UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in Matematica

Master's Thesis

LARGE DEVIATIONS

Author:
Elia Bisi

Supervisor:
Prof. Francesco Caravenna

Academic year 2012-2013

To my uncle Mario

Statistically, the probability of any one of us being here is so small that you'd think the mere fact of existing would keep us all in a contented dazzlement of surprise. We are alive against the stupendous odds of genetics, infinitely outnumbered by all the alternates who might, except for luck, be in our places.

Even more astounding is our statistical improbability in physical terms. The normal, predictable state of matter throughout the universe is randomness, a relaxed sort of equilibrium, with atoms and their particles scattered around in an amorphous muddle. We, in brilliant contrast, are completely organized structures, squirming with information at every covalent bond. We make our living by catching electrons at the moment of their excitement by solar photons, swiping the energy released at the instant of each jump and storing it up in intricate loops for ourselves. We violate probability, by our nature. To be able to do this systemically, and in such wild varieties of form, from viruses to whales, is extremely unlikely; to have sustained the effort successfully for the several billion years of our existence, without drifting back into randomness, was nearly a mathematical impossibility.

— Lewis Thomas, *The Lives of a Cell: Notes of a Biology Watcher* (1974)

Introduction

Large deviations theory is a part of probability theory that studies "rare events". One might naïvely think that it is not very interesting to deal with events whose probability is very small; however, rare events might be of great importance, as the following list of "real world" examples shows:

- a gambler hits the jackpot;
- an insurance company goes bankrupt;
- an investment portfolio has a large loss;
- a statistical estimator gives a wrong information;
- a physical or chemical system shows an atypical configuration;
- a buffer overflow occurs in a computer software.

In all these examples, we observe an *atypical* situation, that is a deviation from the average behavior.

In basic probability theory, one is used to study the *typical* behavior of a model: the first important example is the empirical average of a sequence $\{X_i\}_{i\geq 1}$ of independent and identically distributed real-valued random variables (copies of a random variable X), i.e. $\frac{S_n}{n}$, where $S_n = X_1 + \cdots + X_n$ is the nth partial sum of the random variables. In this respect, the two basic results are the law of large numbers, which states that $\frac{S_n}{n}$ approximates asymptotically the actual mean $\overline{x} = \mathbb{E}(X)$, and the central limit theorem, which describes the limit behavior of the probability distribution of $\frac{S_n}{n}$, suitably rescaled. For example, if $\varepsilon > 0$, by the (weak) law of large numbers

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \overline{x}\right| > \varepsilon\right) \xrightarrow{n\to\infty} 0,$$

i.e. the probability that $\frac{S_n}{n}$ is *not* in any fixed small neighborhood of $\overline{x}$ is asymptotically vanishing. The event $\left\{\left|\frac{S_n}{n} - \overline{x}\right| > \varepsilon\right\}$, which is "rare" since its probability converges to 0 as $n \to \infty$, is an example of large deviations event. The Swedish mathematician, statistician and actuary Harald Cramér, who was interested in estimating the decay of the probability of such an event, proved in 1938 the following result:

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \overline{x}\right| \geq \varepsilon\right) = e^{-nc+o(n)} \qquad \text{as } n \to \infty, \tag{1}$$

where c is a positive constant that of course depends on ε and on the law of X; in other words, the decay of the probability is exponential in n. An interesting answer is then: how does the constant c depend on ε and on the law of X? The answer to this problem involves the study of the so-called rate function, which, in turn, relates to the exponential moments of X. The estimate (1) is known as Cramér's Theorem and it is the subject of chapter 1 of this work, together with the analysis of the properties of the rate function.

Analyzing large deviations probabilities in the case of empirical averages leads us to look for an efficient and handy way to state results such as Cramér's Theorem in more general contexts. Thus in chapter 2 we deal with the abstract theory of large deviations, which dates back to the fundamental papers by Varadhan[†] and Donsker in the 1960s and 1970s. This theory allows to unify and extend Cramér's Theorem and other important results under a suitable theoretical framework: essentially, a *large deviations principle* is a property of a sequence of probability measures that on certain sets decay exponentially with respect to a given *rate* (a positive sequence that gives the scale of the exponential decay) and a *rate function* (a nonnegative function that describes how the exponential decay varies among sets). In the case of Cramér's Theorem, such measures are the laws of the empirical averages and the rate is n. A sequence of probability measures satisfying a large deviations principle in general needs to be defined on a *topological space*, and some topological restrictions are required to develop the theory: the handiest environment to work with is a separable completely metrizable topological space (i.e. a so-called *Polish space*). In this abstract framework, it is possible to prove some general results that can be applied in a variety of contexts. In the course of the chapter, we always highlight the strong formal analogy between large deviations principle and weak convergence of a sequence of probability measures: in many cases, a weak convergence result has a large deviations analogue (which is often more difficult to prove).

Chapter 3 is devoted to discuss the two most important large deviations results about sequences of independent and identically distributed random variables. The first one is Cramér's Theorem for empirical averages, which we can reformulate in the light of abstract theory, proving a strengthened version of it: the laws of the empirical averages satisfy a large deviations principle. We then extend it to multivariate random variables (Cramér's Theorem in $\mathbb{R}^d$): this is the only part of our work where we do not give complete proofs, for the sake of conciseness. The second important result concerns *empirical distributions*, defined as follows. If $\{X_i\}_{i\in\mathbb{N}}$ is a sequence of independent and identically distributed random variables (with mean

[†]Varadhan was awarded in 2007 the prestigious Abel prize "for his fundamental contributions to probability theory and in particular for creating a unified theory of large deviations".

$\overline{x}$ and common distribution μ) taking values in a Polish space $\mathbf{X}$, the empirical distributions are

$$L_n \doteq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i},$$

viewed as random distributions (here δ_x is the Dirac measure concentrated on x). Empirical distributions are important, for instance, in statistics, since they approximate the actual distribution of the X_i: just as the empirical averages converge to $\overline{x}$ almost surely, similarly the empirical distributions converge weakly to μ almost surely (this is the strong law of large numbers for empirical distributions). Sanov's Theorem is the fundamental large deviations result that we present about empirical distributions: it states that their laws satisfy a large deviations principle. A discrete version of this theorem goes back to Boltzmann, who in 1877 in a work on entropy carried out a fundamental calculation which "represents a revolutionary moment in human culture during which both statistical mechanics and the theory of large deviations were born" (see [Ell]). The rate function for the large deviations principle satisfied by the empirical distributions is the so-called *relative entropy*, a fundamental concept in statistical mechanics as well as in information theory.

A fascinating aspect of large deviations theory is that it involves several fields of mathematics. Probability theory is surely the most important, but topology results, both elementary and advanced, are often required, as well as tools of convex analysis, variational calculus, and even functional analysis. This deep mathematical structure is balanced by a rich variety of applications, both inside and outside of mathematics: also for this reason large deviations theory is one of the most active research topics in probability theory. Applied probability, statistics, engineering, queuing theory, information theory, operations research, statistical mechanics, chemistry, financial mathematics and actuarial risk theory are some fields of application. Our work is principally focused on the theoretical aspects, however in chapter 4 we have also included three significant examples where large deviations theory plays an important role. In the first application, in actuarial risk theory, we derive the Cramér-Lundberg inequality for the ruin probability of an insurance company. The second example concerns statistical hypothesis testing about the probability distribution of a random variable: we obtain the asymptotical optimality of Neyman-Pearson decision tests. Finally, we analyze the Curie-Weiss model in statistical mechanics, giving a basic interpretation of the ferromagnetic behavior of certain materials.

CONTENTS

Chapter 1

Cramér's Theorem

Suppose we are an insurance company and we should choose the premium x for the n holders of our insurance policy in a given year. Let X_i denote the claim amount of policy holder i, for $i = 1,\dots,n$, and $S_n = \sum_{i=1}^n X_i$ the total claim amount. We assume that all the risks of loss, covered by the insurance company, of the policy holders are 'equivalent' and independent from each other, so that X_i is a sequence of independent and identically distributed (i.i.d.) random variables, which are copies of X. First of all, it is natural to choose $x > \mathbb{E}(X)$. Assuming that the number n of policy holders is very large, we would like to obtain an asymptotic estimation of $\mathbb{P}(S_n \geq nx)$, the probability that the total claim amount payed out exceeds the total amount of premiums collected. We shall choose the premium x so that this probability is small enough.

The great Swedish mathematician, statistician and actuary Harald Cramér[†] was interested in these insurance modeling problems and he was the first, in 1938, to estimate $\mathbb{P}(S_n \geq nx)$ asymptotically, where S_n is a sum of i.i.d. random variables and x is larger than their mean (see [Cra]). The result of this investigation is known as Cramér's Theorem and it is the subject of this chapter, which is mainly based on [Swa, §2.1 and §2.2] and [Hol, ch.I].

1.1 First examples

Let $(\Omega,\mathcal{E},\mathbb{P})$ be a probability space and $X \in L^2(\Omega,\mathcal{E},\mathbb{P})$, with

$$\overline{x} \doteq \mathbb{E}(X), \qquad \sigma^2 \doteq \mathrm{Var}(X).$$

[†]To understand the key role that Cramér had in putting probability theory on firm mathematical foundations, let us refer to a statement he made in 1926, when this subject was not yet an accepted branch of mathematics: "The probability concept should be introduced by a purely mathematical definition, from which its fundamental properties and the classical theorems are deduced by purely mathematical operations." (see [Blo]).

Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. copies of X, and for all $n \geq 1$ let

$$S_n \doteq \sum_{i=1}^{n} X_i$$

be the random variables that represent their partial sums. The two standard results in probability theory about the so-called *empirical averages* $\left\{\frac{S_n}{n}\right\}_{n\geq 1}$ are:

- Strong law of large numbers:

$$\frac{S_n}{n} \xrightarrow{n\to\infty} \overline{x} \qquad \mathbb{P}\text{-a.s..}$$

- Central limit theorem:

$$\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \overline{x}\right) \xrightarrow{n\to\infty} Z, \qquad Z \sim \mathcal{N}(0,1),$$

  where $\rightharpoonup$ denotes weak convergence.

The central limit theorem gives us information about some kind of deviations of the empirical average $\frac{S_n}{n}$ from the mean $\overline{x}$. More precisely, the probability that the empirical average deviates from the mean by an amount of $c/\sqrt{n}$ (where c is a positive constant) converges to a fixed number $\neq 0, 1$, which is related to the standard normal distribution function:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \overline{x}\right| \geq \frac{c}{\sqrt{n}}\right) = \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - \overline{x}\right)\right| \geq \frac{c}{\sigma}\right) = \mathbb{P}\left(|Z| \geq \frac{c}{\sigma}\right).$$

If we replace the asymptotically small amount $c/\sqrt{n}$ with a fixed constant $\varepsilon > 0$, we obtain a more unlikely event (for large n). Indeed, since almost sure convergence implies convergence in probability, the strong law of large numbers guarantees that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \overline{x}\right| \geq \varepsilon\right) = 0.$$

Such events are 'rare', because their probability tends to 0 as $n \to \infty$, hence these deviations of the empirical average from the mean are said to be *large*. The main aim of large deviations theory is to estimate the rate of decay of the probability of these ´rare' events. Cramér's Theorem, as we will see, states that the decay is exponential in n, in the sense that

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \overline{x}\right| \geq \varepsilon\right) = e^{-nc+o(n)},$$

where c is a constant that depends on ε and on the law of X. We now illustrate two examples of explicit calculation of this exponential decay, before going into the analysis of the general case.

Example 1.1. We consider a sequence of tosses of a fair coin. Let X_i be the random variable that takes either the value 0 or 1 according to whether the result of the toss i is tail or head; thus, $\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \frac{1}{2}$ and $\mathbb{E}(X_i) = \frac{1}{2}$. Since $S_n \le n$, if $x > 1$ then $\mathbb{P}\Big(\frac{S_n}{n} \ge x\Big) = 0$. Let $x \in \Big(\frac{1}{2}, 1\Big]$. Since $S_n \sim B\Big(n, \frac{1}{2}\Big)$,

$$\mathbb{P}\Big(\frac{S_n}{n} \ge x\Big) = \sum_{\lceil nx \rceil \le k \le n} \binom{n}{k} \frac{1}{2^n},$$

where $\lceil nx \rceil$ is the smallest integer $\ge nx$. Therefore, we obtain the estimation

$$2^{-n} Q_n(x) \le \mathbb{P}\Big(\frac{S_n}{n} \ge x\Big) \le 2^{-n}(n+1) Q_n(x),$$

where

$$Q_n(x) \doteq \max_{\lceil nx \rceil \le k \le n} \binom{n}{k}.$$

This proves that

$$-\log 2 + \frac{1}{n} \log Q_n(x) \le \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \ge x\Big) \le -\log 2 + \frac{\log(n+1)}{n} + \frac{1}{n} \log Q_n(x). \tag{1.1}$$

Since $x > \frac{1}{2}$

$$Q_n(x) = \binom{n}{\lceil nx \rceil},$$

and by Stirling's formula $n! = n^n e^{-n} \sqrt{2\pi n}(1 + o(1))$, hence

$$\begin{aligned}
&\frac{1}{n} \log Q_n(x) = \frac{1}{n} \log \binom{n}{\lceil nx \rceil} = \frac{1}{n} \log \frac{n!}{\lceil nx \rceil!(n - \lceil nx \rceil)!} \\
&= \frac{1}{n} \log\left(\frac{n^n e^{-n} \sqrt{2\pi n}(1 + o(1))}{\lceil nx \rceil^{\lceil nx \rceil} e^{-\lceil nx \rceil} \sqrt{2\pi \lceil nx \rceil}(n - \lceil nx \rceil)^{(n - \lceil nx \rceil)} e^{-(n - \lceil nx \rceil)} \sqrt{2\pi(n - \lceil nx \rceil)}(1 + o(1))} \right) \\
&= \frac{1}{n}\left(\log\left(\left(\frac{\lceil nx \rceil}{n}\right)^{-\lceil nx \rceil} \left(\frac{n - \lceil nx \rceil}{n}\right)^{-(n - \lceil nx \rceil)} \right) + \log(1 + o(1)) \right. \\
&\quad \left. + \frac{1}{2}\log(2\pi n) - \frac{1}{2}\log(2\pi \lceil nx \rceil) - \frac{1}{2}\log(2\pi(n - \lceil nx \rceil)) - \log(1 + o(1)) \right) \\
&= \left(-\frac{\lceil nx \rceil}{n} \log\left(\frac{\lceil nx \rceil}{n}\right) - \frac{n - \lceil nx \rceil}{n} \log\left(\frac{n - \lceil nx \rceil}{n}\right) + o(1) \right) \\
&\xrightarrow{n \to \infty} -x \log x - (1 - x)\log(1 - x),
\end{aligned}$$

since $\frac{\lceil nx \rceil}{n} \xrightarrow{n \to \infty} x$ (indeed, $nx \le \lceil nx \rceil \le nx + 1$). Therefore, passing to the limit as $n \to \infty$ in (1.1), we obtain

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \ge x\Big) = -(\log 2 + x \log x + (1 - x)\log(1 - x)). \qquad \diamond$$

Example 1.2. Assume that X_i are i.i.d. standard normal random variables, i.e. they are absolutely continuous with density function

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{t^2}{2}\Big).$$

Then $\frac{S_n}{\sqrt{n}}$ is Standard Normal again, hence if $x > 0$

$$\mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) = \mathbb{P}\Big(\frac{S_n}{\sqrt{n}} \geq \sqrt{n}x\Big) = \int_{\sqrt{n}x}^{+\infty} f(t)\,dt.$$

We first note that $f'(t) = -tf(t)$, hence for all $k \in \mathbb{Z} \setminus \{0\}$, $t > 0$

$$\frac{d}{dt}(t^k f(t)) = kt^{k-1} f(t) + t^k(-tf(t)) = f(t)(kt^{k-1} - t^{k+1}).$$

If $y > 0$, we have

$$\begin{aligned}
(y^{-1} - y^{-3}) f(y) = [(t^{-3} - t^{-1}) f(t)]_y^{+\infty} &= \int_y^{+\infty} f(t)(-3t^{-4} - t^{-2} - (-t^{-2} - 1))\,dt \\
&= \int_y^{+\infty} f(t)(1 - 3t^{-4})\,dt \leq \int_y^{+\infty} f(t)\,dt, \\
y^{-1} f(y) = -[t^{-1} f(t)]_y^{+\infty} &= -\int_y^{+\infty} f(t)(-t^{-2} - 1)\,dt \\
&= \int_y^{+\infty} f(t)(1 + t^{-2})\,dt \geq \int_y^{+\infty} f(t)\,dt,
\end{aligned}$$

hence

$$\frac{y^2 - 1}{y^3} f(y) \leq \int_y^{+\infty} f(t)\,dt \leq \frac{1}{y} f(y). \tag{1.2}$$

If we set $y \doteq \sqrt{n}x$, we have

$$\int_y^{+\infty} f(t)\,dt = \mathbb{P}\Big(\frac{S_n}{n} \geq x\Big),$$

and as $n \to \infty$

$$\begin{aligned}\frac{y^2-1}{y^3} f(y) &= \frac{1}{\sqrt{2\pi}} \frac{nx^2-1}{n^{3/2}x^3} \exp\Big(-n\frac{x^2}{2}\Big)\\ &= \exp\Big(-n\frac{x^2}{2} + \log\Big(\frac{1}{\sqrt{2\pi}} \frac{nx^2-1}{n^{3/2}x^3}\Big)\Big)\\ &= \exp\Big(-n\frac{x^2}{2} + o(n)\Big),\\ \frac{1}{y} f(y) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n}x} \exp\Big(-n\frac{x^2}{2}\Big)\\ &= \exp\Big(-n\frac{x^2}{2} + \log\Big(\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n}x}\Big)\Big)\\ &= \exp\Big(-n\frac{x^2}{2} + o(n)\Big).\end{aligned}$$

Thus, by (1.2), $\mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) = \exp\Big(-n\frac{x^2}{2} + o(n)\Big)$, i.e.

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) = -\frac{x^2}{2}$$

for all $x > 0$. ◊

1.2 The Legendre transform

In Cramér's Theorem, we will run into the Legendre transform of the so-called logarithmic moment generating function. Therefore, in this section we are interested in defining the Legendre transform and studying its properties.

Definition 1.3. The *Legendre transform* of $f\colon \mathbb{R} \to \overline{\mathbb{R}}$ is the function $f^*\colon \mathbb{R} \to \overline{\mathbb{R}}$ defined by

$$f^*(x) = \sup_{t\in\mathbb{R}} [tx - f(t)]. \tag{1.3}$$

We now recall the definitions of convex set, convex function and concave function.

Definition 1.4. Let $\mathbf{X}$ be a real vector space. A subset $C \subseteq \mathbf{X}$ is said to be *convex* if

$$px_1 + (1-p)x_2 \in C \qquad \forall x_1, x_2 \in C, \quad \forall p \in (0,1).$$

Let $C \subseteq \mathbf{X}$ be convex.

- A function $f\colon C \to (-\infty, +\infty]$ is said to be *convex* if

  $$f(px_1 + (1-p)x_2) \leq pf(x_1) + (1-p)f(x_2) \qquad \forall x_1, x_2 \in C, \quad \forall p \in (0,1).$$

  In addition, if the latter inequality is strict for all $x_1, x_2 \in C$ such that $x_1 \neq x_2$ and for all $p \in (0,1)$, f is said to be *strictly convex*.

- A function $f : C \to [-\infty, +\infty)$ is said to be *concave* (resp., *strictly concave*) if $-f$ is convex (resp., $-f$ is strictly convex).

Remark 1.5. We will need to deal with convex and concave functions that may also take infinite values. However, in our definition we assume that convex functions never take $-\infty$, since $pf(x_1) + (1-p)f(x_2)$ is not well-defined if $f(x_1) = +\infty$ and $f(x_2) = -\infty$ (or vice versa). As a consequence, we assume that concave functions never take $+\infty$. ◊

Definition 1.6. A function $f : \mathbb{R} \to \overline{\mathbb{R}}$ is said to be *lower semi-continuous* if it satisfies one of the following equivalent conditions:

(i) For all $a \in \mathbb{R}$, the level set $\{t \in \mathbb{R} : \ f(t) \le a\}$ is closed.

(ii) For all $t_n \to t$, $\liminf_{n\to\infty} f(t_n) \ge f(t)$.

We will examine the properties of lower semi-continuous functions (including the equivalence of the two conditions of our definition) in detail in chapter 2, because they are a fundamental tool in the abstract large deviations theory.

Lemma 1.7. The Legendre transform f^* of any function $f \not\equiv +\infty$ is convex and lower semi-continuous.

Proof. Since f is not identically $+\infty$, $tx - f(t) > -\infty$ for some $t \in \mathbb{R}$ and for all $x \in \mathbb{R}$, hence $f^*(x) > -\infty$ for all $x \in \mathbb{R}$. Moreover, if $x_1, x_2 \in \mathbb{R}$ and $p \in (0,1)$, then for all $t \in \mathbb{R}$

$$t(px_1 + (1-p)x_2) - f(t) = p(tx_1 - f(t)) + (1-p)(tx_2 - f(t)) \le pf^*(x_1) + (1-p)f^*(x_2),$$

hence, passing to the supremum over all $t \in \mathbb{R}$, we obtain

$$f^*(px_1 + (1-p)x_2) \le pf^*(x_1) + (1-p)f^*(x_2).$$

This proves that f^* is convex.

Let now $x \in \mathbb{R}$ and assume that $x_n \to x$. Then $\forall t \in \mathbb{R}$

$$\liminf_{n\to\infty} f^*(x_n) \ge \liminf_{n\to\infty} [tx_n - f(t)] = tx - f(t),$$

hence, passing to the supremum over all $t \in \mathbb{R}$, we obtain

$$\liminf_{n\to\infty} f^*(x_n) \ge \sup_{t\in\mathbb{R}} [tx - f(t)] = f^*(x).$$

This proves that f^* is lower semi-continuous. □

Definition 1.8. Let $\mathbf{X}$ be a set. We call *(effective) domain* of a function $f : \mathbf{X} \to \overline{\mathbb{R}}$ the set

$$\mathcal{D}_f = \{x \in \mathbf{X} : \ f(x) \in \mathbb{R}\}.$$

Remark 1.9. If $\mathbf{X}$ is a real vector space and $f: \mathbf{X} \to (-\infty, +\infty]$ is convex, then $\mathcal{D}_f$ is a convex subset of $\mathbf{X}$. Indeed, if $x_1, x_2 \in \mathcal{D}_f$ and $p \in (0,1)$, since f is convex and $f(x_1), f(x_2) < +\infty$ we obtain

$$f(px_1 + (1-p)x_2) \le pf(x_1) + (1-p)f(x_2) < +\infty.$$

Therefore, $px_1 + (1-p)x_2 \in \mathcal{D}_f$ and $\mathcal{D}_f$ is convex. In particular, if $\mathbf{X} = \mathbb{R}$ then $\mathcal{D}_f$ is an interval with endpoints t_- and t_+ for some $-\infty \le t_- < t_+ \le +\infty$. ◊

Lemma 1.10. Let $f: \mathbb{R} \to (-\infty, +\infty]$ be convex and lower semi-continuous. Then f is continuous on $\overline{\mathcal{D}}_f$; in particular, if t_- and t_+ are the endpoints of $\mathcal{D}_f$,

(i) If $t_+ < +\infty$, then $\lim_{t \nearrow t_+} f(t) = f(t_+)$.

(ii) If $t_- > -\infty$, then $\lim_{t \searrow t_-} f(t) = f(t_-)$.

Proof. It is well-known that a convex *real*-valued function defined on an open interval is continuous, hence f is continuous on $\mathring{\mathcal{D}}_f = (t_-, t_+)$. We only have to show that f is continuous on the boundary of (t_-, t_+), i.e. (i) and (ii). Assume that $t_+ < +\infty$. Since f is lower semi-continuous, $\liminf_{t \nearrow t_+} f(t) \ge f(t_+)$. Since f is convex, if we fix $t_0 \in (t_-, t_+)$, we have

$$f(pt_0 + (1-p)t_+) \le pf(t_0) + (1-p)f(t_+) \qquad \forall p \in (0,1).$$

Passing to the limit superior as $p \searrow 0$,

$$\limsup_{t \nearrow t_+} f(t) = \limsup_{p \searrow 0} f(pt_0 + (1-p)t_+) \le f(t_+).$$

We conclude that $\lim_{t \nearrow t_+} f(t) = f(t_+)$. The proof of the second statement is similar. □

Definition 1.11. A function $f: \mathbb{R} \to \overline{\mathbb{R}}$ is said to be *steep* if f is differentiable on $\mathring{\mathcal{D}}_f$ and for all $t_0 \in \partial \mathring{\mathcal{D}}_f$

$$\lim_{t \to t_0,\, t \in \mathring{\mathcal{D}}_f} |f'(t)| = +\infty.$$

We now introduce the function space Conv_∞, which includes the logarithmic moment generating function of any random variable (under certain hypotheses concerning its law) and its Legendre transform, as we will see later.

Definition 1.12. We denote by Conv_∞ the space of the functions $f: \mathbb{R} \to (-\infty, +\infty]$ such that

(i) $\mathring{\mathcal{D}}_f \ne \varnothing$.

(ii) f is lower semi-continuous.

(iii) $f \in C^\infty(\mathring{\mathcal{D}}_f)$.

(iv) f is convex and $f'' > 0$ on $\mathring{\mathcal{D}}_f$.

(v) f is steep.

Remark 1.13. If $f \in \mathrm{Conv}_\infty$, in particular f is convex, therefore by Remark 1.9 $\mathring{\mathcal{D}}_f = (t_-, t_+)$, with $-\infty \le t_- < t_+ \le +\infty$. In particular, the steepness of f means that

$$\begin{aligned} t_+ < +\infty \quad &\Longrightarrow \quad \lim_{t \nearrow t_+} f'(t) = +\infty, \\ t_- > -\infty \quad &\Longrightarrow \quad \lim_{t \searrow t_-} f'(t) = -\infty. \end{aligned}$$

We stress that, if $t_+ = +\infty$ (or $t_- = -\infty$), we do *not* require anything about the limit of $f'(t)$ as $t \to t_+ = +\infty$ (as $t \to t_- = -\infty$, respectively): indeed, in this case $t_+ \notin \partial\mathring{\mathcal{D}}_f$ ($t_- \notin \partial\mathring{\mathcal{D}}_f$, respectively), since $\partial\mathring{\mathcal{D}}_f \subseteq \mathbb{R}$. ◊

In the following proposition, we examine the properties of the Legendre transform of a function of class Conv_∞.

