $\mu \in \mathcal{M}_1(E)$, we define $\Gamma_\mu : B_b(E) \to \mathbb{R}$ by

$$\Gamma_\mu(\phi) := \log Z_\mu(\phi) \quad \text{with} \quad Z_\mu(\phi) := \int_E e^{\phi(x)}\mu(\mathrm{d}x). \tag{5.5}$$

Proposition 5.7 (Variational formula) *Let E be a Polish space and let $\mu, \nu \in \mathcal{M}_1(E)$. Then*

$$H(\nu|\mu) = \sup_{\phi\in B_b(E)} \big[\langle\nu,\phi\rangle - \Gamma_\mu(\phi)\big] = \sup_{\phi\in \mathcal{C}_b(E)} \big[\langle\nu,\phi\rangle - \Gamma_\mu(\phi)\big]. \tag{5.6}$$

The proof of Proposition 5.7 will be split into a number of steps. For each $\phi \in B_b(E)$, we let μ_ϕ denote the probability measure

$$\mu_\phi(\mathrm{d}x) := \frac{1}{Z_\mu(\phi)} e^{\phi(x)}\mu(\mathrm{d}x).$$

The following lemma says that the functions $H(\,\cdot\,|\mu)$ and Γ_μ are dual in a sense that is reminiscent of Lemma 4.24.

Lemma 5.8 (Duality relation) *One has*

$$\langle\nu,\phi\rangle \leq H(\nu|\mu) + \Gamma_\mu(\phi) \qquad \big(\nu \in \mathcal{M}_1(E),\ \phi \in B_b(E)\big), \tag{5.7}$$

with equality if and only if $\nu = \mu_\phi$.

Proof We trivially have a strict inequality 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

$$\begin{aligned}\int \nu(\mathrm{d}x)\phi(x) - H(\nu|\mu) &= \int \nu(\mathrm{d}x)\Big(\log\big(e^{\phi(x)}\big) - \log\big(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x)\big)\Big)\\ &= \int \nu(\mathrm{d}x)\log\Big(e^{\phi(x)}(\frac{\mathrm{d}\nu}{\mathrm{d}\mu})^{-1}(x)\Big) \leq \log\Big(\int \nu(\mathrm{d}x)e^{\phi(x)}h(x)\Big)\\ &\leq \log\Big(\int \mu(\mathrm{d}x)e^{\phi(x)}\Big) = \Gamma_\mu(\phi).\end{aligned}$$

This proves (5.7). 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 $e^{\phi(x)}(\frac{\mathrm{d}\nu}{\mathrm{d}\mu})^{-1}$ is a.s. constant w.r.t. ν. Since the logarithm is a strictly increasing function and e^ϕ 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 (5.7) if and only if $\mu \ll \nu$ and

$$\nu(\mathrm{d}x) = \frac{1}{Z}e^{\phi(x)}\mu(\mathrm{d}x),$$

where Z is some constant. Since ν is a probability measure, we must have $Z = Z(\phi)$. ∎

Lemma 5.9 (First variational formula) *One has*

$$H(\nu|\mu) = \sup_{\phi\in B_b(E)}\big[\langle\nu,\phi\rangle - \Gamma_\mu(\phi)\big] \tag{5.8}$$

Proof Lemma 5.8 implies that

$$H(\nu|\mu) \geq \sup_{\phi\in B_b(E)}\big[\langle\nu,\phi\rangle - \Gamma_\mu(\phi)\big] \qquad \big(\nu \in \mathcal{M}_1(E)\big).$$

We need to show that this is in fact an equality. We first treat the case that $\nu \ll \mu$. Let $\psi := \log(\mathrm{d}\nu/\mathrm{d}\mu)$, so that

$$H(\nu|\mu) = \int \mathrm{d}\mu e^\psi\psi.$$

Note that the function $z \mapsto e^z z$ is bounded from below. We set $\psi_m := m \wedge \psi$ and $\psi_{n,m} := -n \vee \psi_m$, and consider the expression

$$\langle \nu, \psi_{n,m} \rangle - \Gamma_\mu(\psi_{n,m}) = \int \mathrm{d}\mu \, e^{\psi} \psi_{n,m} - \log \int \mathrm{d}\mu \, e^{\psi_{n,m}}.$$