Proposition 1.14. If $f \in \mathrm{Conv}_\infty$, then:

(i) $f^*(x) = \sup_{t \in \mathring{\mathcal{D}}_f} [tx - f(t)] \qquad \forall x \in \mathbb{R}$.

(ii) $f' \colon \mathring{\mathcal{D}}_f \to \mathring{\mathcal{D}}_{f^*}$ is a strictly increasing bijection and

$$(f')^{-1} = (f^*)'.$$

Moreover, for any $x \in \mathring{\mathcal{D}}_{f^*}$ the supremum in the definition of f^* (see formula (1.3)) is attained at $t = (f')^{-1}(x)$, that is

$$f^*(x) = (f')^{-1}(x) \cdot x - f\big((f')^{-1}(x)\big) \qquad \forall x \in \mathring{\mathcal{D}}_{f^*}. \tag{1.4}$$

(iii) $f^* \in \mathrm{Conv}_\infty$.

(iv) $f^{**} = f$.

Proof. Let $g_x(t) \doteq tx - f(t)$ for all $x, t \in \mathbb{R}$, and let $\mathring{\mathcal{D}}_f \doteq (t_-, t_+)$.

(i) $g_x(t) = -\infty$ if $t \in \mathbb{R} \setminus \overline{\mathcal{D}}_f$; by Lemma 1.10, f is continuous on $\overline{\mathcal{D}}_f$, hence g_x is so, and

$$f^*(x) = \sup_{t \in \mathbb{R}} g_x(t) = \sup_{t \in (t_-, t_+)} g_x(t).$$

(ii) Let

$$x_- \doteq \lim_{t \searrow t_-} f'(t), \qquad x_+ \doteq \lim_{t \nearrow t_+} f'(t). \tag{1.5}$$

Since $f'' > 0$ on $\mathring{\mathcal{D}}_f$ (by Definition 1.12 (iv)), f' is strictly increasing on $\mathring{\mathcal{D}}_f$, hence a bijection $\mathring{\mathcal{D}}_f \to (x_-, x_+)$.

Let now $x \in (x_-, x_+)$: to compute $f^*(x)$, by (i) it suffices to study $g_x(t)$ for $t \in (t_-, t_+)$. On this interval, g_x is differentiable, with derivative $g_x'(t) = x - f'(t)$. Since f' is strictly increasing on (t_-, t_+), g_x assumes its maximum at the unique point t such that $x = f'(t)$, that is $t = (f')^{-1}(x)$, hence

$$f^*(x) = \sup_{t \in (t_-, t_+)} g_x(t) = (f')^{-1}(x) \cdot x - f\Big((f')^{-1}(x)\Big). \tag{1.6}$$

Therefore, f^* is differentiable on (x_-, x_+), with derivative

$$\begin{aligned}(f^*)'(x) &= \Big((f')^{-1}\Big)'(x) \cdot x + (f')^{-1}(x) - f'\Big((f')^{-1}(x)\Big)\Big((f')^{-1}\Big)'(x) \\ &= \frac{1}{f''\Big(f^{-1}(x)\Big)} x + (f')^{-1}(x) - x \frac{1}{f''\Big(f^{-1}(x)\Big)} = (f')^{-1}(x).\end{aligned}$$

This proves that $(f^*)' = (f')^{-1}$ on (x_-, x_+).

It remains to prove that $(x_-, x_+) = \mathring{\mathcal{D}}_{f^*}$. If $x_+ < +\infty$, then $t_+ = +\infty$ since f is steep. Therefore, for $x \in (x_+, +\infty)$, g_x is differentiable in a neighborhood of $+\infty$, with derivative

$$g_x'(t) = x - f'(t) > x - x_+ > 0,$$

since $f'(t) < x_+$ for all $t \in (t_-, t_+)$. So g_x' is larger than a positive constant in a neighborhood of $+\infty$, hence $\lim_{t \to +\infty} g_x(t) = +\infty$ and

$$f^*(x) = \sup_{t \in \mathbb{R}} g_x(t) = +\infty.$$

Similarly, if $x \in (-\infty, x_-)$, then $f^*(x) = +\infty$. This proves that $\mathcal{D}_{f^*} \subseteq [x_-, x_+]$, i.e. $\mathring{\mathcal{D}}_{f^*} \subseteq (x_-, x_+)$. Moreover, if $x \in (x_-, x_+)$, by (1.6) $f^*(x) \in \mathbb{R}$, hence $(x_-, x_+) \subseteq \mathring{\mathcal{D}}_{f^*}$. We conclude that $(x_-, x_+) = \mathring{\mathcal{D}}_{f^*}$.

(iii) We need to verify the conditions of Definition 1.12. $f < +\infty$ on $\mathring{\mathcal{D}}_f \neq \varnothing$, hence $\sup_{t \in \mathbb{R}}[tx - f(t)] > -\infty$ for all $x \in \mathbb{R}$, which proves that f^* takes values in $(-\infty, +\infty]$. It follows from (ii) that $\mathring{\mathcal{D}}_{f^*} = (x_-, x_+) \neq \varnothing$ and $f^* \in C^\infty(\mathring{\mathcal{D}}_{f^*})$. The lower semi-continuity and the convexity of f^* follow from Lemma 1.7. Since $(f^*)' = (f')^{-1}$ on $\mathring{\mathcal{D}}_{f^*}$ and $f'' > 0$ on $\mathring{\mathcal{D}}_f$, for all $x \in \mathring{\mathcal{D}}_{f^*}$

$$(f^*)''(x) = ((f^*)')'(x) = ((f')^{-1})'(x) = \frac{1}{f''(f^{-1}(x))} > 0.$$

Finally, if

$$x_+ = \sup \mathring{\mathcal{D}}_{f^*} = \lim_{t \nearrow t_+} f'(t) < +\infty,$$

then $t_+ = +\infty$ since f is steep. Therefore

$$\lim_{x \nearrow x_+} (f^*)'(x) = \lim_{x \nearrow x_+} (f')^{-1}(x) = t_+ = +\infty.$$

Similarly, if $x_- > -\infty$, then $\lim_{x \searrow x_-} (f^*)'(x) = -\infty$. This proves that f^* is steep.

(iv) Since $f^* \in \mathrm{Conv}_\infty$ by (iii), we can apply (ii) to f^*, obtaining that $(f^*)'$ is a bijection $\mathring{\mathcal{D}}_{f^*} \to \mathring{\mathcal{D}}_{f^{**}}$; but $(f^*)' = (f')^{-1}$, hence $\mathring{\mathcal{D}}_{f^{**}} = \mathring{\mathcal{D}}_f = (t_-, t_+)$. This proves that $f^{**} = +\infty = f$ on $\mathbb{R} \setminus [t_-, t_+]$.

Let now $t_0 \in (t_-, t_+)$ and $x_0 \doteq f'(t_0) \in (x_-, x_+)$. If we apply (1.4) first to the Legendre transform of f^* at the point t_0 and then to the Legendre transform of f at the point x_0, we obtain

$$\begin{aligned} f^{**}(t_0) &= ((f^*)')^{-1}(t_0) \cdot t_0 - f^*\Big(((f^*)')^{-1}(t_0)\Big) = f'(t_0) \cdot t_0 - f^*(f'(t_0)) \\ &= x_0 t_0 - f^*(x_0) = x_0 t_0 - \Big((f')^{-1}(x_0) \cdot x_0 - f\Big((f')^{-1}(x_0)\Big)\Big) \\ &= x_0 t_0 - (t_0 x_0 - f(t_0)) = f(t_0). \end{aligned}$$

This proves that $f^{**} = f$ on (t_-, t_+).

Assume now that $t_+ < +\infty$. Since $f, f^{**} \in \mathrm{Conv}_\infty$, by Lemma 1.10

$$\lim_{t \nearrow t_+} f(t) = f(t_+), \qquad \lim_{t \nearrow t_+} f^{**}(t) = f^{**}(t_+).$$

We already know that $f^{**} = f$ on (t_-, t_+), hence

$$\lim_{t \nearrow t_+} f(t) = \lim_{t \nearrow t_+} f^{**}(t),$$

which proves that $f^{**}(t_+) = f(t_+)$. An analogous argument shows that, if $t_- > -\infty$, $f^{**}(t_-) = f(t_-)$. We conclude that $f^{**} = f$ on $\mathbb{R}$. □

1.3 The rate function

In this section we study the logarithmic moment generating function and its Legendre transform, the rate function, which will appear in Cramér's Theorem.

From now on, we let X denote a fixed real-valued random variable on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$.

Definition 1.15. We call *moment generating function* of the random variable X the function $M\colon \mathbb{R} \to (0, +\infty]$ given by

$$M(t) \doteq \mathbb{E}(e^{tX}).$$

We call *logarithmic moment generating function* of X the function $\log M$, taking values in $(-\infty,+\infty]$. We call *rate function* the Legendre transform of $\log M$, given by

$$I(x) \doteq \sup_{t\in\mathbb{R}}[tx - \log M(t)].$$

M does not need to be finite everywhere; however, the next theorem states that, where M is finite, it is even analytic. This theorem also justifies the denomination of M as moment generating function.

First we note that, if $M(t) < +\infty$, we can define a new probability measure $\mathbb{P}_t$ on $(\Omega,\mathcal{E})$, absolutely continuous with respect to $\mathbb{P}$, by the following Radon-Nikodym derivative:

$$\frac{d\mathbb{P}_t}{d\mathbb{P}} = \frac{e^{tX}}{M(t)}.$$

We let $\mathbb{E}_t$ and Var_t denote expectation and variance respectively, with respect to $\mathbb{P}_t$. It is clear that $\mathbb{P}_0 = \mathbb{P}$, since

$$\frac{d\mathbb{P}_0}{d\mathbb{P}} = \frac{e^{0X}}{\mathbb{E}(e^{0X})} = 1.$$

Theorem 1.16. M is analytic (hence, C^∞) on $\mathring{\mathcal{D}}_M$, with derivatives

$$M^{(k)}(t) = M(t)\mathbb{E}_t(X^k) = \mathbb{E}(X^k e^{tX}) \qquad \forall t \in \mathring{\mathcal{D}}_M, \forall k \in \mathbb{N}_0. \tag{1.7}$$

In particular, if M is finite in a neighborhood of 0, then $X \in L^p(\Omega,\mathcal{E},\mathbb{P})$ for all $p \in [1,+\infty)$ and its moments are given by

$$\mathbb{E}(X^k) = M^{(k)}(0) \qquad \forall k \in \mathbb{N}_0.$$

Proof. Let $t \in \mathring{\mathcal{D}}_M$, with $(t-\varepsilon, t+\varepsilon) \subseteq \mathring{\mathcal{D}}_M$: this means that $\mathbb{E}(e^{(t+h)X}) < +\infty$ for all $h \in (-\varepsilon,\varepsilon)$, i.e.

$$\mathbb{E}_t(e^{hX}) = \mathbb{E}\left(\frac{e^{(t+h)X}}{M(t)}\right) < +\infty \qquad \forall h \in (-\varepsilon,\varepsilon). \tag{1.8}$$

We claim that

$$M(t+h) = \sum_{k=0}^{\infty} \frac{M(t)\mathbb{E}_t(X^k)}{k!} h^k \qquad \forall h \in (-\varepsilon,\varepsilon). \tag{1.9}$$

Indeed, if $h \in (-\varepsilon,\varepsilon)$

$$\frac{M(t+h)}{M(t)} = \frac{\mathbb{E}(e^{(t+h)X})}{M(t)} = \mathbb{E}\left(e^{hX}\frac{e^{tX}}{M(t)}\right) = \mathbb{E}_t(e^{hX}) = \mathbb{E}_t\left(\sum_{k=0}^{\infty}\frac{(hX)^k}{k!}\right)$$

$$= \mathbb{E}_t\left(\lim_{N\to\infty}\sum_{k=0}^{N}\frac{(hX)^k}{k!}\right) = \lim_{N\to\infty}\mathbb{E}_t\left(\sum_{k=0}^{N}\frac{(hX)^k}{k!}\right) = \lim_{N\to\infty}\sum_{k=0}^{N}\frac{\mathbb{E}_t(X^k)}{k!}h^k = \sum_{k=0}^{\infty}\frac{\mathbb{E}_t(X^k)}{k!}h^k,$$

where the passage to the limit under $\mathbb{E}_t$ holds by the dominated convergence theorem, since the sequence of random variables

$$Y_N \doteq \sum_{k=0}^{N} \frac{(hX)^k}{k!}$$

satisfies

$$|Y_N| \leq \sum_{k=0}^{N} \frac{|hX|^k}{k!} \leq \sum_{k=0}^{\infty} \frac{|hX|^k}{k!} = e^{|hX|} \leq e^{hX} + e^{-hX} \in L^1(\Omega, \mathcal{E}, \mathbb{P}_t)$$

by (1.8). It follows from (1.9) that M can be locally written as a convergent power series, i.e. it is analytic, on $\mathring{\mathcal{D}}_M$, and its derivatives are given by (1.7). □

In the next results, we will need the concept of support of a random variable.

Definition 1.17. We call *support* of a probability measure μ on $\mathbb{R}$ (or $\mathbb{R}^d$) the set

$$\operatorname{supp}(\mu) \doteq \bigcap_{\substack{C \text{ closed}, \\ \mu(C)=1}} C, \tag{1.10}$$

i.e. the smallest closed S such that $\mu(S) = 1$[†]. We call *support* of a real-valued (or multivariate) random variable X the support of the law of X, and we denote it by $\operatorname{supp}(X)$.

We now show some useful properties of the logarithmic moment generating function and the rate function.

Lemma 1.18. The function $\log M$ satisfies:

(i) For all $t \in \mathring{\mathcal{D}}_M$

$$(\log M)'(t) = \mathbb{E}_t(X), \qquad (\log M)''(t) = \operatorname{Var}_t(X).$$

(ii) Let $m_- \doteq \inf(\operatorname{supp} X)$ and $m_+ \doteq \sup(\operatorname{supp} X)$. If $m_+ < +\infty$, $\log M$ has an oblique asymptote with slope m_+ as $t \to +\infty$, i.e.

$$\lim_{t \to +\infty} \frac{\log M(t)}{t} = m_+.$$

[†]Since $S = \operatorname{supp}(\mu)$ is an intersection of closed sets, it is a closed subset (in particular, a Borel subset) of $\mathbb{R}^d$, but since the intersection is *uncountable*, it is not obvious that $\mu(S) = 1$. However, it is well-known that the euclidean topology on $\mathbb{R}^d$ has a countable basis; if we let $\mathcal{A}$ denote the (countable) collection of the complements in $\mathbb{R}^d$ of such a basis, any closed C is the intersection of a subcollection of $\mathcal{A}$. In particular, any closed C such that $\mu(C) = 1$ is the intersection of a subcollection of $\mathcal{A}$ of measure one sets. By (1.10), S is also the intersection of a subcollection of $\mathcal{A}$ of measure one sets. Since $\mathcal{A}$ is countable, this proves that $\mu(S) = 1$.

If $m_- > -\infty$, $\log M$ has an oblique asymptote with slope m_- as $t \to -\infty$, i.e.

$$\lim_{t\to-\infty} \frac{\log M(t)}{t} = m_-.$$

(iii) If X is not a.s. constant, $0 \in \mathring{\mathcal{D}}_M$ and $\log M$ is steep, then $\log M \in \mathrm{Conv}_\infty$.

Proof. (i) If $t \in \mathring{\mathcal{D}}_M$, by Theorem 1.16 $M'(t) = M(t)\mathbb{E}_t(X)$ and $M''(t) = M(t)\mathbb{E}_t(X^2)$, hence

$$\begin{aligned}
(\log M)'(t) &= \frac{M'(t)}{M(t)} = \mathbb{E}_t(X),\\
(\log M)''(t) &= \left(\frac{M'}{M}\right)'(t) = \frac{M''(t)M(t) - M'(t)^2}{M(t)^2}\\
&= \frac{M(t)\mathbb{E}_t(X^2)M(t) - M(t)^2\mathbb{E}_t(X)^2}{M(t)^2}\\
&= \mathbb{E}_t(X^2) - \mathbb{E}_t(X)^2 = \mathrm{Var}_t(X).
\end{aligned}$$

(ii) If $m_+ < +\infty$, then $\forall t > 0$

$$\log M(t) = \log \mathbb{E}(e^{tX}) \leq \log(e^{tm_+}) = tm_+ \in \mathbb{R}.$$

Therefore, $\mathring{\mathcal{D}}_M$ is unbounded above and

$$\limsup_{t\to+\infty} \frac{\log M(t)}{t} \leq m_+. \tag{1.11}$$

On the other hand, $\forall \varepsilon > 0$ and $\forall t > 0$

$$M(t) = \mathbb{E}(e^{tX}) \geq \mathbb{E}(e^{tX}\mathbb{1}_{\{X\geq m_+-\varepsilon\}}) \geq \mathbb{E}(e^{t(m_+-\varepsilon)}\mathbb{1}_{\{X\geq m_+-\varepsilon\}}) = e^{t(m_+-\varepsilon)}\mathbb{P}(X \geq m_+ - \varepsilon).$$

Since $m_+ = \sup(\operatorname{supp} X)$, $\mathbb{P}(X \geq m_+ - \varepsilon) > 0$, hence

$$\log M(t) \geq t(m_+ - \varepsilon) + \log \mathbb{P}(X \geq m_+ - \varepsilon) \qquad \forall t > 0.$$

Dividing by t and passing to the limit inferior as $t \to +\infty$,

$$\liminf_{t\to+\infty} \frac{\log M(t)}{t} \geq m_+ - \varepsilon + \liminf_{t\to+\infty} \frac{\log \mathbb{P}(X \geq m_+ - \varepsilon)}{t} = m_+ - \varepsilon.$$

Since ε is arbitrary, we obtain

$$\liminf_{t\to+\infty} \frac{\log M(t)}{t} \geq m_+. \tag{1.12}$$

From (1.11) and (1.12), it follows that

$$\lim_{t\to+\infty} \frac{\log M(t)}{t} = m_+.$$

The proof of the second statement is similar.

(iii) We need to verify the conditions of Definition 1.12.

By hypothesis, $0 \in \mathring{\mathcal{D}}_M = \mathring{\mathcal{D}}_{\log M} \neq \varnothing$.

If $t_n \to t$, by Fatou's Lemma

$$\begin{aligned}\log M(t) = \log \mathbb{E}(e^{tX}) &= \log\Big(\mathbb{E}\Big(\liminf_{n\to\infty} e^{t_n X}\Big)\Big)\\ &\leq \log\Big(\liminf_{n\to\infty} \mathbb{E}\Big(e^{t_n X}\Big)\Big) = \liminf_{n\to\infty} \log M(t_n),\end{aligned}$$

hence $\log M$ is lower semi-continuous.

$M \in C^\infty(\mathring{\mathcal{D}}_M)$ by Theorem 1.16, hence also $\log M \in C^\infty(\mathring{\mathcal{D}}_M)$.

For all $t_1, t_2 \in \mathbb{R}$, $p \in (0,1)$

$$\begin{aligned}\log M(pt_1 + (1-p)t_2) &= \log \mathbb{E}\Big(e^{(pt_1+(1-p)t_2)X}\Big)\\ &= \log \mathbb{E}\Big((e^{t_1X})^p (e^{t_2X})^{(1-p)}\Big)\\ &\leq \log\left[\mathbb{E}\Big((e^{t_1X})^{p\frac{1}{p}}\Big)^p \mathbb{E}\Big((e^{t_2X})^{(1-p)\frac{1}{1-p}}\Big)^{1-p}\right]\\ &= p\log \mathbb{E}(e^{t_1X}) + (1-p)\log \mathbb{E}(e^{t_2X})\\ &= p\log M(t_1) + (1-p)\log M(t_2),\end{aligned}$$

by applying Hölder's inequality to the random variables $(e^{t_1X})^p$ and $(e^{t_2X})^{(1-p)}$ with the conjugate exponents $\frac{1}{p}, \frac{1}{1-p} \in (1,+\infty)$. This proves that $\log M$ is convex on $\mathbb{R}$. Moreover, if $t \in \mathring{\mathcal{D}}_M$, by (i) $(\log M)''(t) = \mathrm{Var}_t(X) > 0$, since X is not almost surely constant.

Finally, $\log M$ is steep by hypothesis. □

Theorem 1.19 (Properties of the rate function)**.** Let X be a real-valued random variable that is not a.s. constant. Assume that the logarithmic moment generating function of X is finite in a neighborhood of 0 (in particular, $X \in L^p(\Omega, \mathcal{E}, \mathbb{P})$ for all $p \in [1,+\infty)$) and steep. Let $\overline{x} \doteq \mathbb{E}(X)$ and $\sigma^2 \doteq \mathrm{Var}(X)$. Then the rate function I satisfies:

(i) $\mathring{\mathcal{D}}_I = (x_-, x_+) \neq \varnothing$.

(ii) I is convex.

(iii) I is lower semi-continuous.

(iv) $I \in C^\infty(\mathring{\mathcal{D}}_I)$.

(v) I is steep.

(vi) $I^* = \log M$.

(vii) $\lim_{x\to\pm\infty} I(x) = +\infty$.

(viii) I has compact level sets.

(ix) I is decreasing on $(-\infty,\overline{x}]$, and strictly decreasing on $(x_-,\overline{x}]$.
I is increasing on $[\overline{x},+\infty)$, and strictly increasing on $[\overline{x},x_+)$.

(x) $I \geq 0$ and $I(x) = 0$ if and only if $x = \overline{x}$.

(xi) If $x > \overline{x}$, $I(x) = \sup_{t>0}[tx - \log M(t)]$;
if $x < \overline{x}$, $I(x) = \sup_{t<0}[tx - \log M(t)]$.

(xii) $I'' > 0$ on $\mathring{\mathcal{D}}_I$ and $I''(\overline{x}) = \frac{1}{\sigma^2}$.

(xiii) $\mathring{\mathcal{D}}_I = (\inf(\operatorname{supp} X), \sup(\operatorname{supp} X))$.

(xiv) If $x_+ < +\infty \implies I(x_+) = -\log \mathbb{P}(X = x_+)$;
if $x_- > -\infty \implies I(x_-) = -\log \mathbb{P}(X = x_-)$.

Proof. Since $\log M \in \mathrm{Conv}_\infty$ by Lemma 1.18 (iii), the same holds for its Legendre transform $I = (\log M)^*$ by Proposition 1.14 (iii); in particular, (i), (ii), (iii), (iv) and (v) follow immediately. Moreover, (vi) follows from Proposition 1.14 (iv).

(vii) For all $t \in \mathbb{R}$, $I(x) \geq tx - \log M(t)$ $\forall x \in \mathbb{R}$, hence

$$\frac{I(x)}{|x|} \geq t \operatorname{sign} x - \frac{\log M(t)}{|x|} \qquad \forall x \neq 0.$$

Since $0 \in \mathring{\mathcal{D}}_M$, there exist $t_1, t_2 \in \mathring{\mathcal{D}}_M$, $t_1 > 0$, $t_2 < 0$. Therefore, passing to the limit inferior as $x \to \pm\infty$ with $t = t_1, t_2$ in the latter inequality, we obtain

$$\liminf_{x\to+\infty} \frac{I(x)}{|x|} \geq t_1 > 0, \qquad \liminf_{x\to-\infty} \frac{I(x)}{|x|} \geq -t_2 > 0.$$

This proves that $I(x) \to +\infty$ as $x \to \pm\infty$.

(viii) Let $a \in \mathbb{R}$. By (vii), there exists $M \in \mathbb{R}$ such that $I(x) > a$ if $|x| > M$, hence the level set $\{I \leq a\} \subseteq [-M, M]$ is bounded. Since I is lower semi-continuous by (iii), $\{I \leq a\}$ is also closed, hence it is compact.

(ix) I is differentiable on $\mathring{\mathcal{D}}_I$ by (iv), and by Proposition 1.14 (ii) $I' = ((\log M)')^{-1}$. Since $0 \in \mathring{\mathcal{D}}_M$, $I'(x) = 0$ if and only if

$$x = (\log M)'(0) = \mathbb{E}_0(X) = \overline{x}$$

by Lemma 1.18 (i). Since $I \in \mathrm{Conv}_\infty$, I' is strictly increasing on $\mathring{\mathcal{D}}_I$, hence

$$I'(x)\begin{cases} < 0 & \text{if } x \in (x_-, \overline{x}),\\ = 0 & \text{if } x = \overline{x},\\ > 0 & \text{if } x \in (\overline{x}, x_-),\end{cases}$$

which proves that I is strictly decreasing on $[x_-, \overline{x})$ and strictly increasing on $(\overline{x}, x_+]$. By Lemma 1.10, I is continuous on $\overline{\mathcal{D}}_I$, and $I(x) = +\infty$ for $x \notin \overline{\mathcal{D}}_I$, hence I is decreasing on $(-\infty, \overline{x}]$ and increasing for $[\overline{x}, +\infty)$.

(x) $\log M$ is the Legendre transform of I by (vi), hence

$$\inf_{x\in\mathbb{R}} I(x) = -\sup_{x\in\mathbb{R}}[x\cdot 0 - I(x)] = -I^*(0) = -\log M(0) = -\log \mathbb{E}(e^{0X}) = 0.$$

In particular, $I \geq 0$. By (ix), I attains its unique minimum at $\overline{x}$, hence $I(\overline{x}) = 0$ and $I(x) > 0$ for all $x \neq \overline{x}$.

(xi) If $x > \overline{x}$, for all $t \leq 0$

$$tx - \log M(t) \leq t\overline{x} - \log M(t) \leq I(\overline{x}) = 0.$$

Since $x \neq \overline{x}$, $I(x) > 0$, hence

$$I(x) = \sup_{t\in\mathbb{R}}[tx - \log M(t)] = \sup_{t>0}[tx - \log M(t)].$$

The proof for $x < \overline{x}$ is analogous.

(xii) Since $I \in \mathrm{Conv}_\infty$, $I'' > 0$ on $\mathring{\mathcal{D}}_I$. Since $I'(\overline{x}) = 0$ and $(I')^{-1} = \log M$,

$$I''(\overline{x}) = (I')'(\overline{x}) = \frac{1}{((I')^{-1})'(0)} = \frac{1}{(\log M)''(0)}$$

and by Lemma 1.18 (i) $(\log M)''(0) = \mathrm{Var}_0(X) = \sigma^2$.

(xiii) Let $m_- \doteq \inf(\operatorname{supp} X)$ and $m_+ \doteq \sup(\operatorname{supp} X)$. Our aim is to prove that $\mathring{\mathcal{D}}_I = (m_-, m_+)$.

If $m_+ < +\infty$, by Lemma 1.18 (ii)

$$\lim_{t\to+\infty} \frac{\log M(t)}{t} = m_+.$$

Therefore, if $x \in (m_+, +\infty)$,

$$I(x) = \sup_{t\in\mathbb{R}}[tx - \log M(t)] \geq \lim_{t\to+\infty}\left[t\left(x - \frac{\log M(t)}{t}\right)\right] = +\infty,$$

since $x - \frac{\log M(t)}{t} \xrightarrow{t\to+\infty} x - m_+ > 0$. Thus we have proved that $(m_+, +\infty) \subseteq \mathcal{D}_I^c$. Similarly one can prove that $(-\infty, m_-) \subseteq \mathcal{D}_I^c$. Therefore, $\mathcal{D}_I \subseteq [m_-, m_+]$, i.e. $\mathring{\mathcal{D}}_I \subseteq (m_-, m_+)$.

To prove the other inclusion, let $x \in (m_-, m_+)$. If $x = \overline{x}$, $I(\overline{x}) = 0 < +\infty$. If $x \in (\overline{x}, m_+)$, for all $t > 0$ by Markov's inequality

$$0 < \mathbb{P}(X \geq x) = \mathbb{P}(e^{tX} \geq e^{tx}) \leq \frac{\mathbb{E}(e^{tX})}{e^{tx}} = e^{-[tx - \log M(t)]},$$

hence

$$0 < \mathbb{P}(X \geq x) \leq e^{-\sup_{t>0}[tx - \log M(t)]} = e^{-I(x)}$$

by (xi) (it will be useful to remember this argument for the proof of the upper bound in Cramér's Theorem). This proves that $I(x) < +\infty$. The proof for $x \in (m_-, \overline{x})$ is analogous. We conclude that $(m_-, m_+) \subseteq \mathcal{D}_I$, hence $(m_-, m_+) \subseteq \mathring{\mathcal{D}}_I$.

(xiv) If $x_+ = \sup(\operatorname{supp} X) < +\infty$, then

$$\begin{aligned}
I(x_+) &= \sup_{t\in\mathbb{R}}[tx_+ - \log M(t)] = -\inf_{t\in\mathbb{R}}[\log e^{-tx_+} + \log \mathbb{E}(e^{tX})] \\
&= -\inf_{t\in\mathbb{R}} \log\Big(e^{-tx_+}\mathbb{E}(e^{tX}\mathbb{1}_{\{X\leq x_+\}})\Big) = -\log \inf_{t\in\mathbb{R}} \mathbb{E}\Big(e^{t(X-x_+)}\mathbb{1}_{\{X\leq x_+\}}\Big) \\
&= -\log\Big[\inf_{t\in\mathbb{R}} \mathbb{E}\Big(e^{t(X-x_+)}\mathbb{1}_{\{X<x_+\}}\Big) + \mathbb{P}(X = x_+)\Big].
\end{aligned}$$

The collection of random variables $\{e^{t(X-x_+)}\mathbb{1}_{\{X<x_+\}}\}_{t\in\mathbb{R}}$ is decreasing in t, so we can apply monotone convergence theorem and obtain

$$\inf_{t\in\mathbb{R}} \mathbb{E}\Big(e^{t(X-x_+)}\mathbb{1}_{\{X<x_+\}}\Big) = \lim_{t\to+\infty} \mathbb{E}\Big(e^{t(X-x_+)}\mathbb{1}_{\{X<x_+\}}\Big) = \mathbb{E}\Big(\lim_{t\to+\infty} e^{t(X-x_+)}\mathbb{1}_{\{X<x_+\}}\Big) = 0.$$

We conclude that $I(x_+) = -\log \mathbb{P}(X = x_+)$. Similarly one can prove that, if $x_- > -\infty$, $I(x_-) = -\log \mathbb{P}(X = x_-)$. □

1.4 Cramér's Theorem

Now that we know the properties of the rate function I, we can prove the following result, that is the aim of this chapter.

Theorem 1.20 (Cramér's Theorem)**.** Let X be a real-valued random variable such that its logarithmic moment generating function $\log M$ is finite in a neighborhood of 0 and steep, and let $\overline{x} \doteq \mathbb{E}(X)$. Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. copies of X, and let

$$S_n \doteq \sum_{i=1}^{n} X_i.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) = -I(x) \qquad \forall x > \overline{x}, \tag{1.13}$$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \leq x\Big) = -I(x) \qquad \forall x < \overline{x}, \tag{1.14}$$

where I is the rate function, defined by

$$I(x) \doteq (\log M)^*(x) = \sup_{t\in\mathbb{R}}[tx - \log M(t)].$$

Proof. By symmetry, it suffices to prove (1.13). Indeed, if (1.13) holds, then we can apply it to the random variables $\{-X_i\}_{i\geq 1}$, whose mean is $-\overline{x}$: for all $x < \overline{x}$ (i.e. $-x > -\overline{x}$)

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\Big(\frac{S_n}{n} \leq x\Big) &= \lim_{n\to\infty} \mathbb{P}\Big(\frac{(-X_1)+\cdots+(-X_n)}{n} \geq -x\Big) \\ &= -\sup_{t\in\mathbb{R}}[t(-x) - \log \mathbb{E}(e^{t(-X)})] \\ &= -\sup_{t\in\mathbb{R}}[tx - \log \mathbb{E}(e^{tX})] = -I(x),\end{aligned}$$

which proves (1.14).

Let $x > \overline{x}$. If $X = \overline{x}$ almost surely, then

$$I(x) = \sup_{t\in\mathbb{R}}[tx - \log e^{t\overline{x}}] = \sup_{t\in\mathbb{R}}[t(x-\overline{x})] = \begin{cases} 0 & x = \overline{x}, \\ +\infty & x \neq \overline{x}, \end{cases}$$

and $\frac{S_n}{n} = \overline{x}$ almost surely, hence (1.13) is trivially satisfied.

Therefore, we may suppose that X is not almost surely constant. By Theorem 1.19 (xiii), $\mathring{\mathcal{D}}_I \doteq (x_-, x_+)$, with $x_- < x_+$ given by

$$x_- \doteq \inf(\operatorname{supp} X), \qquad x_+ \doteq \sup(\operatorname{supp} X).$$

We distinguish three cases:

(i) If $x > x_+$ (in particular, $x_+ < +\infty$),

$$\mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) \leq \mathbb{P}\Big(\bigcup_{i=1}^{n} \{X_i \geq x\}\Big) \leq \sum_{i=1}^{n} \mathbb{P}(X_i \geq x) = 0.$$

Moreover, $I(x) = +\infty$ since $x \in \mathbb{R} \setminus \overline{\mathcal{D}_I}$, and (1.13) follows.

(ii) If $x = x_+$ (in particular, $x_+ < +\infty$),

$$\mathbb{P}\Big(\frac{S_n}{n} \geq x_+\Big) = \mathbb{P}\Big(\frac{S_n}{n} = x_+\Big) = \mathbb{P}\Big(\bigcap_{i=1}^{n} \{X_i = x_+\}\Big) = \mathbb{P}(X = x_+)^n.$$

Therefore
$$\frac{1}{n}\log\mathbb{P}\Big(\frac{S_n}{n}\geq x_+\Big)=\log\mathbb{P}(X=x_+)=-I(x_+)$$
by Theorem 1.19 (xiv).

(iii) Let now $\overline{x}<x<x_+$. By Markov's inequality, for all $t>0$
$$\begin{aligned}\mathbb{P}\Big(\frac{S_n}{n}\geq x\Big)&=\mathbb{P}(tS_n\geq tnx)=\mathbb{P}(e^{tS_n}\geq e^{tnx})\\ \leq\frac{\mathbb{E}(e^{tS_n})}{e^{tnx}}&=\left(\frac{\mathbb{E}(e^{tX})}{e^{tx}}\right)^n=\left(\frac{e^{\log M(t)}}{e^{tx}}\right)^n=e^{-n[tx-\log M(t)]},\end{aligned}$$
hence
$$\mathbb{P}\Big(\frac{S_n}{n}\geq x\Big)\leq\inf_{t>0}e^{-n[tx-\log M(t)]}=e^{-n\sup_{t>0}[tx-\log M(t)]}=e^{-nI(x)}$$
by Theorem 1.19 (xi). This proves that
$$\frac{1}{n}\log\mathbb{P}\Big(\frac{S_n}{n}\geq x\Big)\leq-I(x)\qquad\forall n\geq1. \tag{1.15}$$
To prove (1.13), we should now estimate the limit inferior from below. Since $x\in\mathring{\mathcal{D}}_I$, by Proposition 1.14 (ii)
$$I(x)=tx-\log M(t),$$
where $t=((\log M)')^{-1}(x)=I'(x)\in\mathring{\mathcal{D}}_M$, i.e. $x=(\log M)'(t)=\mathbb{E}_t(X)$ by Lemma 1.18 (i). We note that, since I' is strictly increasing on $\mathring{\mathcal{D}}_I$ and $x>\overline{x}$, we have that $t=I'(x)>I'(\overline{x})=0$. Recalling that $\frac{d\mathbb{P}_t}{d\mathbb{P}}=\frac{e^{tX}}{M(t)}$, $\forall\varepsilon>0$ we can write
$$\begin{aligned}\mathbb{P}\Big(\frac{S_n}{n}\geq x\Big)&\geq\mathbb{P}\Big(x\leq\frac{S_n}{n}\leq x+\varepsilon\Big)=\mathbb{E}\Big(\mathbb{1}_{\{0\leq\frac{S_n}{n}-x\leq\varepsilon\}}\frac{e^{tS_n}}{e^{tS_n}}\Big)\\ &\geq\frac{M(t)^n}{e^{tn(x+\varepsilon)}}\mathbb{E}\Big(\mathbb{1}_{\{0\leq\frac{S_n}{n}-x\leq\varepsilon\}}\frac{e^{tS_n}}{M(t)^n}\Big)\\ &=e^{-t\varepsilon n}e^{-n[tx-\log M(t)]}\mathbb{E}\left(\mathbb{1}_{\{0\leq\frac{S_n}{n}-x\leq\varepsilon\}}\prod_{i=1}^n\frac{e^{tX_i}}{M(t)}\right)\\ &=e^{-t\varepsilon n}e^{-nI(x)}\mathbb{E}_t\Big(\mathbb{1}_{\{0\leq\frac{S_n}{n}-x\leq\varepsilon\}}\Big)=e^{-t\varepsilon n}e^{-nI(x)}\mathbb{P}_t\Big(0\leq\frac{S_n}{n}-x\leq\varepsilon\Big).\end{aligned}$$
If we put $\sigma_t^2\doteq\mathrm{Var}_t(X)>0$ and we consider $\varepsilon\doteq\frac{\sigma_t}{\sqrt{n}}$,
$$\mathbb{P}_t\Big(0\leq\frac{S_n}{n}-x\leq\varepsilon\Big)=\mathbb{P}_t\Big(0\leq\frac{\sqrt{n}}{\sigma_t}\Big(\frac{S_n}{n}-x\Big)\leq1\Big)\xrightarrow{n\to\infty}\int_0^1\frac{1}{\sqrt{2\pi}}e^{-s^2/2}\,ds\doteq\theta>0$$

by the central limit theorem, since $\mathbb{E}_t(X) = x$. This proves that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) \geq \liminf_{n\to\infty} \frac{1}{n} \log\Big[e^{-t\frac{\sigma_t}{\sqrt{n}}n} e^{-nI(x)} \mathbb{P}_t\Big(0 \leq \frac{S_n}{n} - x \leq \varepsilon\Big)\Big]$$
$$= \liminf_{n\to\infty}\Big(-t\frac{\sigma_t}{\sqrt{n}} - I(x) + \frac{1}{n}\log(\theta + o(1))\Big) = -I(x). \quad \square$$

Remark 1.21. In Cramér's Theorem, the hypothesis that $\log M$ is steep and finite in a neighborhood of 0 is not actually necessary (see [DeZe, Th.2.2.3] for a general proof). However, we note that:

- If $\log M$ is not steep, some properties of the rate function are missing: I is neither steep nor *strictly* convex on its domain.
- If $\log M$ is not finite in a neighborhood of zero, further very important properties of the rate function are missing. If $\mathcal{D}_M = \{0\}$, then $I \equiv 0$; in this case, Cramér's Theorem just tells us that the large deviations probabilities decay slower than exponentially, but it gives us no information on how slow the decay is. If M is finite in a right-neighborhood of 0 but not in a left-neighborhood, $I(x) = 0$ for all $x \leq \overline{x}$, and I satisfies the usual properties only for $x \geq \overline{x}$ (in particular, $I(x) > 0$ for $x > \overline{x}$); in this case, Cramér's Theorem gives us interesting information only about the decay of upwards large deviations probabilities. Conversely, if M is finite in a left-neighborhood of 0 but not in a right-neighborhood, $I(x) = 0$ for all $x \geq \overline{x}$, and I satisfies the usual properties only for $x \leq \overline{x}$ (in particular, $I(x) > 0$ for $x < \overline{x}$); in this case, Cramér's Theorem gives us interesting information only about the decay of downwards large deviations probabilities. Therefore, in all these cases, I does not have infinite limits at infinity (in particular, I no longer has compact level sets), and $I(x)$ might be zero even for $x \neq \overline{x}$. ◊

Remark 1.22. Cramér's Theorem tells us that, under the condition that the law of X has sufficiently light tails (M is finite in a neighborhood of 0), the probability that the empirical average $\frac{S_n}{n}$ deviates by any given constant from its mean decays exponentially in n:

$$\mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) = e^{-nI(x)+o(n)} \qquad \forall x > \overline{x}, \tag{1.16}$$

$$\mathbb{P}\Big(\frac{S_n}{n} \leq x\Big) = e^{-nI(x)+o(n)} \qquad \forall x < \overline{x}. \tag{1.17}$$

We note that, in fact, our proof shows that for $\overline{x} < x < \sup(\operatorname{supp} X)$

$$e^{-nI(x)+O(\sqrt{n})} \leq \mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) \leq e^{-nI(x)},$$

where $O(\sqrt{n}) = -c\sqrt{n} + \log\theta + o(1)$ and c is a positive constant. ◊

Remark 1.23. This remark will be often useful. Assume that α_n, β_n and s_n are positive sequences and $s_n \to \infty$ (for example $s_n = n$). Since

$$\frac{1}{s_n}\log(\alpha_n \vee \beta_n) \le \frac{1}{s_n}\log(\alpha_n + \beta_n) \le \frac{1}{s_n}\log(2(\alpha_n \vee \beta_n)),$$

we obtain

$$0 \le \frac{1}{s_n}\Big(\log(\alpha_n + \beta_n) - \log(\alpha_n \vee \beta_n)\Big) \le \frac{\log 2}{s_n} \to 0.$$

This proves that

$$\lim_{n\to\infty} \frac{1}{s_n}\Big(\log(\alpha_n + \beta_n) - \log(\alpha_n \vee \beta_n)\Big) = 0. \tag{1.18}$$

We can also rewrite the latter equality as

$$\frac{1}{s_n}\log(\alpha_n + \beta_n) = \left(\frac{1}{s_n}\log\alpha_n\right) \vee \left(\frac{1}{s_n}\log\beta_n\right) + o(1). \tag{1.19}$$

Applying the exponential function to both sides of the latter equality, we also obtain

$$(\alpha_n + \beta_n)^{1/s_n} = \left(\alpha_n^{1/s_n} \vee \beta_n^{1/s_n}\right)(1 + o(1)). \tag{1.20}$$

◊

Remark 1.24. Cramér's Theorem also allows us to evaluate the decay of two-sided deviations of $\frac{S_n}{n}$ from its mean. If $x < \overline{x}$ and $x' > \overline{x}$,

$$\begin{aligned}
\frac{1}{n}\log\mathbb{P}\left(\left\{\frac{S_n}{n} \le x\right\} \cup \left\{\frac{S_n}{n} \ge x'\right\}\right) &= \frac{1}{n}\log\left(\mathbb{P}\left(\frac{S_n}{n} \le x\right) + \mathbb{P}\left(\frac{S_n}{n} \ge x'\right)\right) \\
&= \frac{1}{n}\log\left(e^{-nI(x)+o(n)} + e^{-nI(x')+o(n)}\right) \\
&= \left(\frac{-nI(x) + o(n)}{n}\right) \vee \left(\frac{-nI(x') + o(n)}{n}\right) + o(1) \\
&\xrightarrow{n\to\infty} (-I(x)) \vee (-I(x')) = -(I(x) \wedge I(x')),
\end{aligned}$$

having applied (1.19) with $\alpha_n = e^{-nI(x)+o(n)}$, $\beta_n = e^{-nI(x')+o(n)}$ and $s_n = n$. In particular, for all $\varepsilon > 0$

$$\lim_{n\to\infty} \frac{1}{n}\log\mathbb{P}\left(\left|\frac{S_n}{n} - \overline{x}\right| \ge \varepsilon\right) = -(I(\overline{x} - \varepsilon) \wedge I(\overline{x} + \varepsilon)) < 0. \tag{1.21}$$

◊

Remark 1.25. Cramér's Theorem implies, in particular, the strong law of large numbers. By (1.21), for all $\varepsilon > 0$

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \overline{x}\right| \ge \varepsilon\right) = \sum_{n=1}^{\infty} e^{-nc+o(n)},$$

where $c = I(\overline{x}-\varepsilon) \wedge I(\overline{x}+\varepsilon) > 0$, and such a series converges by the root test, since

$$\sqrt[n]{e^{-nc+o(n)}} \xrightarrow{n\to\infty} e^{-c} < 1.$$

By Borel-Cantelli Lemma, it follows that $\frac{S_n}{n} \to \overline{x}$ almost surely. ◊

We now compute the rate function for some probability distributions, recalling that

$$I(x) = tx - \log M(t) \qquad \forall x \in \mathring{\mathcal{D}}_I = (x_-, x_+),$$

where t is such that $(\log M)'(t) = x$, $x_- \doteq \inf(\operatorname{supp} X)$, $x_+ \doteq \sup(\operatorname{supp} X)$ and I is continuous on $\overline{\mathcal{D}}_I$.

Example 1.26. Assume that X has a Bernoulli distribution with parameter $p \in (0,1)$, i.e.

$$\mathbb{P}(X=0) = 1-p, \qquad \mathbb{P}(X=1) = p.$$

Then $\mathring{\mathcal{D}}_I = (0,1)$ and the moment generating function is

$$M(t) = \mathbb{E}(e^{tX}) = (1-p)e^{t\cdot 0} + pe^{t\cdot 1} = 1-p+pe^t.$$

Therefore $\mathring{\mathcal{D}}_M = \mathbb{R}$ and

$$\log M(t) = \log(1-p+pe^t), \qquad (\log M)'(t) = \frac{pe^t}{1-p+pe^t}.$$

We have $x = (\log M)'(t)$ if and only if $t = \log\left(\frac{(1-p)x}{p(1-x)}\right)$, hence for all $x \in (0,1)$

$$\begin{aligned} I(x) &= \log\left(\frac{(1-p)x}{p(1-x)}\right)x - \log\left(1-p+p\frac{(1-p)x}{p(1-x)}\right) \\ &= \log\left(\frac{x}{p}\right)x - \log\left(\frac{1-x}{1-p}\right)x + \log\left(\frac{1-x}{1-p}\right) \\ &= x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right). \end{aligned}$$

We conclude that

$$I(x) = \begin{cases} x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right) & x \in (0,1), \\ -\log(1-p) & x = 0, \\ -\log p & x = 1, \\ +\infty & x \in (-\infty,0)\cup(1,+\infty). \end{cases}$$

In particular, this agrees with what we obtained in Example 1.1 for $p = \frac{1}{2}$. ◊

Example 1.27. Assume that X has a Poisson distribution with parameter $\lambda > 0$, i.e.

$$\mathbb{P}(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad \forall x \in \mathbb{N}_0.$$

Then $\mathring{\mathcal{D}}_I = (0, +\infty)$ and the moment generating function is

$$M(t) = \mathbb{E}(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(e^t\lambda)^x}{x!} = e^{-\lambda}e^{e^t\lambda} = e^{\lambda(e^t-1)}.$$

Therefore $\mathring{\mathcal{D}}_M = \mathbb{R}$ and

$$\log M(t) = \lambda(e^t - 1), \qquad (\log M)'(t) = \lambda e^t.$$

We have $x = (\log M)'(t)$ if and only if $t = \log(\frac{x}{\lambda})$, hence for all $x \in (0, +\infty)$

$$I(x) = \log\left(\frac{x}{\lambda}\right)x - \lambda\left(\frac{x}{\lambda} - 1\right) = \log\left(\frac{x}{\lambda}\right)x - x + \lambda.$$

We conclude that

$$I(x) = \begin{cases} +\infty & x < 0 \\ \lambda & x = 0 \\ \log\left(\frac{x}{\lambda}\right)x - x + \lambda & x > 0. \end{cases} \qquad \Diamond$$

Example 1.28. Assume that X has an exponential distribution with parameter $\lambda > 0$, i.e. it is absolutely continuous with density function

$$f(x) = \lambda e^{-\lambda x}\mathbb{1}_{[0,+\infty)}(x).$$

Then $\mathring{\mathcal{D}}_I = (0, +\infty)$ and the moment generating function is

$$M(t) = \mathbb{E}(e^{tX}) = \int_0^{\infty} e^{tx}\lambda e^{-\lambda x}\,dx = \lambda\int_0^{\infty} e^{-(\lambda - t)x}\,dx = \begin{cases} \frac{\lambda}{\lambda - t} & t < \lambda \\ +\infty & t \geq \lambda. \end{cases} \tag{1.22}$$

Therefore, $\mathring{\mathcal{D}}_M = (-\infty, \lambda) \ni 0$ and for all $t \in (-\infty, \lambda)$

$$\log M(t) = \log\left(\frac{\lambda}{\lambda - t}\right) = \log\lambda - \log(\lambda - t), \qquad (\log M)'(t) = \frac{1}{\lambda - t}.$$

We note that $\log M$ is steep, since

$$\lim_{t \nearrow \lambda}(\log M)'(t) = +\infty.$$

We have $x = (\log M)'(t)$ if and only if $t = \lambda - x^{-1}$, hence for all $x \in (0, +\infty)$

$$I(x) = (\lambda - x^{-1})x - \log\left(\frac{\lambda}{\lambda - (\lambda - x^{-1})}\right) = \lambda x - \log(\lambda x) - 1.$$

We conclude that

$$I(x) = \begin{cases} +\infty & x \leq 0 \\ \lambda x - \log(\lambda x) - 1 & x > 0. \end{cases} \qquad \Diamond$$

EXAMPLE 1.29. Assume that X has a normal distribution with mean $\overline{x}$ and variance σ^2, i.e. it is absolutely continuous with density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\overline{x})^2}{2\sigma^2}\Big).$$

Then $\mathring{\mathcal{D}}_I = \mathbb{R}$ and the moment generating function is

$$\begin{aligned}
M(t) = \mathbb{E}(e^{tX}) &= \int_{\mathbb{R}} \exp(tx) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\overline{x})^2}{2\sigma^2}\Big) dx \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x^2 - 2x\overline{x} + \overline{x}^2 - 2\sigma^2 tx)\Big) dx \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x^2 - 2(\overline{x}+\sigma^2 t)x + (\overline{x}+\sigma^2 t)^2) - \frac{\overline{x}^2}{2\sigma^2} + \frac{(\overline{x}+\sigma^2 t)^2}{2\sigma^2}\Big) dx \\
&= \exp\Big(\frac{1}{2\sigma^2}(-\overline{x}^2 + \overline{x}^2 + 2\overline{x}\sigma^2 t + \sigma^4 t^2)\Big) \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-(\overline{x}+\sigma^2 t))^2}{2\sigma^2}\Big) dx \\
&= \exp\Big(\overline{x}t + \frac{\sigma^2}{2} t^2\Big),
\end{aligned}$$

where the latter equality holds since the integrand is the density of a normal random variable with mean $\overline{x} + \sigma^2 t$ and variance σ^2. Therefore, $\mathring{\mathcal{D}}_M = \mathbb{R}$ and

$$\log M(t) = \overline{x}t + \frac{\sigma^2}{2} t^2 \qquad (\log M)'(t) = \overline{x} + \sigma^2 t.$$

We have $x = (\log M)'(t)$ if and only if $t = \frac{x-\overline{x}}{\sigma^2}$, hence for all $x \in \mathbb{R}$

$$\begin{aligned}
I(x) &= \frac{x-\overline{x}}{\sigma^2} x - \Big(\overline{x}\frac{x-\overline{x}}{\sigma^2} + \frac{\sigma^2}{2}\frac{(x-\overline{x})^2}{\sigma^4}\Big) \\
&= \frac{1}{\sigma^2}\Big((x-\overline{x})x - \overline{x}(x-\overline{x}) - \frac{(x-\overline{x})^2}{2}\Big) \\
&= \frac{1}{\sigma^2}\Big((x-\overline{x})^2 - \frac{(x-\overline{x})^2}{2}\Big) = \frac{(x-\overline{x})^2}{2\sigma^2}.
\end{aligned}$$

In particular, if $X \sim \mathcal{N}(0,1)$, the rate function is $I(x) = \frac{x^2}{2}$, which agrees with what we obtained in Example 1.2. ◊

Chapter 2

Abstract Theory

The abstract theory of large deviations was first formulated in 1966 by the Indian American mathematician Srinivasa Varadhan (see [Var]), who was awarded the prestigious Abel prize for 2007 "for his fundamental contributions to probability theory and in particular for creating a unified theory of large deviations". The American mathematician Monroe David Donsker also made a significant contribution, publishing several papers together with Varadhan. The abstract theory, which we present in this chapter (mainly based on [Swa, ch.1] and [Hol, ch.III]), allows to unify and extend Cramér's theorem and other important results under the same theoretical framework, as we will see in the next chapter.