Letting first $n \to \infty$ and then $m \to \infty$, using first dominated convergence and then monotone convergence, we see that our expression converges to $H(\mu_\psi|\mu)$. Since $\psi_{n,m} \in B_b(E)$, this shows that

$$H(\nu|\mu) \leq \sup_{\phi \in B_b(E)} \big[\langle \nu, \phi \rangle - \Gamma_\mu(\phi)\big] \qquad \big(\nu \in \mathcal{M}_1(E),\ \nu \ll \mu\big).$$

If ν is not absolutely continuous w.r.t. μ, then we can find a measurable set A such that $\nu(A) > 0$ but $\mu(A) = 0$. Then

$$\langle \nu, c1_A \rangle - \Gamma_\mu(c1_A) = c\nu(A).$$

Since c is arbitrary, it follows that $H(\nu|\mu) = \infty = \sup_{\phi \in B_b(E)} \big[\langle \nu, \phi \rangle - \Gamma_\mu(\phi)\big]$. ∎

To also prove the second variational formula in Proposition 5.7, we will need an abstract result from measure theory. Let E be a metrizable space and let ϕ_n, ϕ be bounded real functions on E. We write $\phi_n \xrightarrow{\text{bp}} \phi$ if the ϕ_n converge in a *bounded pointwise* way to ϕ, i.e., $\phi_n(x) \to \phi(x)$ for each $x \in E$ and moreover $\sup_n \sup_{x \in E} |\phi_n(x)| < \infty$. We say that a set $\mathcal{A}$ of bounded real functions on E is *closed under bounded pointwise convergence* if $\mathcal{A} \ni \phi_n \xrightarrow{\text{bp}} \phi$ implies $\phi \in \mathcal{A}$. The *bounded pointwise closure* of a set $\mathcal{B} \subset B_b(E)$ is defined as

$$\mathrm{bpclos}(\mathcal{B}) := \bigcap \big\{\mathcal{A} : \mathcal{A} \supset \mathcal{B},\ \mathcal{A} \text{ is closed under bounded pointwise convergence}\big\}.$$

It is easy to see that $\mathrm{bpclos}(\mathcal{B})$ is closed under bounded pointwise convergence and that it is in fact the smallest set containing $\mathcal{B}$ with this property. We note that in general, $\mathrm{bpclos}(\mathcal{B})$ is not the same as

$$\mathrm{bp}(\mathcal{B}) := \big\{\phi \in B_b(E) : \exists\ \phi_n \in \mathcal{B} \text{ s.t. } \phi_n \xrightarrow{\text{bp}} \phi\big\},$$

nor is $\mathrm{bpclos}(\mathcal{B})$ equal to $\mathrm{bp}^2(\mathcal{B}) := \mathrm{bp}(\mathrm{bp}(\mathcal{B}))$, or even to $\bigcup_{n=1}^\infty \mathrm{bp}^n(\mathcal{B})$. This is similar to the σ-field generated by a given collection of sets, which can also not be defined in a constructive way. We cite the following lemma from [EK86, Prop. 3.4.2].

Lemma 5.10 (Bounded pointwise closure) *Let E be a metrizable space. Then* $\mathrm{bpclos}(\mathcal{C}_b(E)) = B_b(E)$.

Remark It is well-known, and not hard to prove, that $B_b(E)$ is closed under bounded pointwise convergence. Since $\mathcal{C}_b(E) \subset B_b(E)$, this immediately implies that $\mathrm{bpclos}(\mathcal{C}_b(E)) \subset B_b(E)$. To prove the other inequality, one first proves that indicator functions of open sets are bounded pointwise limits of continuous functions and then uses the Dynkin class theorem. For details we refer to [EK86, Prop. 3.4.2].