2.1 Semi-continuous functions

Semi-continuous functions play a fundamental role in large deviations theory, so in this section we intend to define them and study their main properties.

Proposition 2.1 (Topologies of semi-continuity). The following collections of sets are non-Hausdorff topologies on $\overline{\mathbb{R}}$:

(i) $\mathcal{O}_{up} = \Big\{[-\infty, a): \quad -\infty < a \leq +\infty\Big\} \cup \Big\{\varnothing, \overline{\mathbb{R}}\Big\}$.

(ii) $\mathcal{O}_{low} = \Big\{(a, +\infty]: \quad -\infty \leq a < +\infty\Big\} \cup \Big\{\varnothing, \overline{\mathbb{R}}\Big\}$.

Proof. (i) Let us prove the three axioms of topology:

- $\varnothing, \overline{\mathbb{R}} \in \mathcal{O}_{up}$ by definition.
- If $-\infty < a_1, a_2 \leq +\infty$:

$$[-\infty, a_1) \cap [-\infty, a_2) = [-\infty, a_1 \wedge a_2) \in \mathcal{O}_{up}.$$

- If $-\infty < a_i \leq +\infty$ for all i in an index set S:

$$\bigcup_{i \in S} [-\infty, a_i) = \left[-\infty, \sup_{i \in S} a_i\right) \in \mathcal{O}_{up}.$$

The topology $\mathcal{O}_{up}$ is not Hausdorff, since open neighborhoods of two different points always intersect: if $b, c \in \overline{\mathbb{R}}$ with $b < c$, the open neighborhoods of c are $[-\infty, a)$ with $c < a \leq +\infty$ and $\overline{\mathbb{R}}$, all of which intersect $\{b\}$.

(ii) The proof is analogous. □

We call $\mathcal{O}$ the Euclidean topology on $\overline{\mathbb{R}}$, i.e. the topology whose base is composed of the open intervals (a, b), $-\infty < a < b < +\infty$, and the neighborhoods of $-\infty$ and $+\infty$ (namely, the sets $[-\infty, a)$ and $(a, +\infty]$, $a \in \mathbb{R}$). Recall that this topology is not only Hausdorff, but it is also metrizable: to be more precise, it is induced by a metric whose balls form the base we have just defined, and whose restriction to $\mathbb{R}$ generates the Euclidean topology of $\mathbb{R}$ (see [Soa, §3.9 and §3.10.3]).

Definition 2.2. Let $(\mathbf{X}, \tau)$ be a topological space and let $f : \mathbf{X} \to \overline{\mathbb{R}}$ be a function.

- f is said to be *upper semi-continuous* if it is continuous as function

$$(\mathbf{X}, \tau) \to (\overline{\mathbb{R}}, \mathcal{O}_{up}).$$

- f is said to be *lower semi-continuous* if it is continuous as function

$$(\mathbf{X}, \tau) \to (\overline{\mathbb{R}}, \mathcal{O}_{low}).$$

- f is said to be *continuous* if it is continuous as function

$$(\mathbf{X}, \tau) \to (\overline{\mathbb{R}}, \mathcal{O}).$$

Remark 2.3.

- Assume that $f^{-1}([-\infty, a))$ is open for all $a \in \mathbb{R}$; then

$$f^{-1}\big([-\infty, +\infty)\big) = f^{-1}\left(\bigcup_{n=1}^{\infty} [-\infty, n)\right) = \bigcup_{n=1}^{\infty} f^{-1}\big([-\infty, n)\big)$$

is open since it is union of opens. Therefore, to prove that f is upper semi-continuous, it suffices to prove that

$$f^{-1}\big([-\infty, a)\big) = \big\{x \in \mathbf{X} : \quad f(x) < a\big\}$$

is open for all $a \in \mathbb{R}$, or similarly that

$$f^{-1}\big([a, +\infty]\big) = \big\{x \in \mathbf{X} : \quad f(x) \geq a\big\}$$

is closed for all $a \in \mathbb{R}$.

- In the same way, to prove that f is lower semi-continuous, it suffices to prove that
$$f^{-1}\Big((a,+\infty]\Big) = \Big\{x \in \mathbf{X} : \quad f(x) > a\Big\}$$
is open for all $a \in \mathbb{R}$, or similarly that
$$f^{-1}\Big([-\infty,a]\Big) = \Big\{x \in \mathbf{X} : \quad f(x) \le a\Big\}$$
is closed for all $a \in \mathbb{R}$. ◊

Remark 2.4. A function f is lower semi-continuous if and only if $-f$ is upper semi-continuous. ◊

We define the following sets:

$$U(\mathbf{X}) \doteq \Big\{f : \mathbf{X} \to \overline{\mathbb{R}} : \quad f \text{ upper semi-continuous}\Big\},$$
$$L(\mathbf{X}) \doteq \Big\{f : \mathbf{X} \to \overline{\mathbb{R}} : \quad f \text{ lower semi-continuous}\Big\},$$
$$C(\mathbf{X}) \doteq \Big\{f : \mathbf{X} \to \overline{\mathbb{R}} : \quad f \text{ continuous}\Big\}.$$

We also define their subsets $U_b(\mathbf{X})$, $L_b(\mathbf{X})$ and $C_b(\mathbf{X})$ (respectively, $U_{b,+}(\mathbf{X})$, $L_{b,+}(\mathbf{X})$ and $C_{b,+}(\mathbf{X})$) by requiring their elements to be bounded functions (respectively, bounded nonnegative functions).

We can now list some properties of semi-continuous functions.

Proposition 2.5. Let $(\mathbf{X},\tau)$ be a topological space. Then:

(i) $C(\mathbf{X}) = U(\mathbf{X}) \cap L(\mathbf{X})$. Namely, a function $\mathbf{X} \to \overline{\mathbb{R}}$ is continuous if and only if it is both upper and lower semi-continuous.

(ii) An upper semi-continuous function attains its maximum on every nonempty compact subset of $\mathbf{X}$ (resp., a lower semi-continuous function attains its minimum on every nonempty compact subset of $\mathbf{X}$).

(iii) If $f \in U(\mathbf{X})$ and $\lambda \ge 0$, then $\lambda f \in U(\mathbf{X})$ (resp., if $f \in L(\mathbf{X})$ and $\lambda \ge 0$, then $\lambda f \in L(\mathbf{X})$).

(iv) Let $f,g\colon \mathbf{X} \to \overline{\mathbb{R}}$ be such that $f+g$ is well-defined, i.e. $\{f=+\infty\} \cap \{g=-\infty\} = \varnothing$ and $\{f=-\infty\} \cap \{g=+\infty\} = \varnothing$. If $f,g \in U(\mathbf{X})$, then $f+g \in U(\mathbf{X})$ (resp., if $f,g \in L(\mathbf{X})$, then $f+g \in L(\mathbf{X})$).

(v) If $f,g \in U(\mathbf{X})$ and $f,g \ge 0$, then $fg \in U(\mathbf{X})$ (resp., if $f,g \in L(\mathbf{X})$ and $f,g \ge 0$, then $fg \in L(\mathbf{X})$).

(vi) If $f,g \in U(\mathbf{X})$, then $f \vee g \in U(\mathbf{X})$ and $f \wedge g \in U(\mathbf{X})$ (resp., if $f,g \in L(\mathbf{X})$, then $f \vee g \in L(\mathbf{X})$ and $f \wedge g \in L(\mathbf{X})$).

(vii) If $f_k \searrow f$ and $f_k \in U(\mathbf{X})\ \forall k$, then $f \in U(\mathbf{X})$ (resp., if $f_k \nearrow f$ and $f_k \in L(\mathbf{X})\ \forall k$, then $f \in L(\mathbf{X})$).

Proof. (i) Since $\mathcal{O} \supseteq \mathcal{O}_{up} \cup \mathcal{O}_{low}$, it is obvious that

$$C(\mathbf{X}) \subseteq U(\mathbf{X}) \cap L(\mathbf{X}).$$

Conversely, if $f \in U(\mathbf{X}) \cap L(\mathbf{X})$ then $\forall a, b \in \mathbb{R}$ we have

$$f^{-1}\Big([-\infty, b)\Big) \in \tau,$$
$$f^{-1}\Big((a, +\infty]\Big) \in \tau,$$

hence

$$f^{-1}\Big((a,b)\Big) = f^{-1}\Big([-\infty, b)\Big) \cap f^{-1}\Big((a, +\infty]\Big) \in \tau.$$

Therefore, $f^{-1}(O)$ is open for every O in the base of the topology $\mathcal{O}$, and so for every $O \in \mathcal{O}$: this proves that f is continuous.

We note that, to prove all the other claims except (v), thanks to Remark 2.4 it suffices to consider upper semi-continuous functions.

(ii) Let $K \subseteq \mathbf{X}$ be nonempty and compact and $f \in U(\mathbf{X})$. Let a_n be an increasing sequence converging to $\sup_{x \in K} f(x) \in \overline{\mathbb{R}}$ and let

$$K_n \doteq \{x \in K : \ f(x) \geq a_n\} = K \cap f^{-1}\Big([a_n, +\infty]\Big).$$

Then K_n is closed in K, since it is the intersection of K with a closed subset of $\mathbf{X}$; in particular, K_n is compact as closed subset of a compact. Moreover, by the definition of sup, for all n there exists $x_n \in K$ such that $f(x_n) \geq a_n$, so every K_n is nonempty. The intersection of $\{K_n\}_{n \in \mathbb{N}}$, that is a decreasing sequence of nonempty, compact and closed[†] subsets of K, is nonempty (see [Man, Prop.4.43]), hence there exists $\overline{x} \in \bigcap_{n=1}^{\infty} K_n$: for such $\overline{x} \in K$, $f(\overline{x}) \geq a_n$ for all n, hence $f(\overline{x}) = \sup_{x \in K} f(x)$. This proves that f attains its maximum on K.

(iii) If $f \in U(\mathbf{X})$ and $\lambda > 0$, then for all $a \in \mathbb{R}$

$$\{\lambda f < a\} = \{f < \lambda^{-1} a\}$$

is open since $f \in U(\mathbf{X})$, hence also $\lambda f \in U(\mathbf{X})$. Moreover, $0 \cdot f = 0$ is clearly an upper semi-continuous function.

(iv) If $f, g : \mathbf{X} \to \overline{\mathbb{R}}$ and $f + g$ is well-defined, then for all $a \in \mathbb{R}$ we claim that

$$\{f + g < a\} = \bigcup_{\substack{(c,d) \in \mathbb{R}^2 \\ c+d=a}} \{f < c\} \cap \{g < d\}. \tag{2.1}$$

Indeed, assume that $f(x) + g(x) < a$. If $f(x) = -\infty$ and $g(x) \in \mathbb{R}$, then x is in the right-hand set, considering $c \doteq a - g(x) - 1$ and $d \doteq g(x) + 1$ (similarly if

[†]If the space is not Hausdorff, it is essential to require every K_n to be closed, because a compact subset may be non-closed.

$f(x) \in \mathbb{R}$ and $g(x) = -\infty$). If $f(x) = g(x) = -\infty$, then x is in the right-hand set, considering $c \doteq a$ and $d \doteq 0$. If $f(x), g(x) \in \mathbb{R}$ and we call $s \doteq f(x) + g(x)$, we have

$$f(x) < f(x) + \frac{a-s}{2} \doteq c, \qquad g(x) < g(x) + \frac{a-s}{2} \doteq d,$$

and

$$c + d = f(x) + \frac{a-s}{2} + g(x) + \frac{a-s}{2} = s + 2\frac{a-s}{2} = a,$$

hence x is in the right hand set again. Conversely, if there exists $(c,d) \in \mathbb{R}^2$ such that $c + d = a$, $f(x) < c$ and $g(x) < d$, hence $f(x) + g(x) < c + d = a$ and x is in the left-hand set.

If $f, g \in U(\mathbf{X})$, then $\{f < c\}$ and $\{g < d\}$ are open for all $c, d \in \mathbb{R}$, hence for all $a \in \mathbb{R}$ $\{f + g < a\}$ is open by (2.1), and $f + g \in U(\mathbf{X})$.

(v) Let $f, g \in U(\mathbf{X})$, $f \geq 0$ and $g \geq 0$. To prove that $fg \in U(\mathbf{X})$, we will show that, for all $a \in \mathbb{R}$, $\{fg < a\}$ is open. If $a \leq 0$ the set is empty and hence open. If $a > 0$, it turns out that

$$\{fg < a\} = \bigcup_{\substack{c,d>0 \\ cd=a}} \{f < c\} \cap \{g < d\}. \tag{2.2}$$

Indeed, assume that $f(x)g(x) < a$. If $f(x) = 0$, then x is in the right-hand set, considering $c \doteq (g(x)+1)^{-1}$ and $d \doteq g(x) + 1$ (the same if $g(x) = 0$). If $f(x), g(x) > 0$, then

$$f(x) < \sqrt{\frac{a}{f(x)g(x)}} f(x) \doteq c, \qquad g(x) < \sqrt{\frac{a}{f(x)g(x)}} g(x) \doteq d,$$

and

$$cd = \sqrt{\frac{a}{f(x)g(x)}} f(x) \sqrt{\frac{a}{f(x)g(x)}} g(x) = a,$$

hence x is in the right-hand set. Conversely, if there exist $c, d > 0$ such that $cd = a$, $f(x) < c$ and $g(x) < d$, hence $f(x)g(x) < cd = a$ and x is in the left-hand set. Since $f, g \in U(\mathbf{X})$, $\{f < c\}$ and $\{g < d\}$ are open for all c, d, hence $\{fg < a\}$ is open by (2.2).

Let now $f, g \in L(\mathbf{X})$, $f \geq 0$ and $g \geq 0$. To prove that $fg \in L(\mathbf{X})$, we will show that, for all $a \in \mathbb{R}$, $\{fg > a\}$ is open. If $a < 0$, the set is $\mathbf{X}$ and hence it is open. If $a = 0$, then

$$\{fg > 0\} = \{f > 0\} \cap \{g > 0\}$$

is open, since $\{f > 0\}$ and $\{g > 0\}$ are open. If $a > 0$, the claim follows from the equality

$$\{fg > a\} = \bigcup_{\substack{c,d>0 \\ cd=a}} \{f > c\} \cap \{g > d\},$$

which can be proved in a similar way to (2.2).

(vi) If $f, g \in U(\mathbf{X})$, then for all $a \in \mathbb{R}$

$$\{f \vee g < a\} = \{f < a\} \cap \{g < a\},$$
$$\{f \wedge g < a\} = \{f < a\} \cup \{g < a\}$$

are open since $\{f < a\}$ and $\{g < a\}$ are open, hence also $f \vee g, f \wedge g \in U(\mathbf{X})$.

(vii) If $f_k \searrow f$ and $f_k \in U(\mathbf{X})$ $\forall k$, then for all $a \in \mathbb{R}$

$$\{f < a\} = \bigcup_{k \in \mathbb{N}} \{f_k < a\}$$

is open because every f_k is upper semi-continuous. □

Definition 2.6. A topological space is said to be *first-countable* if each point has a countable basis of neighborhoods.

In particular, a metric space $\mathbf{X}$ is first countable, because every point $x \in \mathbf{X}$ has the countable basis of neighborhoods $\{B_{n^{-1}}(x)\}_{n \geq 1}$. It turns out that, if X is first-countable and $C \subseteq \mathbf{X}$, $x \in \overline{C}$ if and only if there exists a sequence $x_n \in C$ such that $x_n \to x$ (see [Man, Prop.6.18]); hence, C is closed if and only if, for all sequence $x_n \in C$ such that $x_n \to x$, $x \in C$. In case of first-countable spaces, we can give an intuitive characterization of semi-continuous (and continuous) functions in terms of sequences.

Proposition 2.7. Let $\mathbf{X}$ be a first-countable space and $f : \mathbf{X} \to \overline{\mathbb{R}}$ a function. Then:

(i) f is upper semi-continuous if and only if

$$\limsup_{n \to \infty} f(x_n) \leq f(x) \qquad \forall x_n \to x. \tag{2.3}$$

(ii) f is lower semi-continuous if and only if

$$\liminf_{n \to \infty} f(x_n) \geq f(x) \qquad \forall x_n \to x. \tag{2.4}$$

(iii) f is continuous if and only if

$$\lim_{n \to \infty} f(x_n) = f(x) \qquad \forall x_n \to x. \tag{2.5}$$

Proof. (i) Assume that f is upper semi-continuous, $x_n \to x$ and let

$$a \doteq \limsup_{n \to \infty} f(x_n).$$

Then there exists a subsequence x_{n_k} such that $f(x_{n_k}) \to a$. We distinguish three cases:

- If $a = -\infty$, we do not have anything to prove.
- If $a \in \mathbb{R}$, let $\varepsilon > 0$; then there exists K such that $f(x_{n_k}) \geq a - \varepsilon \ \forall k \geq K$, i.e.

$$x_{n_k} \in f^{-1}\Big([a-\varepsilon, +\infty]\Big) \qquad \forall k \geq K.$$

  Since f is upper semi-continuous, the latter set is closed, and so it contains $\lim_{n\to\infty} x_{n_k} = x$. In other words, $f(x) \geq a - \varepsilon$; since $\varepsilon > 0$ is arbitrary, we conclude that $f(x) \geq a$.
- If $a = +\infty$, let $M \in \mathbb{R}$; then there exists K such that $f(x_{n_k}) \geq M \ \forall k \geq K$. As before, this implies that $f(x) \geq M$; since $M \in \mathbb{R}$ is arbitrary, we conclude that $f(x) = +\infty = a$.

Thus, we have proved the first implication. Conversely, assume (2.3) and let $a \in \mathbb{R}$: we will prove that $f^{-1}\Big([a, +\infty]\Big)$ is closed. Since $\mathbf{X}$ is a first-countable space, it suffices to prove that, if $x_n \to x$ and $x_n \in f^{-1}\Big([a, +\infty]\Big) \ \forall n$ (that is $f(x_n) \geq a \ \forall n$), then also $x \in f^{-1}\Big([a, +\infty]\Big)$. But this is true because, thanks to (2.3),

$$f(x) \geq \limsup_{n\to\infty} f(x_n) \geq a.$$

(ii) It follows from Remark 2.4 and the first part of this proposition.

(iii) By Proposition 2.5 (i) and the first two statements of this proposition, f is continuous if and only if (2.3) and (2.4) hold, i.e. if and only if (2.5) holds. □

Remark 2.8. From the proof of Proposition 2.7 it is clear that the hypothesis of first-countability is not needed to conclude that (2.3), (2.4) and (2.5) are *necessary* conditions for upper semi-continuity, lower semi-continuity and continuity respectively. ◊

We now want to study how to approximate semi-continuous functions with simple semi-continuous functions. First of all, we recall the definition of simple function.

Definition 2.9. Let $\mathbf{X}$ be a topological space. A function $f : \mathbf{X} \to \mathbb{R}$ is said to be *simple* if it assumes only finitely many values. We also define

$$S(\mathbf{X}) \doteq \Big\{ f : \mathbf{X} \to \mathbb{R} : \quad f \text{ simple} \Big\}.$$

Lemma 2.10. Let $\mathbf{X}$ be a topological space.

(i) $f \in U_{b,+}(\mathbf{X}) \cap S(\mathbf{X})$ if and only if it can be written in one of the following ways:

$$f = \sum_{i=1}^{N} c_i \mathbb{1}_{C_i} \qquad\qquad f = \max_{1\leq i\leq N} c_i' \mathbb{1}_{C_i}$$

where $c_i, c_i' \geq 0$ and C_i is closed for all i.

(ii) $f \in L_{b,+}(\mathbf{X}) \cap S(\mathbf{X})$ if and only if it can be written in one of the following ways:

$$f = \sum_{i=1}^{N} o_i \mathbb{1}_{O_i} \qquad f = \max_{1\le i\le N} o_i' \mathbb{1}_{O_i}$$

where $o_i, o_i' \ge 0$ and O_i is open for all i.

Proof. (i) If f is simple, nonnegative and bounded, then we can write

$$f = \sum_{i=1}^{N} a_i \mathbb{1}_{A_i} \tag{2.6}$$

where $0 \le a_1 < a_2 < \cdots < a_N < +\infty$ and $\{A_i\}_{i=1}^N$ is a partition of $\mathbf{X}$. Let $a_0 \doteq 0$ and

$$g \doteq \sum_{i=1}^{N} (a_i - a_{i-1}) \mathbb{1}_{\{f \ge a_i\}},$$
$$h \doteq \max_{1\le i\le N} a_i \mathbb{1}_{\{f\ge a_i\}}.$$

If f is also upper semi-continuous, we have that $\{f \ge a_i\}$ is closed, $a_i - a_{i-1} \ge 0$ and $a_i \ge 0$ for all i. Moreover, if $x \in A_i$, then $f(x) = a_i$ and

$$f(x) \ge a_1, \dots, f(x) \ge a_i, f(x) < a_{i+1}, \dots f(x) < a_N.$$

Therefore

$$g(x) = (a_1 - a_0) + (a_2 - a_1) + \cdots + (a_i - a_{i-1}) + 0 + \cdots + 0 = a_i - a_0 = a_i,$$
$$h(x) = \max\{a_1, a_2, \dots, a_i, 0, \dots, 0\} = a_i$$

and $f = g = h$ on A_i. Since $\{A_i\}_{i=1}^N$ is a partition of $\mathbf{X}$, $f = g = h$, which proves the first implication.

Conversely, we begin observing that $\mathbb{1}_C$ is upper semi-continuous if C is closed, since

$$\{\mathbb{1}_C \ge a\} = \begin{cases} \mathbf{X} & \text{if } a \le 0, \\ C & \text{if } 0 < a \le 1, \\ \varnothing & \text{if } a > 1. \end{cases}$$

If $f = \sum_{i=1}^N c_i \mathbb{1}_{C_i}$, where $c_i \ge 0$ and C_i is closed for all i, then f is upper semi-continuous by (iii) and (iv) of Proposition 2.5; therefore, $f \in U_{b,+}(\mathbf{X}) \cap S(\mathbf{X})$. If $f = \max_{1\le i\le N} c_i \mathbb{1}_{C_i}$, where $c_i \ge 0$ and C_i is closed for all i, then f is upper semi-continuous by (iii) and (vi) of Proposition 2.5; hence, $f \in U_{b,+}(\mathbf{X}) \cap S(\mathbf{X})$.

(ii) The proof of the second claim is analogous and it suffices to note that:

- $\{f \geq a_i\} = \{f > a_{i-1}\}$ is open for all i, if f is a lower semi-continuous function defined as in (2.6).
- $\mathbb{1}_O$ is lower semi-continuous, if O is open. □

Proposition 2.11 (Approximation with simple semi-continuous functions). Let $\mathbf{X}$ be a topological space.

(i) For each $f \in U(\mathbf{X})$, there exists a sequence $f_k \in U(\mathbf{X}) \cap S(\mathbf{X})$ such that $f_k \searrow f$. An analogous statement holds for $f \in U_{b,+}(\mathbf{X})$.

(ii) For each $f \in L(\mathbf{X})$, there exists a sequence $f_k \in L(\mathbf{X}) \cap S(\mathbf{X})$ such that $f_k \nearrow f$. An analogous statement holds for $f \in L_{b,+}(\mathbf{X})$.

Proof. (i) Let $f \in U(\mathbf{X})$, $r_- \doteq \inf_{x\in\mathbf{X}} f(x)$ and $r_+ \doteq \sup_{x\in\mathbf{X}} f(x)$. If $r_- = r_+$, f is constant and the claim is obvious. Let now $r_- < r_+$ and let $\{a_k\}_{k\in\mathbb{N}}$ be an enumeration of the countable set $\mathbb{Q} \cap (r_-, r_+)$. Rearranging $a_0, \dots, a_k$, we may suppose $a_0 < \dots < a_k$ and define the following simple function

$$f_k(x) \doteq \begin{cases} a_0 & f(x) < a_0, \\ a_i & a_{i-1} \leq f(x) < a_i, \quad i = 1, \dots, k, \\ r_+ & a_k \leq f(x). \end{cases}$$

Then $f_k \in U(\mathbf{X})$, since each set $\{f_k \geq a\}$, different from $\emptyset$ and $\mathbf{X}$, can be written as $\{f \geq a_i\}$ for a suitable i, which is closed because f is upper semi-continuous. Moreover, since $\{a_0, \dots, a_k\} \subseteq \{a_0, \dots, a_{k+1}\}$, we have that $f_k \geq f_{k+1}$ for all k. It remains to show that, for all $x \in \mathbf{X}$, $f_k(x) \to f(x)$ as $k \to \infty$. We distinguish the following three cases:

- If $f(x) = r_+$, $f_k(x) = r_+ \ \forall k$, hence the claim is obvious.
- If $-\infty < f(x) < r_+$, let $\varepsilon > 0$: there exists $\bar{k}$ such that $f(x) < a_{\bar{k}} < f(x) + \varepsilon$, since $\{a_k\}_{k\in\mathbb{N}}$ is dense in (r_-, r_+). Then by the definition of f_k

$$f(x) < f_k(x) \leq a_{\bar{k}} < f(x) + \varepsilon \qquad \forall k \geq \bar{k},$$

  which proves that $f_k(x) \to f(x)$.
- If $f(x) = -\infty = r_-$, then by the definition of f_k

$$\lim_{k\to\infty} f_k(x) = \lim_{k\to\infty} \inf_{1\leq i\leq k} a_i = \inf\{a_i\}_{i\in\mathbb{N}} = \inf(-\infty, r_+) \cap \mathbb{Q} = -\infty = f(x).$$

In particular, if $f \in U_{b,+}(\mathbf{X})$, clearly also $f_k \in U_{b,+}(\mathbf{X})$.

(ii) It follows from Remark 2.4 and the first part of this proposition. □

We conclude this section with another useful result about approximation of indicator functions of closed and open sets.

Proposition 2.12 (Approximation of indicator functions)**.** Let $\mathbf{X}$ be a metrizable space.