Proof of Proposition 5.7 The first equality in (5.6) has already been proved in Lemma 5.9, so it remains to prove the second equality. We define

$$\left.\begin{aligned} I(\nu) &:= \sup_{\phi\in B_b(E)} \big[\langle\nu,\phi\rangle - \Gamma_\mu(\phi)\big], \\ I'(\nu) &:= \sup_{\phi\in \mathcal{C}_b(E)} \big[\langle\nu,\phi\rangle - \Gamma_\mu(\phi)\big], \end{aligned}\right\} \quad \big(\nu\in\mathcal{M}_1(E)\big).$$

Since the supremum over a smaller set is smaller, we see immediately that $I' \leq I$. Let us define

$$\mathcal{G} := \big\{\phi \in B_b(E) : \langle\nu,\phi\rangle - \Gamma_\mu(\phi) \leq I'(\nu)\ \forall \nu \in \mathcal{M}_1(E)\big\}.$$

Trivially $\mathcal{C}_b(E) \subset \mathcal{G}$. Our aim is to show that $\mathcal{G} = B_b(E)$, which implies $I \leq I'$ and hence $I = I'$. It follows from (5.5) and the dominated convergence theorem that $\phi_n \stackrel{\mathrm{bp}}{\longrightarrow} \phi$ implies $\Gamma_\mu(\phi_n) \to \Gamma_\mu(\phi)$. Also, for each $\nu \in \mathcal{M}_1(E)$, $\phi_n \stackrel{\mathrm{bp}}{\longrightarrow} \phi$ implies $\langle\nu,\phi_n\rangle \to \langle\nu,\phi\rangle$. It follows that the set $\mathcal{G}$ is closed under bounded pointwise convergence, so $\mathcal{G} = B_b(E)$ by Lemma 5.10. ∎

Lemma 5.11 (Good rate function) *For each $\mu \in \mathcal{M}_1(E)$, the function $H(\,\cdot\,|\mu)$ is a good rate function.*

The proof of Lemma 5.11 makes use of the following exercise.

Exercise 5.12 (Conditions for uniform integrability) Let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space. A set C of real measurable functions on Ω is called *uniformly integrable* if for each $\varepsilon > 0$ there exists a $K < \infty$ such that

$$\sup_{f\in C} \int 1_{\{|f|\geq K\}} |f| \mathrm{d}\mu \leq \varepsilon.$$

Show by counterexample that the condition

$$\sup_{f\in C}\int |f|\mathrm{d}\mu < \infty$$

does *not* imply uniform integrability of C. Show on the other hand that if there exists a nonnegative function $\psi : [0,\infty) \to [0,\infty)$ such that $\lim_{r\to\infty} \psi(r)/r = \infty$ and

$$\sup_{f\in C}\int \psi(|f|)\mathrm{d}\mu < \infty,$$

then C is uniformly integrable. (In fact, by the De la Vallée-Poussin theorem, this latter condition is also necessary and ϕ can be chosen such that it is moreover increasing and convex, but we will not need this deeper converse.)

Proof of Lemma 5.11 We must show that for each $r < \infty$, the level set

$$L_r := \big\{\nu \in \mathcal{M}_1(E) : H(\nu|\mu) \le r\big\}$$

is a compact subset of $\mathcal{M}_1(E)$. We observe that for each $\phi \in \mathcal{C}_b(E)$, the function $\nu \mapsto \langle\nu,\phi\rangle - \Gamma_\mu(\phi)$ is continuous with respect to the topology on $\mathcal{M}_1(E)$. Therefore, since the supremum of a collection of continuous functions is lower semi-continuous, Proposition 5.7 implies that $H(\,\cdot\,|\mu)$ is lower semi-continuous and hence L_r is closed. It therefore suffices to show that L_r is relatively compact.

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

$$\{\nu \in \mathcal{M}(E) : \nu \ll \mu\} \cong \{f \in L^1(\mu) : f \ge 0\},$$

and we may identify L_r with the set

$$L_r' := \Big\{f \in L^1(\mu) : f \ge 0,\ \int f\mathrm{d}\mu = 1,\ \int f\log f\mathrm{d}\mu \le r\Big\}.$$

Applying Exercise 5.12 to the function Φ defined in (5.4), we see that the set L_r' is uniformly integrable.