(i) If $C \subseteq \mathbf{X}$ is a closed set, then there exists a sequence $f_k \in C_{b,+}(\mathbf{X})$ such that $f_k \searrow \mathbb{1}_C$.

(ii) If $O \subseteq \mathbf{X}$ is an open set, then there exists a sequence $f_k \in C_{b,+}(\mathbf{X})$ such that $f_k \nearrow \mathbb{1}_O$.

Proof. Let d be any distance that induces the topology on $\mathbf{X}$.

(i) Let C closed and $f_k(x) \doteq (1 - kd(x,C)) \vee 0$. Then the f_k are continuous, take values in $[0,1]$ and are decreasing. Moreover, if $x \in C$ then $d(x,C) = 0$ and $1 - kd(x,C) = 1\ \forall k$, hence $\lim_{k\to\infty} f_k(x) = 1 = \mathbb{1}_C(x)$; if $x \notin C$ then $d(x,C) > 0$ (since C is closed) and $1 - kd(x,C) \to -\infty$, hence $\lim_{k\to\infty} f_k(x) = 0 = \mathbb{1}_C(x)$. Therefore $f_k \searrow \mathbb{1}_C$.

(ii) If O is open, then O^c is closed, hence by (i) there exists a sequence of functions $f_k \in C_{b,+}(\mathbf{X})$ that take values in $[0,1]$, such that $f_k \searrow \mathbb{1}_C$. Therefore, also $(1 - f_k) \in C_{b,+}(\mathbf{X})$ and $(1 - f_k) \nearrow (1 - \mathbb{1}_{O^c}) = \mathbb{1}_O$. □

2.2 Polish spaces and weak convergence

In this section, we deal with weak convergence of probability measures, which will reveal a strong formal analogy with large deviations principles. We also introduce Polish spaces, which are the most natural and interesting environment to work with, in both contexts.

If $(\mathbf{X},\tau)$ is a topological space, we write $\mathcal{B}(\mathbf{X}) = \sigma(\tau)$ for the Borel σ-algebra on $\mathbf{X}$, i.e. the σ-algebra that is generated by τ. We write $B(\mathbf{X})$ for the vector space of Borel functions (i.e. measurable functions $(\mathbf{X},\mathcal{B}(\mathbf{X})) \to (\overline{\mathbb{R}},\mathcal{B}(\overline{\mathbb{R}}))$), $B_b(\mathbf{X})$ for its vector subspace of bounded functions and $B_{b,+}(\mathbf{X})$ for its subset of nonnegative bounded functions. We also write $\mathfrak{M}(\mathbf{X})$ for the set of all probability measures on $(\mathbf{X},\mathcal{B}(\mathbf{X}))$. If $p \in (0,\infty)$ and $\mu \in \mathfrak{M}(\mathbf{X})$, we define the quantity

$$\|f\|_{p,\mu} \doteq \left(\int_{\mathbf{X}} |f(x)|^p\, \mu(dx)\right)^{1/p} \in [0,+\infty] \qquad \forall f \in B(\mathbf{X}).$$

We also set

$$\|f\|_{\infty,\mu} \doteq \inf\{c : \quad |f(x)| \le c \quad \text{for } \mu\text{-a.e. } x \in \mathbf{X}\} \in [0,+\infty] \qquad \forall f \in B(\mathbf{X}).$$

It is well-known that, for any $p \in [1,+\infty]$:

- If $L^p(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ denotes the vector space of Borel functions f such that $\|f\|_{p,\mu}$ is finite, $\|\cdot\|_{p,\mu}$ is a seminorm on $L^p(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$.

- If $L^p(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ is considered as a quotient space, by identifying functions that are equal μ-a.s., then $\|\cdot\|_{p,\mu}$ is a norm on $L^p(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ that makes it into a Banach space.

Definition 2.13. Let $(\mathbf{X}, \tau)$ be a topological space. We call *weak topology* on $\mathfrak{M}(\mathbf{X})$, and we denote it by τ_w, the coarsest topology that makes continuous all maps $\mathfrak{M}(\mathbf{X}) \to \mathbb{R}$, $\mu \to \|f\|_{1,\mu}$ for $f \in C_{b,+}(\mathbf{X})$†. We say that a sequence $\mu_n \in \mathfrak{M}(\mathbf{X})$ *converges weakly* to a $\mu \in \mathfrak{M}(\mathbf{X})$, and we write $\mu_n \rightharpoonup \mu$, if the limit holds in the weak topology, i.e. if

$$\lim_{n\to\infty} \|f\|_{1,\mu_n} = \|f\|_{1,\mu} \qquad \forall f \in C_{b,+}(\mathbf{X}). \tag{2.7}$$

Example 2.14. Let $\mathbf{X} = \mathbb{R}$ and μ_n the uniform distribution on $\left[-\frac{1}{n}, \frac{1}{n}\right]$. We expect that $\mu_n \rightharpoonup \delta_0$. If $f \in C_{b,+}(\mathbb{R})$,

$$\|f\|_{1,\mu_n} = \frac{n}{2} \int_{-\frac{1}{n}}^{\frac{1}{n}} f(x)\,dx = f(t_n)$$

with $t_n \in \left[-\frac{1}{n}, \frac{1}{n}\right]$, by the mean value theorem for integrals, and

$$\lim_{n\to\infty} f(t_n) = f(0) = \|f\|_{1,\delta_0}.$$

Therefore, $\mu_n \rightharpoonup \delta_0$. Note that the convergence of the norms is not guaranteed for measurable bounded nonnegative *discontinuous* functions: for instance,

$$\|\mathbb{1}_{(-\infty,0]}\|_{1,\mu_n} = \frac{n}{2} \int_{-\frac{1}{n}}^{0} dx = \frac{1}{2},$$

but $\|\mathbb{1}_{(-\infty,0]}\|_{1,\delta_0} = 1$. ◊

Definition 2.15. Let $(\mathbf{X}, \tau)$ be a topological space.

- $(\mathbf{X}, \tau)$ is said to be *separable* if it contains a countable dense subset.

- $(\mathbf{X}, \tau)$ is said to be *completely metrizable* if there exists at least one metric d on $\mathbf{X}$ such that d induces the topology τ and $(\mathbf{X}, d)$ is a complete metric space.

- $(\mathbf{X}, \tau)$ is said to be *Polish* †† if it is a separable completely metrizable topological space.

---

†It is equivalent to define τ_w as the coarsest topology that makes continuous all maps $\mathfrak{M}(\mathbf{X}) \to \mathbb{R}$, $\mu \to \int_{\mathbf{X}} f\,d\mu$ for $f \in C_b(\mathbf{X})$. Indeed, if $f \in C_b(\mathbf{X})$, then $f^+, f^- \in C_{b,+}(\mathbf{X})$ and $f = f^+ - f^-$.

††These spaces owe their weird name to the nationality of the first mathematicians who extensively studied them, such as Sierpiński, Kuratowski and Tarski.

We now give several examples of Polish spaces, which we will often use in this work.

Example 2.16. $\mathbb{R}^d$, equipped with the euclidean topology, is a Polish space. Indeed, it contains the countable dense subset $\mathbb{Q}^d$ and it is complete for the Euclidean metric $d(x,y) = \|x-y\|_2$, which induces the Euclidean topology.

However, note that $\mathbb{R}^d$ is not complete in other metrics that induce the same topology. For instance, consider $\mathbb{R}$ equipped with the metric

$$d'(x,y) = |\arctan(x) - \arctan(y)|.$$

Since $\arctan\colon \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2})$ is an homeomorphism, if $x_n, x \in \mathbb{R}$ it turns out that

$$|x_n - x| \to 0 \qquad \Longleftrightarrow \qquad |\arctan(x_n) - \arctan(x)| \to 0,$$

hence

$$d(x_n, x) \to 0 \qquad \Longleftrightarrow \qquad d'(x_n, x) \to 0.$$

This shows that d and d' induce the same topology on $\mathbb{R}$ (i.e., the Euclidean topology). On the other hand, $(\mathbb{R}, d')$ is not complete: consider the sequence $x_n = n$ and any arbitrary $\varepsilon > 0$. If one takes N such that

$$\frac{\pi}{2} - \arctan(N) < \varepsilon,$$

it turns out that

$$d'(x_n, x_m) = |\arctan(n) - \arctan(m)| \leq \frac{\pi}{2} - \arctan(N) < \varepsilon \qquad \forall n, m \geq N.$$

Therefore, x_n is a Cauchy sequence for d'; however, x_n does not converge in the Euclidean topology of $\mathbb{R}$. ◊

Example 2.17. In general, any separable Banach space is Polish, for example:

(i) If $\mathbf{X}$ is a Polish space, then the space $L^p(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$, equipped with the topology of the p-norm, is a separable Banach space for $p \in [1, \infty)$ (see [Bre, Th.4.8 and 4.13]), hence it is Polish.

(ii) If $\mathbf{X}$ is a compact space, by Weiestrass's Theorem every continuous function $\mathbf{X} \to \mathbb{R}$ is bounded, hence $C(\mathbf{X})$ can be equipped with the supremumnorm, given by

$$\|f\|_\infty \doteq \sup_{x \in \mathbf{X}} |f(x)| \qquad \forall f \in C(\mathbf{X}).$$

If $\mathbf{X}$ is also metrizable, it turns out that $(C(\mathbf{X}), \|\cdot\|_\infty)$ is a separable Banach space (see [Kec, Th.4.19]), hence it is Polish. ◊

Example 2.18. If $\mathbf{X}$ is a compact metrizable space, then $\mathbf{X}$ is Polish. Indeed:

- Let d be any metric on $\mathbf{X}$ that generates its topology. Let us prove that $(\mathbf{X}, d)$ is complete. Let x_n be a Cauchy sequence in $\mathbf{X}$ and let $\varepsilon > 0$: there exists $\overline{n}$ large enough such that $d(x_n, x_m) < \frac{\varepsilon}{2}$ for all $n, m \geq \overline{n}$. Since $\mathbf{X}$ is metrizable and compact, there exists a subsequence x_{n_k} that converges to an element $x \in \mathbf{X}$, hence we can choose $\overline{k}$ such that $n_{\overline{k}} \geq \overline{n}$ and $d(x_{n_{\overline{k}}}, x) < \frac{\varepsilon}{2}$. Thus, by the triangular inequality

$$d(x_n, x) \leq d(x_n, x_{n_{\overline{k}}}) + d(x_{n_{\overline{k}}}, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \qquad \forall n \geq \overline{n}.$$

This proves that $x_n \to x$, hence $(\mathbf{X}, d)$ is complete.

- Let us prove that $\mathbf{X}$ is separable. For all $n \geq 1$, $\{B_{1/n}(x)\}_{x \in \mathbf{X}}$ is an open cover of $\mathbf{X}$. Since $\mathbf{X}$ is compact, there exists $D_n \subseteq \mathbf{X}$ finite such that $\{B_{1/n}(x)\}_{x \in D_n}$ is a cover of $\mathbf{X}$ again. If we set $D \doteq \bigcup_{n \geq 1} D_n$, D is countable, since it is a countable union of finite sets. We claim that D is also dense in $\mathbf{X}$. If we fix $\overline{x} \in \mathbf{X}$ and $\varepsilon > 0$, there exists $\overline{n} \geq 1$ such that $1/\overline{n} < \varepsilon$. Since

$$\overline{x} \in \mathbf{X} = \bigcup_{x \in D_{\overline{n}}} B_{1/\overline{n}}(x),$$

there exists $x \in D_{\overline{n}} \subseteq D$ such that $\overline{x} \in B_{1/\overline{n}}(x)$. Therefore, we have found $x \in D$ such that $d(\overline{x}, x) < 1/\overline{n} < \varepsilon$. This proves that D is a dense subset of $\mathbf{X}$. We conclude that $\mathbf{X}$ is separable. ◊

Example 2.19. If $(\mathbf{X}, \tau)$ is a Polish space, then the space $(\mathfrak{M}(\mathbf{X}), \tau_w)$ of the probability measures on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$, equipped with the weak topology, is also Polish (see [EtKu, §3.1 and Th.3.1.7]). ◊

It may be useful to know under which topological conditions a subspace of a Polish space is Polish: the next theorem answers this question.

Definition 2.20. Let $\mathbf{X}$ be a topological space. $\mathbf{Y} \subseteq \mathbf{X}$ is said to be a G_δ-subset if it is a countable intersection of open sets.

Theorem 2.21. If $\mathbf{X}$ is a Polish space, $\mathbf{Y} \subseteq \mathbf{X}$ is Polish if and only it is a G_δ-subset of $\mathbf{X}$.

Proof. See [Bou, Th.1, §6, Ch.IX]. □

Example 2.22. Let $\mathbf{X}$ be a Polish space. We now give some examples of subsets of $\mathbf{X}$ that are, or are not, G_δ-subset, i.e. Polish subspaces by Theorem 2.21.

- If $O \subseteq \mathbf{X}$ is open, then O is obviously G_δ.

- If $C \subseteq \mathbf{X}$ is closed, then C is G_δ. Indeed, if d is any metric on $\mathbf{X}$ and $B_R(x)$ denotes the ball of radius $R > 0$ and center $x \in \mathbf{X}$ in this metric, we claim that
$$C = \bigcap_{k=1}^{\infty} \bigcup_{x \in C} B_{1/k}(x),$$
which proves that C is G_δ, since $\{\bigcup_{x\in C} B_{1/k}(x)\}_{k\geq 1}$ is a countable collection of open sets. To prove the equality, we note that if $\overline{x} \in C$, then $\overline{x}$ is obviously in the right-hand set. Conversely, assume that $\overline{x}$ is in the right-hand set but $x \notin C$; since C is closed, $d(\overline{x}, C) > 0$, hence we can choose $k \geq 1$ such that $1/k < d(\overline{x}, C)$; by hypothesis, there exists $x \in C$ such that $\overline{x} \in B_{1/k}(x)$, hence $d(\overline{x}, x) < 1/k < d(\overline{x}, C)$, which leads to a contradiction.

- Let $\mathbf{X} = \mathbb{R}$. Then, the subset $\mathbb{R} \setminus \mathbb{Q}$ of irrational numbers is G_δ, since
$$\mathbb{R} \setminus \mathbb{Q} = \mathbb{R} \setminus \bigcup_{q\in\mathbb{Q}} \{q\} = \bigcap_{q\in\mathbb{Q}} \mathbb{R} \setminus \{q\},$$
and $\{\mathbb{R} \setminus \{q\}\}_{q\in\mathbb{Q}}$ is a countable collection of open sets. On the other hand, $\mathbb{Q}$ is not G_δ: if there exists a countable collection of open sets $\{O_k\}_{k\geq 1}$ such that $\mathbb{Q} = \bigcap_{k\geq 1} O_k$, then each O_k is dense in $\mathbb{R}$ (since $\mathbb{Q}$ is so), and
$$\varnothing = \mathbb{Q} \cap (\mathbb{R} \setminus \mathbb{Q}) = \bigcap_{k\geq 1} O_k \cap \bigcap_{q\in\mathbb{Q}} \mathbb{R} \setminus \{q\}$$
is a countable intersection of open dense subsets of the complete metric space $\mathbb{R}$, hence it is dense by the Baire category Theorem (see [Bou, Th.1, §5, ch.IX]); this leads to a contradiction, since $\varnothing$ is not dense in $\mathbb{R}$. Therefore, $\mathbb{Q}$ is not a Polish space: since, as a countable set, it is separable, this means in particular that the Euclidean topology of $\mathbb{Q}$ is *not* completely metrizable. ◊

From now on in this section, $\mathbf{X}$ will refer to a Polish space. The rest of this section is devoted to discuss and prove the following characterization of the weak convergence of probability measures on a Polish space.

Theorem 2.23. $\mu_n \rightharpoonup \mu$ in $\mathfrak{M}(\mathbf{X})$ if and only if one of the following equivalent conditions is satisfied.

(i) $\limsup_{n\to\infty} \mu_n(C) \leq \mu(C) \qquad \forall C \subseteq \mathbf{X}$ closed.

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

Remark 2.24. Assume that $\mu_n \rightharpoonup \mu$ and $A \subseteq \mathbf{X}$ is a *continuity set* of μ, i.e. $\mu(\partial A) = 0$. Then, by theorem 2.23
$$\begin{aligned}\limsup_{n\to\infty} \mu_n(A) &\leq \limsup_{n\to\infty} \mu_n(\overline{A}) \leq \mu(\overline{A}) = \mu(\mathring{A} \dot\cup \partial A) \\ &= \mu(\mathring{A}) + \mu(\partial A) = \mu(\mathring{A}) \leq \liminf_{n\to\infty} \mu_n(\mathring{A}) \leq \liminf_{n\to\infty} \mu_n(A),\end{aligned}$$

since $\overline{A}$ is closed and $\mathring{A}$ is open. Therefore, $\mu_n(A)$ has limit $\mu(\overline{A}) = \mu(\mathring{A}) = \mu(A)$. So we have just proved that, if $\mu_n \rightharpoonup \mu$, then

$$\lim_{n\to\infty} \mu_n(A) = \mu(A) \qquad \forall A \subseteq \mathbf{X} \text{ continuity set of } \mu.$$

Let now $\mathbf{X} = \mathbb{R}$; if $(-\infty, x]$ is a continuity set of μ, i.e. $\mu(\partial(-\infty, x]) = \mu(\{x\}) = 0$, then $\lim_{n\to\infty} \mu_n((-\infty, x]) = \mu((-\infty, x])$. In other words, $\lim_{n\to\infty} F_{\mu_n}(x) = F_\mu(x)$ for all points $x \in \mathbb{R}$ at which F_μ is continuous (F_{μ_n} and F_μ denote the distribution functions of μ_n and μ). The latter is a possible definition of weak convergence of probability measures on $\mathbb{R}$, because one can prove that it is also a sufficient condition to guarantee (2.7). ♢

We now define the *good sets* we will need for the proof of Theorem 2.23, and we show their properties. If $\mu_n, \mu \in \mathfrak{M}(\mathbf{X})$, we set

$$G_{up} \doteq \Big\{ f \in U_{b,+}(\mathbf{X}) : \quad \limsup_{n\to\infty} \|f\|_{1,\mu_n} \le \|f\|_{1,\mu} \Big\},$$
$$G_{low} \doteq \Big\{ f \in L_{b,+}(\mathbf{X}) : \quad \liminf_{n\to\infty} \|f\|_{1,\mu_n} \ge \|f\|_{1,\mu} \Big\}.$$

Lemma 2.25. (i) If $f \in G_{up}$ and $\lambda \ge 0$, then $\lambda f \in G_{up}$ (resp., if $f \in G_{low}$ and $\lambda \ge 0$, then $\lambda f \in G_{low}$).

(ii) If $f, g \in G_{up}$, then $f + g \in G_{up}$ (resp., if $f, g \in G_{low}$, then $f + g \in G_{low}$).

(iii) If $f_k \searrow f$ and $f_k \in G_{up}$ $\forall k$, then $f \in G_{up}$ (resp., if $f_k \nearrow f$, $f_k \in G_{low}$ $\forall k$ and f is bounded, then $f \in G_{low}$).

Proof. (i) If $f \in G_{up}$ and $\lambda \ge 0$, then λf is upper semi-continuous (by Proposition 2.5 (iii)), nonnegative and bounded, hence $\lambda f \in U_{b,+}(\mathbf{X})$. Moreover, since $\|\cdot\|_{1,\mu_n}$ and $\|\cdot\|_{1,\mu}$ are norms,

$$\limsup_{n\to\infty} \|\lambda f\|_{1,\mu_n} = |\lambda| \limsup_{n\to\infty} \|f\|_{1,\mu_n} \le |\lambda| \|f\|_{1,\mu} = \|\lambda f\|_{1,\mu},$$

which proves that $\lambda f \in G_{up}$. The proof for $f \in G_{low}$ is similar.

(ii) If $f, g \in G_{up}$, then $f + g$ is upper semi-continuous (by Proposition 2.5 (iv)), nonnegative and bounded, hence $f + g \in U_{b,+}(\mathbf{X})$. The inequality

$$\limsup_{n\to\infty} \|f + g\|_{1,\mu_n} \le \|f + g\|_{1,\mu}$$

holds because for all $\nu \in \mathfrak{M}(\mathbf{X})$ and for nonnegative f, g

$$\|f + g\|_{1,\nu} = \int_{\mathbf{X}} (f + g)\, d\nu = \int_{\mathbf{X}} f\, d\nu + \int_{\mathbf{X}} g\, d\nu = \|f\|_{1,\nu} + \|g\|_{1,\nu}.$$

This proves that $f + g \in G_{up}$. The proof for $f, g \in G_{low}$ is similar.

(iii) Let $f_k \in G_{up}$ $\forall k$, $f_k \searrow f$ (in particular, $f \leq f_k \leq f_1$ $\forall k$). Then, f is upper semi-continuous by Proposition 2.5 (vii), nonnegative and bounded. Moreover,

$$\limsup_{n\to\infty} \|f\|_{1,\mu_n} \leq \limsup_{n\to\infty} \|f_k\|_{1,\mu_n} \leq \|f_k\|_{1,\mu}$$

since $f_k \in G_{up}$. By dominated convergence, $\lim_{k\to\infty} \|f_k\|_{1,\mu} = \|f\|_{1,\mu}$, hence $\limsup_{n\to\infty} \|f\|_{1,\mu_n} \leq \|f\|_{1,\mu}$, which proves that $f \in G_{up}$. The proof for $f_k \nearrow f$, $f_k \in G_{low}$ $\forall k$ and f bounded is similar, using monotone convergence instead of dominated convergence. □

Proof of Theorem 2.23. Obviously, (i) and (ii) are equivalent, by taking complements.

We first assume that $\mu_n \rightharpoonup \mu$ and prove, for example, (i). If $C \subseteq \mathbf{X}$ is closed, by Proposition 2.12 (i) there exists a sequence $f_k \in C_{b,+}(\mathbf{X})$ such that $f_k \searrow \mathbb{1}_C$. Then $f_k \in G_{up}$ by the fact that $\mu_n \rightharpoonup \mu$ and therefore, by Lemma 2.25 (iii), it follows that $\mathbb{1}_C \in G_{up}$, which proves (i).

Conversely, assume that (i) and (ii) hold. For each closed $C \subseteq \mathbf{X}$, $\mathbb{1}_C$ is upper semi-continuous (by Lemma 2.10 (i)), bounded and nonnegative; moreover, condition (i) implies that

$$\limsup_{n\to\infty} \|\mathbb{1}_C\|_{1,\mu_n} \leq \|\mathbb{1}_C\|_{1,\mu},$$

hence $\mathbb{1}_C \in G_{up}$. By Lemma 2.10 (i) again and properties (i) and (ii) of Lemma 2.25, we obtain $S(\mathbf{X}) \cap U_{b,+}(\mathbf{X}) \subseteq G_{up}$. If $f \in U_{b,+}(\mathbf{X})$, then by Proposition 2.11 (i) there exists a sequence $f_k \in S(\mathbf{X}) \cap U_{b,+}(\mathbf{X}) \subseteq G_{up}$ such that $f_k \searrow f$, hence by Lemma 2.25 (iii) $f \in G_{up}$ again. This proves that $U_{b,+}(\mathbf{X}) = G_{up}$; similarly one can prove that $L_{b,+}(\mathbf{X}) = G_{low}$. If now

$$f \in C_{b,+}(\mathbf{X}) = U_{b,+}(\mathbf{X}) \cap L_{b,+}(\mathbf{X}) = G_{up} \cap G_{low},$$

then $\lim_{n\to\infty} \|f\|_{1,\mu_n} = \|f\|_{1,\mu}$. We conclude that $\mu_n \rightharpoonup \mu$. □

2.3 Large deviations principles

We can now talk about large deviations in a general context.

In this section $\mathbf{X}$ will always refer to a Polish space. We call *level sets* of a function $f : \mathbf{X} \to \overline{\mathbb{R}}$ the sets $\{f \leq a\}$, for any $a \in \mathbb{R}$.

Definition 2.26. $I : \mathbf{X} \to \overline{\mathbb{R}}$ is said to be a *rate function* if:

(i) $I \geq 0$ and $I \not\equiv +\infty$.

(ii) I is lower semi-continuous (i.e. it has closed level sets).

I is said to be a *good rate function* if in addition:

(iii) I has compact level sets.

In this work we shall only be concerned with *good* rate functions. We note that in the latter definition (iii) implies (ii): indeed, if a level set of I is compact then it is closed, since $\mathbf{X}$, as a metrizable space, is Hausdorff.

We recall that we let $B_b(\mathbf{X})$ denote the space of bounded Borel functions $\mathbf{X} \to \mathbb{R}$.

Definition 2.27. If I is a rate function, we define $\forall f \in B_b(\mathbf{X})$

$$\|f\|_{\infty,I} \doteq \sup_{x\in\mathbf{X}} e^{-I(x)}|f(x)|.$$

Remark 2.28. We note that $\|\cdot\|_{\infty,I}$ is a seminorm on $B_b(\mathbf{X})$, since $\forall f,g \in B_b(\mathbf{X})$, $\lambda \in \mathbb{R}$

(i) $\|f\|_{\infty,I} \in [0,+\infty)$ (since $I \geq 0$ and f is bounded).

(ii) $\|\lambda f\|_{\infty,I} = |\lambda|\,\|f\|_{\infty,I}$.

(iii) $\|f+g\|_{\infty,I} \leq \|f\|_{\infty,I} + \|g\|_{\infty,I}$.