By Prohorov's theorem (Proposition 3.1), 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 $\sup_{\nu\in L_r} \nu(E\backslash D) \le \varepsilon$. Since L_r' is uniformly integrable, we can find a $K < \infty$ such that $\sup_{f\in L_r'} \int 1_{\{f\ge K\}} f\mathrm{d}\mu \le \frac{1}{2}\varepsilon$. 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. ∎

Remark We have used the variational formula for $H(\nu \,|\, \mu)$ (Proposition 5.7) to prove that $\nu \mapsto H(\nu \,|\, \mu)$ is lower semi-continuous with respect to the topology on $\mathcal{M}_1(E)$. It is possible to give a direct proof of this fact, see [DZ93, Lemma 6.2.16], but this is quite involved.

5.4 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].

Theorem 5.13 (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 $\rho_n := \mathbb{P}[M_n \in \cdot\,]$, 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 $H(\,\cdot\,|\mu)$.

Proof We first consider the case that E is compact. In this case, every continuous real function on E is automatically bounded, so we simply write $\mathcal{C}(E)$ instead of $\mathcal{C}_b(E)$. Since $\mathcal{C}(E)$ is separable, we may choose a countable dense set $\{\phi_i : i \in \mathbb{N}_+\} \subset \mathcal{C}(E)$. For each $i \in \mathbb{N}_+$, we define $\Psi_i : \mathcal{M}_1(E) \to \mathbb{R}$ by $\Psi_i(\nu) := \int \phi_i \mathrm{d}\nu$. 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 $\int \phi \mathrm{d}\nu = \int \phi \mathrm{d}\nu'$ for all $\phi \in \mathcal{C}(E)$ by the fact that $\{\phi_i : i \in \mathbb{N}_+\}$ is dense, and therefore $\nu = \nu'$.

We want to apply Theorem 3.21 about projective limits. Since E is compact, the same is true for $\mathcal{M}_1(E)$, so exponential tightness of the measures ρ_n comes for free. For each $m \geq 1$, let $\vec{\Psi}_m : \mathcal{M}_1(E) \to \mathbb{R}^m$ be defined as $\vec{\Psi}_m(x) = \big(\Psi_1(\nu), \ldots, \Psi_m(\nu)\big)$ $(\nu \in \mathcal{M}_1(E))$. We want to check condition (ii) of Theorem 3.21, i.e., we want to show that for each $m \geq 1$, there exists a good rate function I_m on $\mathbb{R}^m$ such that the measures $\rho_n \circ \vec{\Psi}_m^{-1}$ satfisfy the large deviation principle with speed n and rate function I_m. For this aim, we want to apply the Gärtner-Ellis theorem (Theorem 4.33). Letting $\langle \cdot, \cdot \rangle$ denote the standard inner product on $\mathbb{R}^m$, we observe that for each $\lambda \in \mathbb{R}^m$, one has

$$\begin{aligned}
&\frac{1}{n} \log \int_{\mathbb{R}^m} e^{n\langle \lambda, x\rangle} \rho_n \circ \vec{\Psi}_m^{-1}(\mathrm{d}x) = \frac{1}{n} \log \int_{\mathbb{R}^m} e^{n \sum_{i=1}^m \lambda_i \Psi_i(\nu)} \rho_n(\mathrm{d}\nu) \\
&\quad = \frac{1}{n} \log \mathbb{E}\big[e^{\sum_{k=1}^n \sum_{i=1}^m \lambda_i \Psi_i(\delta_{X_k})}\big] = \frac{1}{n} \log \mathbb{E}\big[e^{\sum_{k=1}^n \sum_{i=1}^m \lambda_i \phi_i(X_k)}\big] \\
&\quad = \frac{1}{n} \log \prod_{k=1}^n \mathbb{E}\big[e^{\sum_{i=1}^m \lambda_i \phi_i(X_k)}\big] = \log \mathbb{E}\big[e^{\sum_{i=1}^m \lambda_i \phi_i(X_1)}\big] = \Gamma_\mu\Big(\sum_{i=1}^m \lambda_i \phi_i\Big),
\end{aligned}$$

where Γ_μ is defined in (5.5). Applying Lemma 5.2 to the image measure $\mu \circ \vec{\Psi}_m^{-1}$, we see that the function

$$\mathbb{R}^m \ni \lambda \mapsto \Gamma_\mu\Big(\sum_{i=1}^m \lambda_i \phi_i\Big) \in [0, \infty)$$

is an element of $\mathrm{Conv}_\infty(\mathbb{R}^m)$, so the Gärtner-Ellis theorem (Theorem 4.33) is applicable and tells us that the the measures $\rho_n \circ \vec{\Psi}_m^{-1}$ satisfy the large deviation principle with speed n and rate function I_m given by

$$I_m(x) = \sup_{\lambda \in \mathbb{R}^m} \Big[\langle \lambda, x\rangle - \Gamma_\mu\Big(\sum_{i=1}^m \lambda_i \phi_i\Big)\Big] \qquad (x \in \mathbb{R}^m).$$

We have now also checked condition (ii) of Theorem 3.21 about projective limits, so we can use that theorem to conclude that the measures ρ_n satfisfy the large deviation principle with speed n and some good rate function I. Lemma 3.22 moreover tells us that

$$I_m(\vec{\Psi}_m(\nu)) \uparrow I(\nu) \quad \text{as } m \uparrow \infty \qquad \big(\nu \in \mathcal{M}_1(E)\big).$$

Let Φ_m denote the linear span of the functions $\phi_1, \ldots, \phi_m$ and let Φ_∞ denote the

linear span of the functions $\{\phi_i : i \in \mathbb{N}_+\}$. Then

$$I_m(\vec{\Psi}_m(\nu)) = \sup_{\lambda\in\mathbb{R}^m} \big[\sum_{i=1}^m \lambda_i \Psi_i(\nu) - \Gamma_\mu\big(\sum_{i=1}^m \lambda_i\phi_i\big)\big]$$
$$= \sup_{\lambda\in\mathbb{R}^m} \big[\sum_{i=1}^m \lambda_i \int \phi_i \,\mathrm{d}\nu - \Gamma_\mu\big(\sum_{i=1}^m \lambda_i\phi_i\big)\big] = \sup_{\phi\in\Phi_m} \big[\int \phi \,\mathrm{d}\nu - \Gamma_\mu(\phi)\big],$$

so our previous formula implies that

$$I(\nu) = \sup_{\phi\in\Phi_\infty} \big[\langle\nu,\phi\rangle - \Gamma_\mu(\phi)\big] \qquad \big(\nu \in \mathcal{M}_1(E)\big),$$

where as in the previous subsection we adopt the notation $\langle\nu,\phi\rangle := \int\phi\,\mathrm{d}\nu$. Since Φ_∞ is dense in $\mathcal{C}(E)$ and the functions $\phi \mapsto \langle\nu,\phi\rangle$ and $\phi \mapsto \Gamma_\mu(\phi)$ are continuous, we can with the help of Proposition 5.7 conclude that

$$I(\nu) = \sup_{\phi\in\mathcal{C}(E)} \big[\langle\nu,\phi\rangle - \Gamma_\mu(\phi)\big] = H(\nu|\mu) \qquad \big(\nu \in \mathcal{M}_1(E)\big).$$

This concludes the proof in the special case that E is compact.

To prove the general statement, let $\overline{E}$ be a metrizable compactification of E. By Proposition 3.9, such a compactification exists and E is a G_δ-subset of E. By what we have already proved, the laws ρ_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 3.10, 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 1.16), using the fact that $H(\,\cdot\,|\mu)$ is a good rate function (which has been proved in Lemma 5.11) and the fact that $H(\,\cdot\,|\,\mu) = \infty$ on $\mathcal{M}_1(\overline{E})\backslash\mathcal{M}_1(E)$, we conclude that the laws ρ_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 5.13 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]. Another very elegant proof of Sanov's theorem can be found in [Csi06].

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.

[DV75a] M.D. Donsker and S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. I. *Commun. Pure Appl. Math.* 28 (1975), 1–47.

[DV75b] M.D. Donsker and S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. II. *Commun. Pure Appl. Math.* 28 (1975), 279–301.

[DV76] M.D. Donsker and S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. II. *Commun. Pure Appl. Math.* 29 (1976), 389–461.

[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