If moreover $I < +\infty$, then $\|\cdot\|_{\infty,I}$ is a norm, since $\forall f \in B_b(\mathbf{X})$ it satisfies

(iv) $\|f\|_{\infty,I} = 0 \quad \implies \quad f = 0.$ ◊

Definition 2.29 (Large deviations principle)**.** Let s_n be a sequence such that $s_n \geq 1$ $\forall n$ and $s_n \to +\infty$, and let I be a good rate function on $\mathbf{X}$. We say that a sequence $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the *large deviations principle* (LDP) with *rate* s_n and *good rate function* I if

$$\lim_{n\to\infty} \|f\|_{s_n,\mu_n} = \|f\|_{\infty,I} \qquad \forall f \in C_{b,+}(\mathbf{X}). \tag{2.8}$$

Example 2.30. We now discuss the most trivial example of LDP (more interesting examples will be given later), i.e. we consider the case of a constant sequence $\mu_n = \mu \in \mathfrak{M}(\mathbf{X})$. Let $1 \leq s_n \to \infty$. We recall that

$$\lim_{p\to\infty} \|f\|_{p,\mu} = \|f\|_{\infty,\mu} \qquad \forall f \in L^\infty(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu),$$

which is also one of the facts that justify the notation $\|\cdot\|_{\infty,\mu}$. Therefore, for any $f \in C_{b,+}(\mathbf{X})$,

$$\lim_{n\to\infty} \|f\|_{s_n,\mu_n} = \lim_{n\to\infty} \|f\|_{s_n,\mu} = \|f\|_{\infty,\mu} = \sup_{x\in\operatorname{supp}(\mu)} f(x) = \sup_{x\in X} e^{-I(x)} f(x) = \|f\|_{\infty,I}$$

(we explicitly remark that the third equality holds because f is continuous), where

$$I(x) = \begin{cases} 0 & x \in \operatorname{supp}(\mu), \\ +\infty & x \notin \operatorname{supp}(\mu). \end{cases}$$

It is clear that I is a *good* rate function if and only if $\operatorname{supp}(\mu) \subseteq \mathbf{X}$ is compact. In this case, $\mu_n = \mu$ satisfies the large deviations principle with any rate s_n and the good rate function I we defined above. ◊

The definition of LDP we have chosen may look weird compared to the 'naive' large deviations theory we saw in chapter 1, but its advantage is that it stresses the similarity between weak convergence of probability measures and large deviations principles. However, the following characterization of LDP, that is its traditional definition, could sound more familiar.

Theorem 2.31. A sequence $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the large deviations principle with rate s_n and good rate function I if and only if the following two conditions are satisfied:

(i) $\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C) \le -\inf_{x\in C} I(x) \qquad \forall C \subseteq \mathbf{X}$ closed.

(ii) $\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(O) \ge -\inf_{x\in O} I(x) \qquad \forall O \subseteq \mathbf{X}$ open.

One can immediately note the analogy between Theorem 2.31 and Theorem 2.23: in fact, we will follow a similar proof pattern. We now comment on the theorem and see some preparatory lemmas before giving the proof.

Remark 2.32. One might ask what happens if we require that a sequence μ_n, instead of the two conditions of Theorem 2.31, satisfies the stronger condition

$$\lim_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) = -\inf_{x\in A} I(x) \qquad \forall A \in \mathcal{B}(\mathbf{X}). \tag{2.9}$$

This would be certainly too restrictive: for instance, if $\mu_n(\{x\}) = 0 \ \forall x \in \mathbf{X}$, $n \in \mathbb{N}$ (it is the case, e.g., of absolutely continuous probability measures on $\mathbb{R}^d$), then (2.9) implies that $\forall x \in \mathbf{X}$

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

Therefore, $I \equiv +\infty$, which is excluded by the definition of rate function.

However, if μ_n satisfies the large deviations principle with rate s_n and rate function I, we can say that (2.9) holds for a special class of Borel sets A, the I-*continuous sets*, i.e. such that

$$\inf_{x\in \overline{A}} I(x) = \inf_{x\in \mathring{A}} I(x). \tag{2.10}$$

Indeed, in this case by Theorem 2.31

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) &\le \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(\overline{A}) \le -\inf_{x\in\overline{A}} I(x) = -\inf_{x\in\mathring{A}} I(x) \\ &\le \liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(\mathring{A}) \le \liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(A).\end{aligned}$$

In many cases, this is a large class of Borel sets: for example, if I is *continuous*, all the sets A such that $\overline{A} = \overline{\mathring{A}}$ (in particular, all the open sets) are I-continuous. To prove this claim, we consider $\overline{x} \in \overline{A} = \overline{\mathring{A}}$: since $\mathbf{X}$ is a first-countable space, there exists a sequence $x_k \in \mathring{A}$ such that $x_k \to \overline{x}$. Therefore, $I(x_k) \geq \inf_{x\in\mathring{A}} I(x)$, and since I is continuous

$$I(\overline{x}) = \lim_{k\to\infty} I(x_k) \geq \inf_{x\in\mathring{A}} I(x).$$

Since $\overline{x} \in \overline{A}$ is arbitrary,

$$\inf_{x\in\overline{A}} I(x) \geq \inf_{x\in\mathring{A}} I(x),$$

and since the other inequality is obvious, (2.10) follows and A is I-continuous. $\diamondsuit$

We recall that $B(\mathbf{X})$, $U(\mathbf{X})$, $L(\mathbf{X})$ and $C(\mathbf{X})$ denote the sets of Borel, upper semi-continuous, lower semi-continuous and continuous functions $\mathbf{X} \to \overline{\mathbb{R}}$ resp.; $B_{b,+}(\mathbf{X})$, $U_{b,+}(\mathbf{X})$, $L_{b,+}(\mathbf{X})$ and $C_{b,+}(\mathbf{X})$ denote their subsets of bounded nonnegative functions.

Lemma 2.33 (Properties of $\|\cdot\|_{\infty,I}$). If I is a good rate function, then:

(i) $\|f \vee g\|_{\infty,I} = \|f\|_{\infty,I} \vee \|g\|_{\infty,I} \qquad \forall f,g \in B_{b,+}(\mathbf{X})$.

(ii) $\|f_k\|_{\infty,I} \nearrow \|f\|_{\infty,I} \qquad \forall f_k \in B_{b,+}(\mathbf{X}),\ f_k \nearrow f$.

(iii) $\|f_k\|_{\infty,I} \searrow \|f\|_{\infty,I} \qquad \forall f_k \in U_{b,+}(\mathbf{X}),\ f_k \searrow f$.

Proof. (i) For all $f,g \in B_{b,+}(\mathbf{X})$

$$\begin{aligned}\|f \vee g\|_{\infty,I} &= \sup_{x\in\mathbf{X}} e^{-I(x)}(f(x) \vee g(x)) \\ &= \Big(\sup_{x\in\mathbf{X}} e^{-I(x)} f(x)\Big) \vee \Big(\sup_{x\in\mathbf{X}} e^{-I(x)} g(x)\Big) = \|f\|_{\infty,I} \vee \|g\|_{\infty,I}.\end{aligned}$$

(ii) Let $f_k \in B_{b,+}(\mathbf{X})$, $f_k \nearrow f$. Since f_k is an increasing sequence, $f_k \leq f\ \forall k$, hence also $\|f_k\|_{\infty,I} \leq \|f\|_{\infty,I}\ \forall k$. If we pass to the limit superior we obtain

$$\limsup_{k\to\infty} \|f_k\|_{\infty,I} \leq \|f\|_{\infty,I}.$$

To estimate the limit inferior from below, let $\varepsilon > 0$ and take $x \in \mathbf{X}$ such that $e^{-I(x)} f(x) \geq \|f\|_{\infty,I} - \varepsilon$. Therefore

$$\liminf_{k\to\infty} \|f_k\|_{\infty,I} \geq \liminf_{k\to\infty} e^{-I(x)} f_k(x) = e^{-I(x)} f(x) \geq \|f\|_{\infty,I} - \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that

$$\liminf_{k\to\infty} \|f_k\|_{\infty,I} \geq \|f\|_{\infty,I}.$$

(iii) Let $f_k \in U_{b,+}(\mathbf{X})$, $f_k \searrow f$. Since f_k is a decreasing sequence, $f_k \geq f$ $\forall k$, hence also $\|f_k\|_{\infty,I} \geq \|f\|_{\infty,I}$ $\forall k$. If we pass to the limit inferior we obtain

$$\liminf_{k\to\infty} \|f_k\|_{\infty,I} \geq \|f\|_{\infty,I}.$$

To estimate the limit superior from above, let $a > \|f\|_{\infty,I} = \sup_{x\in\mathbf{X}} e^{-I(x)} f(x)$. For all k, the function $e^{-I} f_k$ is upper semi-continuous by Proposition 2.5 (v), since e^{-I} and f_k are nonnegative upper semicontinuous functions ($e^{-I} \in U(\mathbf{X})$ because $I \in L(\mathbf{X})$). Therefore, the sets

$$A_k \doteq \{x \in \mathbf{X} : \quad e^{-I(x)} f_k(x) \geq a\}$$

are closed. Moreover, each f_k is bounded, hence there exists $c_k > 0$ such that $0 \leq f_k \leq c_k$; it follows that

$$A_k \subseteq \{x \in \mathbf{X} : \quad e^{-I(x)} c_k \geq a\} = \left\{x \in \mathbf{X} : I(x) \leq -\log\left(\frac{a}{c_k}\right)\right\}.$$

The latter set is compact, since I has compact level sets; therefore, for all k, A_k is a closed subset of a compact, hence it is compact. Thus, A_k is a decreasing sequence of compact subsets of the Hausdorff space $\mathbf{X}$, and their intersection is

$$\bigcap_{k\in\mathbb{N}} A_k = \left\{x \in \mathbf{X} : \quad e^{-I(x)} f_k(x) \geq a \quad \forall k\right\} \subseteq \left\{x \in \mathbf{X} : \quad e^{-I(x)} f(x) \geq a\right\} = \varnothing$$

since $a > \sup_{x\in\mathbf{X}} e^{-I(x)} f(x)$. We conclude (see [Man, Prop.4.43]) that A_k is empty for k large enough; in other words, there exists K such that $\forall k \geq K$ $e^{-I(x)} f_k(x) < a$ $\forall x \in \mathbf{X}$. Therefore, $\forall k \geq K$ $\|f_k\|_{\infty,I} = \sup_{x\in\mathbf{X}} e^{-I(x)} f_k(x) \leq a$, and if we pass to the limit superior we obtain

$$\limsup_{k\to\infty} \|f_k\|_{\infty,I} \leq a.$$

Since $a > \|f\|_{\infty,I}$ is arbitrary, we conclude that

$$\limsup_{k\to\infty} \|f_k\|_{\infty,I} \leq \|f\|_{\infty,I}. \qquad \square$$

Following the same proof strategy of Theorem 2.23, we now define the *good sets* we will need for the proof of Theorem 2.31, and we show their properties. Let $\mu_n \in \mathfrak{M}(\mathbf{X})$, $1 \leq s_n \to \infty$ and let I be a good rate function. We set

$$G_{up} \doteq \left\{f \in U_{b,+}(\mathbf{X}) : \quad \limsup_{n\to\infty} \|f\|_{s_n,\mu_n} \leq \|f\|_{\infty,I}\right\},$$

$$G_{low} \doteq \left\{f \in L_{b,+}(\mathbf{X}) : \quad \liminf_{n\to\infty} \|f\|_{s_n,\mu_n} \geq \|f\|_{\infty,I}\right\}.$$

Lemma 2.34. (i) If $f \in G_{up}$ and $\lambda \geq 0$, then $\lambda f \in G_{up}$ (resp., if $f \in G_{low}$ and $\lambda \geq 0$, then $\lambda f \in G_{low}$).

(ii) If $f, g \in G_{up}$, then $f \vee g \in G_{up}$ (resp., if $f, g \in G_{low}$, then $f \vee g \in G_{low}$).

(iii) If $f_k \searrow f$ and $f_k \in G_{up}$ $\forall k$, then $f \in G_{up}$ (resp., if $f_k \nearrow f$, $f_k \in G_{low}$ $\forall k$ and f is bounded, then $f \in G_{low}$).

Proof. (i) If $f \in G_{up}$ and $\lambda \geq 0$, then λf is upper semi-continuous (by Proposition 2.5 (iii)), nonnegative and bounded, hence $\lambda f \in U_{b,+}(\mathbf{X})$. Moreover, since $\|\cdot\|_{s_n,\mu_n}$ and $\|\cdot\|_{\infty,I}$ are seminorms,

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

which proves that $\lambda f \in G_{up}$. The proof for $f \in G_{low}$ is similar.

(ii) If $f, g \in G_{up}$, then $f \vee g$ is upper semi-continuous (by Proposition 2.5 (vi)), nonnegative and bounded, hence $f \vee g \in U_{b,+}(\mathbf{X})$. Moreover, applying (1.20) with $\alpha_n = \|f\|_{s_n,\mu_n}^{s_n}$ and $\beta_n = \|g\|_{s_n,\mu_n}^{s_n}$, and Lemma 2.33 (i)

$$\begin{aligned}
\limsup_{n\to\infty} \|f \vee g\|_{s_n,\mu_n} &= \limsup_{n\to\infty} \left(\int_{\{f\geq g\}} f^{s_n}\, d\mu_n + \int_{\{f<g\}} g^{s_n}\, d\mu_n \right)^{\frac{1}{s_n}} \\
&\leq \limsup_{n\to\infty} \left(\|f\|_{s_n,\mu_n}^{s_n} + \|g\|_{s_n,\mu_n}^{s_n} \right)^{1/s_n} \\
&= \limsup_{n\to\infty} \left(\|f\|_{s_n,\mu_n} \vee \|g\|_{s_n,\mu_n} \right)(1 + o(1)) \\
&= \left(\limsup_{n\to\infty} \|f\|_{s_n,\mu_n} \right) \vee \left(\limsup_{n\to\infty} \|g\|_{s_n,\mu_n} \right) \\
&\leq \|f\|_{\infty,I} \vee \|g\|_{\infty,I} \\
&= \|f \vee g\|_{\infty,I},
\end{aligned}$$

which proves that $f \vee g \in G_{up}$. If $f, g \in G_{low}$, then similarly $f \vee g \in L_{b,+}(\mathbf{X})$, and since $\|f \vee g\|_{s_n,\mu_n} \geq \|f\|_{s_n,\mu_n} \vee \|g\|_{s_n,\mu_n}$ for all n,

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

by Lemma 2.33 (i), which proves that $f \vee g \in G_{low}$.

(iii) Let $f_k \in G_{up}$ $\forall k$, $f_k \searrow f$ (in particular, $f \leq f_k$ $\forall k$). Then, f is upper semi-continuous by Proposition 2.5 (vii), nonnegative and bounded. Moreover,

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

since $f_k \in G_{up}$. By lemma 2.33 (iii), $\lim_{k\to\infty}\|f_k\|_{\infty,I} = \|f\|_{\infty,I}$, hence $f \in G_{up}$. The proof for $f_k \nearrow f$, $f_k \in G_{low}$ $\forall k$ and f bounded is similar, using Lemma 2.33 (ii). □

Proof of Theorem 2.31. We begin observing that, if $A \in \mathcal{B}(\mathbf{X})$, the following conditions are equivalent:

$$\limsup_{n\to\infty}\|\mathbb{1}_A\|_{s_n,\mu_n} \leq \|\mathbb{1}_A\|_{\infty,I},$$

$$\limsup_{n\to\infty}\left(\int_{\mathbf{X}}\mathbb{1}_A\,d\mu_n\right)^{1/s_n} \leq \sup_{x\in\mathbf{X}} e^{-I(x)}\mathbb{1}_A(x),$$

$$\limsup_{n\to\infty}\Big(\mu_n(A)\Big)^{1/s_n} \leq \sup_{x\in A} e^{-I(x)},$$

$$\limsup_{n\to\infty}\log\left(\Big(\mu_n(A)\Big)^{1/s_n}\right) \leq \sup_{x\in A}(-I(x)),$$

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

Therefore,

$$\limsup_{n\to\infty}\|\mathbb{1}_A\|_{s_n,\mu_n} \leq \|\mathbb{1}_A\|_{\infty,I} \quad\Longleftrightarrow\quad \limsup_{n\to\infty}\frac{1}{s_n}\log\mu_n(A) \leq -\inf_{x\in A} I(x). \tag{2.11}$$

In the same way one can prove that for any $A \in \mathcal{B}(\mathbf{X})$

$$\liminf_{n\to\infty}\|\mathbb{1}_A\|_{s_n,\mu_n} \geq \|\mathbb{1}_A\|_{\infty,I} \quad\Longleftrightarrow\quad \liminf_{n\to\infty}\frac{1}{s_n}\log\mu_n(A) \geq -\inf_{x\in A} I(x). \tag{2.12}$$

Assume that μ_n satisfies the LDP with rate s_n and good rate function I. If $C \subseteq \mathbf{X}$ is closed, by Proposition 2.12 (i) there exists a sequence $f_k \in C_{b,+}(\mathbf{X})$ such that $f_k \searrow \mathbb{1}_C$. By the LDP, $f_k \in G_{up}$, and by Lemma 2.34 (iii) also $\mathbb{1}_C \in G_{up}$. By the equivalence stated in (2.11), condition (i) of the theorem is proved. The proof of condition (ii) is similar, using Proposition 2.12 (ii) and (2.12).

Conversely, assume that (i) and (ii) hold. For each closed $C \subseteq \mathbf{X}$, $\mathbb{1}_C$ is upper semicontinuous (by Lemma 2.10 (i)), bounded and nonnegative; moreover, condition (i) of this theorem and (2.11) imply that

$$\limsup_{n\to\infty}\|\mathbb{1}_C\|_{s_n,\mu_n} \leq \|\mathbb{1}_C\|_{\infty,I}.$$

Therefore $\mathbb{1}_C \in G_{up}$. By Lemma 2.10 (i) again and properties (i) and (ii) of Lemma 2.34, we obtain $S(\mathbf{X}) \cap U_{b,+}(\mathbf{X}) \subseteq G_{up}$. If $f \in U_{b,+}(\mathbf{X})$, then by Proposition 2.11 (i) there exists a sequence $f_k \in S(\mathbf{X}) \cap U_{b,+}(\mathbf{X}) \subseteq G_{up}$ such that $f_k \searrow f$, hence by Lemma 2.34 (iii) $f \in G_{up}$ again. This proves that $U_{b,+}(\mathbf{X}) = G_{up}$; similarly one can prove that $L_{b,+}(\mathbf{X}) = G_{low}$. If we now take

$$f \in C_{b,+}(\mathbf{X}) = U_{b,+}(\mathbf{X}) \cap L_{b,+}(\mathbf{X}) = G_{up} \cap G_{low}, \tag{2.13}$$

then $\lim_{n\to\infty}\|f\|_{s_n,\mu_n} = \|f\|_{\infty,I}$. We conclude that μ_n satisfies the LDP with rate s_n and rate function I. □

The latter proof suggests the following immediate corollary.

Corollary 2.35. $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the large deviations principle with rate s_n and good rate function I if and only if $G_{up} = U_{b,+}(\mathbf{X})$ and $G_{low} = L_{b,+}(\mathbf{X})$.

Proof. If $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the LDP with rate s_n and good rate function I, then conditions (i) and (ii) of Theorem 2.31 hold, hence the same argument of the proof above shows that $G_{up} = U_{b,+}(\mathbf{X})$ and $G_{low} = L_{b,+}(\mathbf{X})$. Conversely, if $G_{up} = U_{b,+}(\mathbf{X})$ and $G_{low} = L_{b,+}(\mathbf{X})$, then μ_n satisfies the LDP with rate s_n and good rate function I by (2.13) of the proof above again. □

It turns out that the good rate function of a LDP attains its minimum on $\mathbf{X}$:

Corollary 2.36. Assume that $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the large deviations principle with rate s_n and good rate function I. Then there exists $\overline{x} \in \mathbf{X}$ such that

$$I(\overline{x}) = \min_{x\in\mathbf{X}} I(x) = 0.$$

Proof. Since $\mathbf{X}$ is both closed and open, Theorem 2.31 implies that

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

Moreover, the set $\{I \le 1\}$ is nonempty (since $\inf_{x\in\mathbf{X}} I(x) = 0$) and compact (since I is a good rate function). I is lower semi-continuous, hence it attains its minimum on $\{I \le 1\}$ by Proposition 2.5 (ii). We conclude that

$$0 = \inf_{x\in\mathbf{X}} I(x) = \inf_{x\in\{I\le 1\}} I(x) = \min_{x\in\{I\le 1\}} I(x) = \min_{x\in\mathbf{X}} I(x). \qquad \square$$

Some natural questions one may consider concern the uniqueness of the good rate function. What can we say about two rate functions I and I' by comparing their supremum seminorms $\|\cdot\|_{\infty,I}$ and $\|\cdot\|_{\infty,I'}$? Is the good rate function unique for a sequence of probability measures satisfying the LDP with a fixed rate?

Definition 2.37. A set $D \subseteq C_{b,+}(\mathbf{X})$ is said to be *rate function determining* if for any two good rate functions I, I'

$$\|f\|_{\infty,I} = \|f\|_{\infty,I'} \quad \forall f \in D \qquad \Longrightarrow \qquad I = I'.$$

Lemma 2.38. Let $D \subseteq C_{b,+}(\mathbf{X})$. Assume that for all $x \in \mathbf{X}$ there exists a sequence $f_k \in D$ such that $f_k \searrow \mathbb{1}_{\{x\}}$. Then D is rate function determining.

Proof. Let I and I' be good rate functions such that $\|f\|_{\infty,I} = \|f\|_{\infty,I'}$ for all $f \in D$, and let $x \in \mathbf{X}$. By hypothesis, there exists a sequence $f_k \in D$ such that $f_k \searrow \mathbb{1}_{\{x\}}$. Since $f_k \in D \subseteq C_{b,+}(\mathbf{X}) \subseteq U_{b,+}(\mathbf{X})$, by Lemma 2.33 (iii)

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

i.e. $I(x) = I'(x)$. Since $x \in \mathbf{X}$ is arbitrary, it follows that $I = I'$. □

Proposition 2.39 (Uniqueness of the rate function). (i) $C_{b,+}(\mathbf{X})$ is rate function determining.

(ii) If a sequence $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the large deviations principle with rate s_n and good rate functions I and I', then $I = I'$.

Proof. (i) Let $x \in \mathbf{X}$. Since $\mathbf{X}$ is a Hausdorff space, $\{x\}$ is closed, hence by Proposition 2.12 (i) there exists a sequence $f_k \in C_{b,+}(\mathbf{X})$ such that $f_k \searrow \mathbb{1}_{\{x\}}$. By Lemma 2.38, $C_{b,+}(\mathbf{X})$ is rate function determining.

(ii) If μ_n satisfies the large deviations principle with rate s_n and good rate functions I and I', then

$$\|f\|_{\infty,I} = \lim_{n\to\infty} \|f\|_{s_n,\mu_n} = \|f\|_{\infty,I'} \qquad \forall f \in C_{b,+}(\mathbf{X}).$$

Since $C_{b,+}(\mathbf{X})$ is rate function determining, $I = I'$. □

We now give some examples of probability measures satisfying, or not, the LDP, using Theorem 2.31.

Example 2.40. Let $\mathbf{X} = \mathbb{R}$ and for any $n \in \mathbb{N}$ let μ_n be the uniform probability measure on $[-n, n]$. Assume that μ_n satisfies the LDP with rate s_n and good rate function I. If $A \in \mathcal{B}(\mathbb{R})$, then

$$\mu_n(A) = \int_A \frac{1}{2n} \mathbb{1}_{[-n,n]}(x)\,dx = \frac{|A \cap [-n,n]|}{2n},$$

where $|\cdot|$ denotes the Lebesgue measure on $\mathbb{R}$. If $A \subseteq \mathbb{R}$ is a bounded interval, then we can find $N \in \mathbb{N}$ such that $A \cap [-n,n] = A$ for all $n \geq N$, hence

$$\frac{1}{s_n} \log \mu_n(A) = \frac{1}{s_n} \log \frac{|A|}{2n} = \frac{\log|A| - \log 2 - \log(n)}{s_n} = -\frac{\log n}{s_n} + o(1).$$

If we set

$$\underline{a} \doteq \liminf_{n\to\infty} \frac{\log n}{s_n}, \qquad \overline{a} \doteq \limsup_{n\to\infty} \frac{\log n}{s_n},$$

we obtain

$$\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) = -\overline{a}, \qquad \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(A) = -\underline{a}. \tag{2.14}$$

Let $\overline{x} \in \mathbb{R}$ and let $C_k \doteq \overline{B_{1/k}(\overline{x})}$. Since C_k is closed and bounded for all k, $C_k \subseteq B_2(\overline{x})$ and $B_2(\overline{x})$ is open and bounded, by (2.14) and the definition of LDP

$$-\underline{a} = \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C_k) \le -\inf_{x\in C_k} I(x) \le -\inf_{x\in B_2(\overline{x})} I(x) \le \liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(B_2(\overline{x})) = -\overline{a}.$$

In particular, $-\underline{a} \le -\overline{a}$, i.e. $\frac{\log n}{s_n}$ has limit $a \doteq \underline{a} = \overline{a}$; moreover, $\inf_{x\in C_k} I(x) = a$ for all k. Since the lower semi-continuous function I attains its minimum on the compact C_k (see Proposition 2.5 (ii)), $\forall k$ there exists $x_k \in C_k$ such that $I(x_k) = a$. Since $x_k \in C_k$, $x_k \to \overline{x}$, hence by Proposition 2.7 (ii)

$$a = \liminf_{n\to\infty} I(x_k) \ge I(\overline{x}) \ge \inf_{x\in C_1} I(x) = a.$$

We conclude that $I(\overline{x}) = a$ for all $\overline{x} \in \mathbb{R}$. If $a \in [0, +\infty)$, it follows that I does not have compact level sets, contradicting Definition 2.26 (iii); if $a = +\infty$, it follows that $I \equiv +\infty$, contradicting Definition 2.26 (i). This proves that μ_n does not satisfy the LDP with rate s_n and good rate function I, for any s_n and any I. ♢

Example 2.41. Let $\mathbf{X} = \mathbb{R}$ and for any $n \in \mathbb{N}$ let μ_n be the uniform probability measure on $\left[-\frac{1}{n}, \frac{1}{n}\right]$. Assume that μ_n satisfies the LDP with rate s_n and good rate function I. If $A \in \mathcal{B}(\mathbb{R})$, then

$$\mu_n(A) = \int_A \frac{n}{2} \mathbb{1}_{[-\frac{1}{n},\frac{1}{n}]}(x)\, dx = \frac{n}{2} \cdot \left| A \cap \left[-\frac{1}{n}, \frac{1}{n}\right]\right|.$$

Let $C \subseteq \mathbb{R}$ be closed and such that $0 \in \mathring{C}$; then clearly $C \cap \left[-\frac{1}{n}, \frac{1}{n}\right] = \left[-\frac{1}{n}, \frac{1}{n}\right]$ and $\mu_n(C) = \frac{n}{2}\frac{2}{n} = 1$ for n large enough. Therefore:

- Since C is closed,

  $$-\inf_{x\in C} I(x) \ge \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C) = 0,$$

  hence $\inf_{x\in C} I(x) = 0$.

- Since C^c is open,

  $$-\inf_{x\in C^c} I(x) \le \liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(C^c) = \liminf_{n\to\infty} \frac{1}{s_n} \log(1 - \mu_n(C)) = -\infty,$$

  hence $\inf_{x\in C^c} I(x) = +\infty$.

If $x \ne 0$, $x \in [-\varepsilon, \varepsilon]^c$ for $\varepsilon > 0$ small enough, hence

$$I(x) \ge \inf_{x\in[-\varepsilon,\varepsilon]^c} I(x) = +\infty \qquad \Longrightarrow \qquad I(x) = +\infty.$$

On the other hand, since $\inf_{x\in[-1,1]} I(x) = 0$, $I(0) = 0$. This proves that, if μ_n satisfies the LDP with rate s_n and rate function I, then

$$I(x) = \begin{cases} 0 & x = 0, \\ +\infty & x \neq 0. \end{cases}$$

In fact, if I is defined as above, then μ_n satisfies the LDP with rate s_n and rate function I, for any $1 \leq s_n \to \infty$: first we have to prove that for all $C \subseteq \mathbb{R}$ closed

$$\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C) \leq \begin{cases} 0 & \text{if } 0 \in C, \\ -\infty & \text{if } 0 \in C^c. \end{cases}$$

The first inequality is always true, since $\mu_n(C) \leq 1$; for the second one, we note that if $0 \in C^c$ then, since C^c is open, $C \cap \left[-\frac{1}{n}, \frac{1}{n}\right] = \varnothing$ and $\mu_n(C) = 0$ for n large enough. Therefore $\frac{1}{s_n} \log \mu_n(C) \to -\infty$ and the claim is proved.

We also have to prove that for all $O \subseteq \mathbb{R}$ open

$$\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(O) \geq \begin{cases} 0 & \text{if } 0 \in O, \\ -\infty & \text{if } 0 \in O^c. \end{cases}$$

The second inequality is always true; for the first one, we note that if $0 \in O$ then, since O is open, $O \cap \left[-\frac{1}{n}, \frac{1}{n}\right] = \left[-\frac{1}{n}, \frac{1}{n}\right]$ and $\mu_n(O) = \frac{2}{n}\frac{n}{2} = 1$ for n large enough. Therefore $\frac{1}{s_n} \log \mu_n(O) \to 0$ and the claim is proved. ◊

Example 2.42. Let $\mathbf{X} = \mathbb{R}$ and for any $n \in \mathbb{N}$ let μ_n be the absolutely continuous probability measure with density function

$$f_n(t) = \frac{n}{2} e^{-n|t|}.$$

For all $A \in \mathcal{B}(\mathbb{R})$ we have

$$\mu_n(A) = \frac{n}{2} \int_A e^{-n|t|}\, dt \leq \frac{n}{2} \sup_{x\in A} e^{-n|x|} = \frac{n}{2} \exp\Big(-n \inf_{x\in A} |x|\Big),$$

hence

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(A) &\leq \limsup_{n\to\infty} \frac{1}{n} \log\Big(\frac{n}{2} \exp\Big(-n \inf_{x\in A}|x|\Big)\Big) \\ &= \limsup_{n\to\infty} \Big(\frac{\log n}{n} - \frac{\log 2}{n} - \inf_{x\in A}|x|\Big) = -\inf_{x\in A}|x|. \end{aligned}$$

This suggests that μ_n might satisfy the LDP with rate n and good rate function $I(x) \doteq |x|$ (I is such that $0 \leq I < +\infty$, it is continuous and it has compact level sets $\{I \leq a\} = [-a, a]$). Indeed, the latter computation in particular shows that for all $C \subseteq \mathbb{R}$ closed

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(C) \leq -\inf_{x\in C}|x|.$$

To prove also condition (ii) of Theorem 2.31, we will show that for each open set $O \subseteq \mathbb{R}$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(O) \geq -|x| \qquad \forall x \in O.$$

If $x \in O$, there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq O$, hence it suffices to prove that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\varepsilon(x)) \geq -|x|.$$

If $x > 0$, we can choose ε such that $B_\varepsilon(x) \subseteq O \cap \mathbb{R}^+$, hence

$$\mu_n(B_\varepsilon(x)) = \frac{n}{2}\int_{x-\varepsilon}^{x+\varepsilon} e^{-nt}\,dt = \frac{n}{2}\Big[-\frac{e^{-nt}}{n}\Big]_{x-\varepsilon}^{x+\varepsilon} = \frac{e^{-n(x-\varepsilon)} - e^{-n(x+\varepsilon)}}{2} = \frac{e^{-n(x-\varepsilon)}}{2}(1-e^{-2n\varepsilon}).$$

Therefore,

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\varepsilon(x)) &= \liminf_{n\to\infty} \frac{1}{n} \log\Big(\frac{e^{-n(x-\varepsilon)}}{2}(1-e^{-2n\varepsilon})\Big) \\ &= \liminf_{n\to\infty}\Big(-(x-\varepsilon) - \frac{\log 2}{n} + \frac{\log(1-e^{-2n\varepsilon})}{n}\Big) \\ &= -x+\varepsilon \geq -x.\end{aligned} \tag{2.15}$$

This proves the claim for $x > 0$. If $x < 0$, we can choose ε such that $B_\varepsilon(x) \subseteq O \cap \mathbb{R}^-$ and, by the symmetry of the distribution μ_n with respect to 0, $\mu_n(B_\varepsilon(x)) = \mu_n(B_\varepsilon(-x))$. Therefore, by applying (2.15) to $-x > 0$ it turns out that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\varepsilon(x)) = \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\varepsilon(-x)) \geq -(-x) = -|x|.$$

Finally, if $x = 0$,

$$\mu_n(B_\varepsilon(0)) = \frac{n}{2}\int_{-\varepsilon}^{\varepsilon} e^{-n|t|}\,dt = n\int_0^{\varepsilon} e^{-nt}\,dt = n\Big[-\frac{e^{-nt}}{n}\Big]_0^{\varepsilon} = 1 - e^{-n\varepsilon} \xrightarrow{n\to\infty} 1,$$

hence

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\varepsilon(0)) = 0 = -|0|.$$

We conclude that μ_n actually satisfies the LDP with rate n and good rate function $I(x) = |x|$. ◊

We now discuss another characterization of the LDP.

Theorem 2.43 (Varadhan-Bryc). A sequence $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the large deviations principle with rate s_n and good rate function I if and only if

$$\lim_{n\to\infty} \frac{1}{s_n} \log \int_{\mathbf{X}} e^{s_n F}\,d\mu_n = \sup_{x\in\mathbf{X}}[F(x) - I(x)] \tag{2.16}$$

for any $F\colon \mathbf{X} \to \overline{\mathbb{R}}$ continuous and bounded from above.

Proof. We first note that any $F\colon \mathbf{X} \to \overline{\mathbb{R}}$ is continuous and bounded from above if and only if $f \doteq e^F \in C_{b,+}(\mathbf{X})$. Therefore, the following sequence of equivalent equalities proves the theorem.

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

$$\lim_{n\to\infty} \log\left(\int_{\mathbf{X}} \left(e^F\right)^{s_n} d\mu_n \right)^{1/s_n} = \sup_{x\in\mathbf{X}} \log\left[e^{-I(x)} e^{F(x)}\right] \qquad \text{(applying log to both sides)},$$

$$\lim_{n\to\infty} \frac{1}{s_n} \log \int_{\mathbf{X}} e^{s_n F}\, d\mu_n = \sup_{x\in\mathbf{X}} [F(x) - I(x)]. \qquad \square$$

Since our definition of LDP is 2.29, the latter theorem comes 'for free'. However, if we take the traditional definition of LDP (the two conditions of Theorem 2.31), this theorem essentially corresponds to our Theorem 2.31. Historically, the 'if' part (the LDP is a sufficient condition for (2.16)) was proved by Varadhan in 1966 (see [Var]); the 'only if' part (the LDP is a necessary condition for (2.16)) was only proved by Bryc in 1990, twenty-four years later (see [Bry]).

2.4 Tightness

We now recall the concept of tightness of a set of probability measures on a Polish space $\mathbf{X}$ and Prohorov's Theorem, which highlights its relationship with weak convergence.

Definition 2.44. Let $\mathbf{X}$ be a Polish space. A subset $\mathfrak{N} \subseteq \mathfrak{M}(\mathbf{X})$ is said to be *tight* if

$$\forall \varepsilon > 0 \quad \exists K \subseteq \mathbf{X} \text{ compact:} \quad \mu(K^c) \le \varepsilon \quad \forall \mu \in \mathfrak{N}.$$

Intuitively, $\mathfrak{N} \subseteq \mathfrak{M}(\mathbf{X})$ is tight if $\mathbf{X}$ can be uniformly approximated with a compact subset, in the sense of the probability measures $\mu \in \mathfrak{N}$.

Example 2.45. If $\mathbf{X}$ is compact, any subset $\mathfrak{N} \subseteq \mathfrak{M}(\mathbf{X})$ is obviously tight: it suffices to consider $K = \mathbf{X}$ in the condition of tightness. ◊

Example 2.46. Let $\mathfrak{N} \subseteq \mathfrak{M}(\mathbf{X})$ be *finite*. Assume that $\mathbf{X}$ is a *sigma-compact*† Polish space (for example, $\mathbf{X} = \mathbb{R}^d$), so that there exists a countable collection of compact subsets $\{K_i\}_{i\in\mathbb{N}}$ such that $\mathbf{X} = \bigcup_{i\in\mathbb{N}} K_i$. If we set $\widetilde{K}_N \doteq \bigcup_{i=1}^N K_i$ for all $N \in \mathbb{N}$, we can also write $\mathbf{X} = \bigcup_{N\in\mathbb{N}} \widetilde{K}_N$. Since $\{\widetilde{K}_N\}_{N\in\mathbb{N}}$ is an increasing sequence of subsets of $\mathbf{X}$, by the continuity of measures

$$\lim_{N\to\infty} \mu(\widetilde{K}_N) = \mu(\mathbf{X}) = 1 \qquad \forall \mu \in \mathfrak{N}.$$

†We recall that a topological space is said to be sigma-compact if it is the union of countably many compact subsets.

Since $\mathfrak{N}$ is finite, for all $\varepsilon > 0$ there exists N such that $\mu(\widetilde{K}_N) \geq 1 - \varepsilon$ for all $\mu \in \mathfrak{N}$, i.e.

$$\mu(\widetilde{K}_N^c) \leq \varepsilon \qquad \forall \mu \in \mathfrak{N}.$$

Since by definition $\widetilde{K}_N$ is a finite union of compact sets, it is compact. Thus we conclude that every finite collection of probability measures on a sigma-compact Polish space is tight. As we will see in Example 2.49, the hypothesis of sigma-compactness is actually unnecessary. ♢

Remark 2.47. If $\mathbf{X} = \mathbb{R}^d$, in the definition of tightness we may replace compact sets with open (or closed) balls. More precisely, we claim that $\mathfrak{N} \subseteq \mathfrak{M}(\mathbb{R}^d)$ is tight if and only if

$$\forall \varepsilon > 0 \quad \exists R > 0 : \quad \mu(B_R(0)^c) \leq \varepsilon \quad \forall \mu \in \mathfrak{N}. \tag{2.17}$$

Indeed, let $\mathfrak{N}$ be tight and $\varepsilon > 0$: then, there exists $K \subseteq \mathbf{X}$ compact such that $\mu(K^c) \leq \varepsilon$ for all $\mu \in \mathfrak{N}$. Since K is compact, it is bounded, hence there exists $R > 0$ such that $K \subseteq B_R(0)$: this proves that $\mu(B_R(0)^c) \leq \mu(K^c) \leq \varepsilon$ for all $\mu \in \mathfrak{N}$. We note that this first implication holds for any Polish space $\mathbf{X}$, since compact sets are always bounded in any metric space.

Conversely, assume that (2.17) holds and let $\varepsilon > 0$. Then, there exists $R > 0$ such that $\mu(B_R(0)^c) \leq \varepsilon$ for all $\mu \in \mathfrak{N}$. By the Heine-Borel Theorem, $\overline{B_R(0)} \subseteq \mathbb{R}^d$ is compact, hence $\mu\left(\overline{B_R(0)}^c\right) \leq \mu(B_R(0)^c) \leq \varepsilon$ for all $\mu \in \mathfrak{N}$, which proves the tightness of $\mathfrak{N}$. We note that this second implication uses the assumption that $\mathbf{X} = \mathbb{R}^d$, because closed balls are not in general compact in any Polish space (for instance, in every infinite-dimensional Banach space closed balls are not compact). ♢

Theorem 2.48 (Prohorov's Theorem)**.** Let $\mathbf{X}$ be a Polish space. $\mathfrak{N} \subseteq \mathfrak{M}(\mathbf{X})$ is pre-compact (i.e. its closure is compact) in the weak topology if and only if $\mathfrak{N}$ is tight.

Proof. See for example [Bil, Th.5.1 and Th.5.2]. □

Example 2.49. Assume that $\mathfrak{N} \subseteq \mathfrak{M}(\mathbf{X})$ is finite. If $\mathbf{X}$ is sigma-compact, we proved in Example 2.46 that $\mathfrak{N}$ is tight. Let now $\mathbf{X}$ be a generic Polish space. Since finite sets are compact in any topology, $\mathfrak{N}$ is compact (in particular, pre-compact) in the weak topology, hence by Prohorov's Theorem $\mathfrak{N}$ is tight. Thus we conclude that *every finite collection of probability measures on a Polish space* $\mathbf{X}$ *is tight* (even if $\mathbf{X}$ is not sigma-compact)[†]. ♢

Remark 2.50. Since $\mathfrak{M}(\mathbf{X})$ is a metric space (see Example 2.19), a subset $\mathfrak{N}$ is pre-compact if and only if any sequence $\{\mu_n\}_{n \geq 1} \subseteq \mathfrak{N}$ has a subsequence that converges weakly in $\mathfrak{M}(\mathbf{X})$. Prohorov's Theorem may be therefore formulated as follows:

[†]This can actually be proved in a more straightforward way, without the need of invoking Prohorov's Theorem: see [Bil, Th.1.3].

$\mathfrak{N} \subseteq \mathfrak{M}(\mathbf{X})$ is tight if and only if any sequence $\{\mu_n\}_{n\geq 1} \subseteq \mathfrak{N}$ has a subsequence that converges weakly in $\mathfrak{M}(\mathbf{X})$. In particular, if $\mathfrak{N} \doteq \{\mu_n\}_{n\geq 1}$ is a countable subset of $\mathfrak{M}(\mathbf{X})$, we have:

- If $\{\mu_n\}_{n\geq 1}$ converges weakly, then it is tight.
- If $\{\mu_n\}_{n\geq 1}$ is tight, then there exists a subsequence $\{\mu_{n_k}\}_{k\geq 1}$ and $\mu \in \mathfrak{M}(\mathbf{X})$ such that $\mu_{n_k} \xrightarrow{k\to\infty} \mu$. ♢

Example 2.51. Let $\mathbf{X} = \mathbb{R}^d$ and $\mathfrak{N} \doteq \{\mu_n\}_{n\geq 1}$, where μ_n is the continuous uniform distribution on $B_{R_n}(0)$ and $\{R_n\}_{n\geq 1}$ is a positive sequence. We want to find conditions on $\{R_n\}_{n\geq 1}$ for the tightness of $\{\mu_n\}_{n\geq 1}$, using Remark 2.47. If we denote the Lebesgue measure on $\mathbb{R}^d$ by $|\cdot|$ as usual, for all $R > 0$

$$\mu_n(B_R(0)) = \int_{\mathbb{R}^d} \frac{1}{|B_{R_n}(0)|} \mathbb{1}_{B_{R\wedge R_n}(0)}(x)\,dx = \frac{|B_{R\wedge R_n}(0)|}{|B_{R_n}(0)|} = \frac{(R\wedge R^n)^d}{(R_n)^d} = \left(\frac{R}{R_n}\wedge 1\right)^d.$$

Thus we may distinguish two cases:

- If $R \doteq \sup_{n\geq 1} R_n < +\infty$, then $\mu_n(B_R(0)) = 1$ for all $n \geq 1$ and $\{\mu_n\}_{n\geq 1}$ is tight, by condition (2.17).
- If $\sup_{n\geq 1} R_n = +\infty$, then there exists a subsequence R_{n_k} such that $R_{n_k} \xrightarrow{k\to\infty} +\infty$, hence
$$\forall R > 0 \quad \mu_{n_k}(B_R(0)) = \left(\frac{R}{R_{n_k}}\wedge 1\right)^d \xrightarrow{k\to\infty} 0,$$
which contradicts condition (2.17) and proves that $\{\mu_n\}_{n\geq 1}$ is not tight.

Therefore, $\{\mu_n\}_{n\geq 1}$ is tight if and only if $\{R_n\}_{n\geq 1}$ is bounded. For instance, if

$$R_n = 2 + (-1)^n,$$

then $\{\mu_n\}_{n\geq 1}$ is tight. According to Remark 2.50, it has subsequences that converge weakly, for example:

$$\mu_{2n} = U(B_3(0)) \rightharpoonup U(B_3(0));$$
$$\mu_{2n+1} = U(B_1(0)) \rightharpoonup U(B_1(0)).$$

However, $\{\mu_n\}_{n\geq 1}$ does not converge weakly, because the weak limits of such subsequences are not equal. ♢

Remark 2.52. In the next section, the following observation will be useful to give an appropriate "large deviations analogue" of tightness: $\{\mu_n\}_{n\geq 1}$ is tight if and only if

$$\forall \varepsilon > 0 \quad \exists K \subseteq \mathbf{X} \text{ compact:} \quad \limsup_{n\to\infty} \|\mathbb{1}_{K^c}\|_{1,\mu_n} \leq \varepsilon. \tag{2.18}$$

Indeed,

$$\|\mathbb{1}_{K^c}\|_{1,\mu_n} = \int_{\mathbf{X}} \mathbb{1}_{K^c}\, d\mu_n = \mu_n(K^c),$$

hence, if $\{\mu_n\}_{n\geq 1}$ is tight, then (2.18) clearly holds. Conversely, assume that (2.18) holds and let $\varepsilon > 0$. By hypothesis, there exists $K_1 \subseteq \mathbf{X}$ compact such that

$$\limsup_{n\to\infty} \|\mathbb{1}_{K_1^c}\|_{1,\mu_n} = \lim_{N\to\infty} \sup_{n\geq N} \mu_n(K_1^c) \leq \frac{\varepsilon}{2},$$

hence there exists $N \geq 1$ such that $\sup_{n\geq N} \mu_n(K_1^c) \leq \varepsilon$. Moreover, since by Example 2.49 the finite collection of measures $\{\mu_1, \dots, \mu_{N-1}\}$ is tight, there exists $K_2 \subseteq \mathbf{X}$ compact such that $\mu_n(K_2^c) \leq \varepsilon$ for all $n = 1, \dots, N-1$. If we consider the compact set $K \doteq K_1 \cup K_2$, it follows that $\mu_n(K^c) \leq \varepsilon$ for all $n \geq 1$, which proves that $\{\mu_n\}_{n\geq 1}$ is tight. ♢

2.5 Exponential tightness

In this section we formulate and discuss a concept in the large deviations context that plays the same role of tightness in the weak convergence context: the so-called *exponential tightness*. The characterization of tightness (2.18) inspires the following definition.

Definition 2.53. Let $\mathbf{X}$ be a Polish space and let s_n be a sequence such that $s_n \geq 1$ and $s_n \to +\infty$. We say that a sequence $\mu_n \in \mathfrak{M}(\mathbf{X})$ is *exponentially tight* with *rate* s_n if

$$\forall \varepsilon > 0 \quad \exists K \subseteq \mathbf{X} \text{ compact:} \quad \limsup_{n\to\infty} \|\mathbb{1}_{K^c}\|_{s_n,\mu_n} \leq \varepsilon. \tag{2.19}$$

Example 2.54. If $\mathbf{X}$ is compact, any sequence $\mu_n \in \mathfrak{M}(\mathbf{X})$ is obviously exponentially tight with any rate s_n: it suffices to consider $K = \mathbf{X}$ in condition (2.19). ♢

Remark 2.55. Since

$$\|\mathbb{1}_{K^c}\|_{s_n,\mu_n} = \left(\int_{\mathbf{X}} \mathbb{1}_{K^c}\, d\mu_n \right)^{\frac{1}{s_n}} = \exp\left(\frac{1}{s_n} \log \mu_n(K^c) \right),$$

it follows that $\mu_n \in \mathfrak{M}(\mathbf{X})$ is exponentially tight with rate s_n if and only if

$$\forall M \in \mathbb{R} \quad \exists K \subseteq \mathbf{X} \text{ compact:} \quad \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(K^c) \leq -M. \tag{2.20}$$

The change between the arbitrary $\varepsilon > 0$ in (2.19) and the arbitrary $M \in \mathbb{R}$ in (2.20) is given by $\varepsilon = \exp(-M)$. ♢

Remark 2.56. The same argument of Remark 2.47 shows that $\mu_n \in \mathfrak{M}(\mathbb{R}^d)$ is exponentially tight with rate s_n if and only if

$$\forall M \in \mathbb{R} \quad \exists R > 0: \quad \limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(B_R(0)^c) \leq -M. \tag{2.21}$$

♢

Example 2.57. Let $\mathbf{X} = \mathbb{R}^d$ and let μ_n be the continuous uniform distribution on $B_{R_n}(0)$, where $\{R_n\}_{n\geq 1}$ is a positive sequence, just like in Example 2.51. We want to find conditions on $\{R_n\}_{n\geq 1}$ for the *exponential* tightness of μ_n, using Remark 2.56. Recalling what we showed in Example 2.51, we have:

- If $R \doteq \sup_{n\geq 1} R_n < +\infty$, then $\mu_n(B_R(0)) = 1$ for all $n \geq 1$; in this case, for any sequence $s_n \to \infty$,
$$\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(B_R(0)^c) = -\infty,$$
hence μ_n is exponentially tight with any rate s_n by condition (2.21).

- If $\sup_{n\geq 1} R_n = +\infty$, then there exists a subsequence n_k such that for all $R > 0$ $\mu_{n_k}(B_R(0)) \xrightarrow{k\to\infty} 0$, hence for any $s_n \to \infty$
$$\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(B_R(0)^c) \geq \limsup_{k\to\infty} \frac{1}{s_{n_k}} \log \mu_{n_k}(B_R(0)^c) = 0,$$
which contradicts condition (2.21) and proves that μ_n is not exponentially tight with rate s_n, for any $s_n \to \infty$.

Therefore, for any fixed $s_n \to \infty$, μ_n is exponentially tight with rate s_n if and only if $\{R_n\}_{n\geq 1}$ is bounded. For instance, if
$$R_n = 2 + (-1)^n,$$
then μ_n is exponentially tight. We note that μ_n has subsequences that satisfy a LDP with rate s_n, since by Example 2.30

- $\mu_{2n} = U(B_3(0))$ satisfies the LDP with rate s_n and good rate function
$$I_{\text{even}}(x) = \begin{cases} 0 & x \in \overline{B_3(0)}, \\ +\infty & x \notin \overline{B_3(0)}. \end{cases}$$

- $\mu_{2n+1} = U(B_1(0))$ satisfies the LDP with rate s_n and good rate function
$$I_{\text{odd}}(x) = \begin{cases} 0 & x \in \overline{B_1(0)}, \\ +\infty & x \notin \overline{B_1(0)}. \end{cases}$$

However, $\{\mu_n\}_{n\geq 1}$ does not satisfy any LDP, because the good rate functions I_{even} and I_{odd} are not equal. ◊

The latter example indicates that an analogue of Prohorov's Theorem (or better of Remark 2.50) may also hold in large deviations theory. In fact, the aim of this section is to prove the following important result, which connects exponential tightness and LDP's, and dates back to a paper by O'Brian and Vervaat in 1991 (see [OBVe]).

Theorem 2.58. Let $\mathbf{X}$ be a Polish space and let μ_n be a sequence in $\mathfrak{M}(\mathbf{X})$.

(i) If μ_n satisfies the LDP with rate s_n and good rate function I, then μ_n is exponentially tight with rate s_n.

(ii) If μ_n is exponentially tight with rate s_n, then there exist a good rate function I and a subsequence μ_{n_k} such that μ_{n_k} satisfies the LDP with rate s_{n_k} and good rate function I.

We start by proving the easiest statement, the first one.

Theorem 2.59. Let $\mathbf{X}$ be a Polish space. If $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the LDP with rate s_n and good rate function I, then μ_n is exponentially tight with rate s_n.

Proof. Since $\mathbf{X}$ is Polish, we can choose:

- A metric d such that $(\mathbf{X}, d)$ is a complete metric space ($B_\delta(x)$ will denote the open ball of center x and radius δ with respect to d).
- A dense sequence $\{x_i\}_{i\geq 1}$ in $\mathbf{X}$.

By the density, it turns out that for any $\delta > 0$

$$\mathbf{X} = \bigcup_{i=1}^{\infty} B_\delta(x_i) = \bigcup_{k=1}^{\infty} O_{\delta,k}$$

where $O_{\delta,k} \doteq \bigcup_{i=1}^{k} B_\delta(x_i)$, $k \geq 1$. Since the sets $\{O_{\delta,k}\}_{k\geq 1}$ increase to $\mathbf{X}$, $\mathbb{1}_{O^c_{\delta,k}} \searrow 0$ as $k \to \infty$; since $O_{\delta,k}$ are open, $\mathbb{1}_{O^c_{\delta,k}} \in U_{b,+}(\mathbf{X})$ by Lemma 2.10. Thus we can apply Lemma 2.33 (iii): for any $\delta > 0$ $\|\mathbb{1}_{O^c_{\delta,k}}\|_{\infty,I} \searrow 0$ as $k \to \infty$. Let $\varepsilon > 0$; then there exists a subsequence $\{k_j\}_{j\geq 1}$ such that

$$\|\mathbb{1}_{O^c_{1/j,k_j}}\|_{\infty,I} \leq \frac{\varepsilon}{2^j} \qquad \forall j \geq 1. \tag{2.22}$$

If K denotes the closure of $\bigcap_{j=1}^{\infty} O_{1/j,k_j}$, we claim that K is compact. Since $(\mathbf{X}, d)$ is a *complete* metric space, a subset is compact if and only if it is closed and *totally bounded*†; thus, it suffices to prove that K is totally bounded. Since

$$\bigcap_{j=1}^{\infty} O_{1/j,k_j} = \bigcap_{j=1}^{\infty} \bigcup_{i=1}^{k_j} B_{1/j}(x_i) \subseteq \bigcup_{i=1}^{k_j} B_{1/j}(x_i) \qquad \forall j \geq 1,$$

†We recall that a subset of a metric space is said to be totally bounded if $\forall \delta > 0$ it can be covered with a finite number of open balls of radius δ.

it follows that $K \subseteq \bigcup_{i=1}^{k_j} B_{2/j}(x_i)$. For any $\delta > 0$, if we choose $j \geq 1$ such that $2/j < \delta$, it turns out that $K \subseteq \bigcup_{i=1}^{k_j} B_\delta(x_i)$: this proves that K is totally bounded, hence compact. To prove the exponential tightness of μ_n, it remains to show that

$$\limsup_{n\to\infty} \|\mathbb{1}_{K^c}\|_{s_n,\mu_n} \leq \varepsilon.$$

Since $K^c \subseteq \left(\bigcap_{j=1}^\infty O_{1/j,k_j}\right)^c = \bigcup_{j=1}^\infty O^c_{1/j,k_j}$, for all $n \geq 1$

$$\begin{aligned}\|\mathbb{1}_{K^c}\|_{s_n,\mu_n} &\leq \|\mathbb{1}_{\bigcup_{j=1}^\infty O^c_{1/j,k_j}}\|_{s_n,\mu_n} \leq \left\|\sum_{j=1}^\infty \mathbb{1}_{O^c_{1/j,k_j}}\right\|_{s_n,\mu_n} = \left\|\lim_{N\to\infty}\sum_{j=1}^N \mathbb{1}_{O^c_{1/j,k_j}}\right\|_{s_n,\mu_n} \\ &= \lim_{N\to\infty}\left\|\sum_{j=1}^N \mathbb{1}_{O^c_{1/j,k_j}}\right\|_{s_n,\mu_n} \leq \lim_{N\to\infty}\sum_{j=1}^N \|\mathbb{1}_{O^c_{1/j,k_j}}\|_{s_n,\mu_n} = \sum_{j=1}^\infty \|\mathbb{1}_{O^c_{1/j,k_j}}\|_{s_n,\mu_n},\end{aligned}$$

where we have applied the monotone convergence theorem to the increasing sequence of non-negative functions $\{\sum_{j=1}^N \mathbb{1}_{O^c_{1/j,k_j}}\}_{N\geq 1}$, and the subadditivity of the norm $\|\cdot\|_{s_n,\mu_n}$. It follows that

$$\limsup_{n\to\infty}\|\mathbb{1}_{K^c}\|_{s_n,\mu_n} \leq \sum_{j=1}^\infty \limsup_{n\to\infty}\|\mathbb{1}_{O^c_{1/j,k_j}}\|_{s_n,\mu_n} \leq \sum_{j=1}^\infty \|\mathbb{1}_{O^c_{1/j,k_j}}\|_{\infty,I} \leq \sum_{j=1}^\infty \frac{\varepsilon}{2^j} = \varepsilon,$$

where the second inequality holds because $\mathbb{1}_{O^c_{1/j,k_j}} \in U_{b,+}(\mathbf{X}) = G_{up}$ by Corollary 2.35, and the third one holds by (2.22). □

Lemma 2.60. Let $\mathbf{X}$ be a Polish space. Let $\mu_n \in \mathfrak{M}(\mathbf{X})$, $1 \leq s_n \to \infty$ and let I be a good rate function. Assume that

(i) μ_n is exponentially tight with rate s_n.

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

Then

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

Proof. Let $M \in \mathbb{R}$. Since μ_n is exponentially tight, by (2.20) there exists $K_M \subseteq \mathbf{X}$ compact such that

$$\limsup_{n\to\infty}\frac{1}{s_n}\log\mu_n(K_M^c) \leq -M.$$

It follows that, if $C \subseteq \mathbf{X}$ is closed,

$$\limsup_{n\to\infty}\frac{1}{s_n}\log\mu_n(C\cap K_M^c) \leq -M.$$

Moreover, $C \cap K_M$ is compact, hence (ii) holds for $C \cap K_M$. Therefore, applying (1.19) with $\alpha_n = \mu_n(C \cap K_M)$ and $\beta_n = \mu_n(C \cap K_M^c)$, we obtain

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C) &= \limsup_{n\to\infty} \frac{1}{s_n} \log(\mu_n(C \cap K_M) + \mu_n(C \cap K_M^c)) \\
&= \limsup_{n\to\infty} \left(\left(\frac{1}{s_n} \log \mu_n(C \cap K_M) \right) \vee \left(\frac{1}{s_n} \log \mu_n(C \cap K_M^c) \right) + o(1) \right) \\
&= \left(\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C \cap K_M) \right) \vee \left(\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(C \cap K_M^c) \right) \\
&\le \left(-\inf_{x \in C \cap K_M} I(x) \right) \vee (-M) = -\left(\left(\inf_{x \in C \cap K_M} I(x) \right) \wedge M \right) \\
&\le -\left(\left(\inf_{x\in C} I(x) \right) \wedge M \right) \xrightarrow{M\to\infty} -\inf_{x \in C} I(x)
\end{aligned}$$

Since $C \subseteq \mathbf{X}$ closed is arbitrary, this proves the claim. □

The latter lemma gives us a criterion to show that a certain sequence of probability measures μ_n satisfies a LDP: if we know that μ_n is exponentially tight, then it suffices to verify the large deviations upper bound for compact sets (not for each closed set). More precisely, the following result holds.

Corollary 2.61. Let $\mathbf{X}$ be a Polish space. Let $\mu_n \in \mathfrak{M}(\mathbf{X})$, $1 \le s_n \to \infty$ and let I be a good rate function. Then μ_n satisfies the LDP with rate s_n and good rate function I if and only if the following three conditions hold:

(i) μ_n is exponentially tight with rate s_n.

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

(iii) $\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(O) \ge -\inf_{x\in O} I(x) \qquad \forall O \subseteq \mathbf{X}$ open.

Proof. Assume that μ_n satisfies the LDP with rate s_n and good rate function I. Then, (i) follows from Theorem 2.59; (ii) and (iii) follow from Theorem 2.31 (recall that every compact is closed in a Hausdorff space).

Assume now that the three conditions hold. By Lemma 2.60, (i) and (ii) guarantee the large deviations upper bound; the large deviations lower bound also holds by (iii). Therefore, μ_n satisfies the large deviations principle with rate s_n and good rate function I by Theorem 2.31. □

Our aim is now to prove the second statement of Theorem 2.58, that is more difficult than the first one by far: the main problem is that we have to “build” a good rate function, that is not given by hypothesis. First of all, we will prove the theorem for compact metrizable spaces (which are Polish, as we proved in Example 2.18); afterwards, we will extend it to generic Polish spaces.

The next lemma shows that, if $\mathbf{X}$ is a compact metrizable space, a function $C(\mathbf{X}) \to \mathbb{R}^+$ satisfying certain properties needs to be a seminorm $\|\cdot\|_{\infty,I}$, for some good rate function I.

Lemma 2.62. Let $\mathbf{X}$ be a compact metrizable space and let $\Lambda\colon C(\mathbf{X}) \to \mathbb{R}^+$ be a function such that

(i) Λ is a seminorm.

(ii) $\Lambda(f) = \Lambda(|f|) \quad \forall f \in C(\mathbf{X})$.

(iii) $\Lambda(f) \le \Lambda(g) \quad \forall f, g \in C(\mathbf{X}),\ 0 \le f \le g$.

(iv) $\Lambda(f \vee g) \le \Lambda(f) \vee \Lambda(g) \quad \forall f, g \in C(\mathbf{X}),\ f, g \ge 0$.

(v) $\Lambda(\mathbf{1}) = 1$, where $\mathbf{1}$ is the function that is identically $= 1$.

Then there exists a good rate function I such that $\Lambda(f) = \|f\|_{\infty,I}$ for all $f \in C(\mathbf{X})$.

Proof. Let d be a metric on $\mathbf{X}$ that generates its topology. Our aim is to prove that there exists a good rate function I such that $\Lambda(f) = \sup_{x\in\mathbf{X}} e^{-I(x)}|f(x)|$ for all $f \in C(\mathbf{X})$. If this happens, $e^{-I(x)} \le \Lambda(f)$ for all $f \in C(\mathbf{X})$ such that $f(x) = 1$. Therefore, it is reasonable to define $I\colon \mathbf{X} \to (-\infty, +\infty]$ by

$$e^{-I(x)} \doteq \inf\{\Lambda(f): \quad f \in C(\mathbf{X}), \quad f \ge 0, \quad f(x) = 1\} \qquad \forall x \in \mathbf{X}.$$

We begin by proving the following claim:

$$f_n \in C(\mathbf{X}), \quad f_n \searrow \mathbb{1}_{\{x\}} \quad \Longrightarrow \quad \Lambda(f_n) \searrow e^{-I(x)}. \tag{2.23}$$

Since $f_n \ge f_{n+1} \ge \mathbb{1}_{\{x\}} \ge 0$ for all n, $\Lambda(f_n) \ge \Lambda(f_{n+1})$ for all n by (iii), i.e. $\Lambda(f_n)$ is a decreasing sequence. Moreover, $f_n(x) \ge 1$ implies that $f_n \ge \frac{f_n}{f_n(x)}$, hence by (iii) again

$$\Lambda(f_n) \ge \Lambda\left(\frac{f_n}{f_n(x)}\right) \ge \inf\{\Lambda(f): \quad f \in C(\mathbf{X}), \quad f \ge 0, \quad f(x) = 1\} = e^{-I(x)}.$$

This proves that

$$\liminf_{n\to\infty} \Lambda(f_n) \ge e^{-I(x)}.$$

To prove the upper bound, let $\varepsilon > 0$ and consider $f \in C(\mathbf{X})$ such that $f \ge 0$, $f(x) = 1$ and $\Lambda(f) \le e^{-I(x)} + \varepsilon$. We define

$$K_n \doteq \{t \in \mathbf{X}: \quad f_n(t) \ge f(t) + \varepsilon\}.$$

K_n is a closed subset of the compact space $\mathbf{X}$, since the functions f_n and f are both continuous, hence it is compact itself. Since $f_n \ge f_{n+1}$ for all n, K_n is decreasing.

Moreover,

$$\bigcap_{n\geq 1} K_n = \{t\in\mathbf{X}:\quad f_n(t)\geq f(t)+\varepsilon\quad \forall n\geq 1\} = \{t\in\mathbf{X}:\quad \lim_{n\to\infty} f_n(t)\geq f(t)+\varepsilon\}$$
$$= \{t\in\mathbf{X}:\quad f(t)-\mathbb{1}_{\{x\}}(t)\leq -\varepsilon\} = \varnothing,$$

since

$$f(t)-\mathbb{1}_{\{x\}}(t) = \begin{cases} f(x)-1=0 & t=x,\\ f(t)\geq 0 & t\neq x.\end{cases}$$

Thus, K_n is a decreasing sequence of compact sets whose intersection is empty. This means that $K_n=\varnothing$, i.e. $f_n < f+\varepsilon$, for n large enough. It follows that

$$\Lambda(f_n)\leq \Lambda(f+\varepsilon)\leq \Lambda(f)+\varepsilon\Lambda(\mathbf{1})\leq \Lambda(f)+\varepsilon\leq e^{-I(x)}+2\varepsilon$$

for n large enough, thanks to (iii), (i) and (v). Passing to the limit superior as $n\to\infty$, we obtain

$$\limsup_{n\to\infty}\Lambda(f_n)\leq e^{-I(x)}+2\varepsilon.$$

Since $\varepsilon>0$ is arbitrary, the upper bound follows. This concludes the proof of (2.23).

We now prove that

$$\Lambda(f) = \sup_{x\in\mathbf{X}} e^{-I(x)}|f(x)| \qquad \forall f\in C(\mathbf{X}). \tag{2.24}$$

By (ii), it suffices to prove the claim for $f\in C(\mathbf{X})$, $f\geq 0$. Let $x\in\mathbf{X}$: if $f(x)=0$, then $0=e^{-I(x)}f(x)\leq\Lambda(f)$, since Λ takes values in $\mathbb{R}^+$; if $f(x)>0$, by definition of I and by (i)

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

hence $e^{-I(x)}f(x)\leq\Lambda(f)$ again. This proves that

$$\sup_{x\in\mathbf{X}} e^{-I(x)}f(x)\leq\Lambda(f).$$

To prove the other inequality, we consider, for all $x\in\mathbf{X}$ and $n\geq 1$, the continuous functions

$$\psi_{n,x}\colon \mathbf{X}\to[0,1],\qquad \psi_{n,x}(t) = 0\vee(1-nd(B_{1/n}(x),t))\qquad \forall t\in\mathbf{X}.$$

We observe that $\psi_{n,x}$ equals 1 on $B_{1/n}(x)$ and 0 on $B_{2/n}(x)^c$. Since $\{B_{1/n}(x)\}_{x\in\mathbf{X}}$ is an open cover of the compact space $\mathbf{X}$, there exists $D_n\subseteq\mathbf{X}$ finite such that

$$\mathbf{X} = \bigcup_{x\in D_n} B_{1/n}(x).$$

Therefore, for all $t \in \mathbf{X}$, there exists $x \in D_n$ such that $t \in B_{1/n}(x)$, hence $\psi_{n,x}(t) = 1$. It follows that $\max_{x\in D_n} \psi_{n,x} = 1$. Let $f_0 \doteq f$. For all $n \geq 1$, we define f_n recursively in the following way: since by (iv)

$$\Lambda(f_{n-1}) = \Lambda\Big(\max_{x\in D_n} \psi_{n,x} f_{n-1}\Big) \leq \max_{x\in D_n} \Lambda(\psi_{n,x} f_{n-1}),$$

we can choose $x_n \in D_n$ such that $\Lambda(f_{n-1}) \leq \Lambda(f_n)$, where $f_n \doteq \psi_{n,x_n} f_{n-1}$. Moreover, $0 \leq \psi_{n,x_n} \leq 1$ implies that $0 \leq f_n \leq f_{n-1}$, hence by (iii) $\Lambda(f_{n-1}) \geq \Lambda(f_n)$. Therefore, $\Lambda(f_{n-1}) = \Lambda(f_n)$. Thus we have defined a decreasing sequence $\{f_n\}_{n\geq 0}$ of continuous nonnegative functions such that $\Lambda(f_{n-1}) = \Lambda(f_n)$ for all $n \geq 1$. We now study the pointwise limit

$$f_\infty \doteq \lim_{n\to\infty} f_n.$$

Assume that there exist $x, x' \in \mathbf{X}$ such that $f_\infty(x) > 0$ and $f_\infty(x') > 0$. Since f_n is decreasing, $f_n(x) > 0$ and $f_n(x') > 0$ for all $n \geq 1$. Since $f_n = \psi_{n,x_n} f_{n-1}$ and $\psi_{n,x_n} = 0$ on $B_{2/n}(x_n)^c$, $x, x' \in B_{2/n}(x_n)$ for all $n \geq 1$, hence

$$d(x,x') \leq d(x,x_n) + d(x_n,x') < \frac{2}{n} + \frac{2}{n} = \frac{4}{n} \xrightarrow{n\to\infty} 0.$$

Therefore, $d(x,x') = 0$, i.e. $x = x'$. This proves that there exists at most one point $\overline{x} \in \mathbf{X}$ such that $f_\infty(x) \neq 0$, i.e. $f_n \searrow c\mathbb{1}_{\{\overline{x}\}}$, $0 \leq c \leq f(\overline{x})$. Recalling that $\Lambda(f) = \Lambda(f_n)$ for all $n \geq 1$ and using (2.23),

$$\Lambda(f) = \lim_{n\to\infty} \Lambda(f_n) = ce^{-I(\overline{x})} \leq f(\overline{x})e^{-I(\overline{x})} \leq \sup_{x\in\mathbf{X}} f(x)e^{-I(x)},$$

which concludes the proof of (2.24).

It remains to show that I is a good rate function. By (v) and (2.24),

$$e^{-I(\overline{x})} \leq \sup_{x\in\mathbf{X}} e^{-I(x)} = \Lambda(\mathbf{1}) = 1 \qquad \forall \overline{x} \in \mathbf{X},$$

hence $I \geq 0$. If $I \equiv +\infty$, then $\Lambda(\mathbf{1}) = \sup_{x\in\mathbf{X}} e^{-I(x)} = 0$, contradicting (v); thus $I \not\equiv +\infty$. To prove the lower semicontinuity of I, we consider, for all $x \in \mathbf{X}$ and $n \geq 1$, the continuous functions

$$\varphi_{n,x} \colon \mathbf{X} \to [0,1], \qquad \varphi_{n,x}(t) = 0 \vee (1 - nd(x,t)) \qquad \forall t \in \mathbf{X}.$$

Since $\{\varphi_{n,x}\}_{n\geq 1}$ is decreasing, $\varphi_{n,x}(x) = 1$ and $\varphi_{n,x}(t) = 0$ for all $t \in B_{1/n}(x)^c$, we obtain $\varphi_{n,x} \searrow \mathbb{1}_{\{x\}}$ as $n \to \infty$; by (2.23), $\Lambda(\varphi_{n,x}) \searrow e^{-I(x)}$ for all $x \in \mathbf{X}$. Consider the composite function

$$\Lambda \circ \varphi_{n,\cdot} \colon \mathbf{X} \xrightarrow{\varphi_{n,\cdot}} C(\mathbf{X}) \xrightarrow{\Lambda} \mathbb{R}^+, \qquad x \to \Lambda(\varphi_{n,x}).$$

It turns out that:

- By the subadditivity of the seminorm Λ and the property (ii), if $f, g \in C(\mathbf{X})$

$$\Lambda(f) \leq \Lambda(f-g) + \Lambda(g) = \Lambda(|f-g|) + \Lambda(g)$$
$$\Lambda(g) \leq \Lambda(g-f) + \Lambda(f) = \Lambda(|f-g|) + \Lambda(f),$$

hence $|\Lambda(f) - \Lambda(g)| \leq \Lambda(|f-g|)$; moreover, by (iii), (i) and (v)

$$|\Lambda(f) - \Lambda(g)| \leq \Lambda(|f-g|) \leq \Lambda(\|f-g\|_\infty \mathbf{1}) = \|f-g\|_\infty \Lambda(\mathbf{1}) = \|f-g\|_\infty.$$

Therefore, $\Lambda\colon C(\mathbf{X}) \to \mathbb{R}^+$ is 1-Lipschitz continuous.

- If $x, x' \in \mathbf{X}$,

$$\begin{aligned}\|\varphi_{n,x} - \varphi_{n,x'}\|_\infty &\leq \sup_{t\in\mathbb{R}} |0 \vee (1 - nd(x,t)) - 0 \vee (1 - nd(x',t))| \\ &\leq \sup_{t\in\mathbb{R}} |(1 - nd(x,t)) - (1 - nd(x',t))| \\ &= n \sup_{t\in\mathbb{R}} |d(x,t) - d(x',t)| \leq nd(x,x'),\end{aligned}$$

where the latter inequality follows from the triangular inequality in $(\mathbf{X}, d)$. This proves that $\varphi_{n,\cdot}\colon \mathbf{X} \to C(\mathbf{X})$ is n-Lipschitz continuous.

Therefore, the composite function $\Lambda \circ \varphi_{n,\cdot}$ is continuous. It follows that e^{-I} is the pointwise limit of a decreasing sequence of continuous functions, hence it is upper semi-continuous by Theorem 2.5 (vii). In other words, I is lower semi-continuous. Finally, any level set of I is closed by lower semi-continuity, hence it is compact since $\mathbf{X}$ is compact. This completes the proof that I is a good rate function. □

We are now ready to prove the second statement of Theorem 2.58 for compact spaces, where exponential tightness is always automatically verified. Therefore, it will turn out that in a compact space *any* sequence of probability measures has a subsequence that satisfies a LDP: this might look weird, but on the other hand such a result reminds the well-known fact that in a compact space *any* sequence has a convergent subsequence.

Theorem 2.63. Let $\mathbf{X}$ be a compact metrizable space (i.e. a compact Polish space). If $\mu_n \in \mathfrak{M}(\mathbf{X})$ and $1 \leq s_n \to \infty$, there exists a subsequence μ_{n_k} and a good rate function I such that μ_{n_k} satisfies the LDP with rate s_{n_k} and rate function I.

Proof. We note that, since μ_n are *probability* measures,

$$\|f\|_{s_n,\mu_n} \leq \|f\|_\infty \qquad \forall f \in C(\mathbf{X}).$$

By Example 2.17 (ii), $(C(\mathbf{X}), \|\cdot\|_\infty)$ is a Polish space, hence there exists a dense sequence $D = \{f_i\}_{i\geq 1}$ in $C(\mathbf{X})$. We set

$$\Pi \doteq \bigtimes_{i=1}^{\infty} [0, \|f_i\|_\infty] \subseteq \mathbb{R}^{\mathbb{N}}.$$

For all $i \geq 1$, $0 \leq \|f_i\|_{s_n,\mu_n} \leq \|f_i\|_\infty$, hence $\{\|f_i\|_{s_n,\mu_n}\}_{i\geq 1} \in \Pi$ for all $n \geq 1$. In other words, $\{\{\|f_i\|_{s_n,\mu_n}\}_{i\geq 1}\}_{n\geq 1}$ is a sequence in Π. Since Π is a countable product of compact metric spaces, there exists a subsequence $\{\{\|f_i\|_{s_{n_k},\mu_{n_k}}\}_{i\geq 1}\}_{k\geq 1}$ that converges to a limit† that we call $\{\Lambda(f_i)\}_{i\geq 1} \in \Pi$. We can see Λ as a function $D \to \mathbb{R}^+$, since

$$\Lambda(f_i) = \lim_{k\to\infty} \|f_i\|_{s_{n_k},\mu_{n_k}} \qquad \forall i \geq 1. \tag{2.25}$$

By the triangular inequality of the seminorm $\|\cdot\|_{s_{n_k},\mu_{n_k}}$,

$$\left| \|f\|_{s_{n_k},\mu_{n_k}} - \|g\|_{s_{n_k},\mu_{n_k}} \right| \leq \|f-g\|_{s_{n_k},\mu_{n_k}} \leq \|f-g\|_\infty \qquad \forall f,g \in C(\mathbf{X}). \tag{2.26}$$

Taking $f = f_i$ and $g = f_j$ $(i,j \geq 1)$, and passing to the limit as $k \to \infty$,

$$|\Lambda(f_i) - \Lambda(f_j)| \leq \|f_i - f_j\|_\infty \qquad \forall i,j \geq 1.$$

Therefore, $\Lambda\colon D \to \mathbb{R}^+$ is 1-Lipschitz continuous. Since D is dense in $C(\mathbf{X})$, there exists a (unique) 1-Lipschitz continuous function $C(\mathbf{X}) \to \mathbb{R}^+$ that extends Λ: without ambiguity, we can call such a function Λ again. We claim that (2.25) holds for any $f \in C(\mathbf{X})$, i.e.

$$\Lambda(f) = \lim_{k\to\infty} \|f\|_{s_{n_k},\mu_{n_k}} \qquad \forall f \in C(\mathbf{X}). \tag{2.27}$$

To prove this, let $\varepsilon > 0$ and $i \geq 1$ such that $\|f - f_i\|_\infty < \frac{\varepsilon}{3}$. Since $\|f_i\|_{s_{n_k},\mu_{n_k}} \to \Lambda(f_i)$ by (2.25), $\left|\|f_i\|_{s_{n_k},\mu_{n_k}} - \Lambda(f_i)\right| < \frac{\varepsilon}{3}$ for k large enough, hence

$$\begin{aligned}
\left|\|f\|_{s_{n_k},\mu_{n_k}} - \Lambda(f)\right| &\leq \left|\|f\|_{s_{n_k},\mu_{n_k}} - \|f_i\|_{s_{n_k},\mu_{n_k}}\right| + \left|\|f_i\|_{s_{n_k},\mu_{n_k}} - \Lambda(f_i)\right| + |\Lambda(f_i) - \Lambda(f)| \\
&\leq \|f - f_i\|_\infty + \left|\|f_i\|_{s_{n_k},\mu_{n_k}} - \Lambda(f_i)\right| + \|f - f_i\|_\infty \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon,
\end{aligned}$$

for k large enough (the second inequality follows from (2.26) and the 1-Lipschitz continuity of Λ).

If we now prove for Λ the properties (i)-(v) of Lemma 2.62, we will obtain that there exists a good rate function I such that $\Lambda(f) = \|f\|_{\infty,I}$ for all $f \in C(\mathbf{X})$, and from (2.27) it will follow that

$$\|f\|_{\infty,I} = \lim_{k\to\infty} \|f\|_{s_{n_k},\mu_{n_k}} \qquad \forall f \in C(\mathbf{X}),$$

†This fact can be justified by invoking Tychonoff's Theorem, but there is a straightforward diagonal argument to show it. Assume that $\{C_i\}_{i\geq 1}$ is a sequence of compact metric spaces, Π is their product and $\{\{x_i(n)\}_{i\geq 1}\}_{n\geq 1}$ is a sequence in Π (in our specific case, $C_i = [0, \|f_i\|_\infty]$ and $x_i(n) = \|f_i\|_{s_n,\mu_n}$). Let $n_{0,k} \doteq k$, $k \geq 1$. Recursively, assume that we have defined $n_{i-1,k}$ and let us define $n_{i,k}$, $i \geq 1$: since the sequence $\{x_i(n_{i-1,k})\}_{k\geq 1}$ takes values in the compact metric space C_i, we can extract a subsequence $\{n_{i,k}\}_{k\geq 1}$ from $\{n_{i-1,k}\}_{k\geq 1}$ such that $x_i(n_{i,k}) \to x_i$ as $k \to \infty$, for some $x_i \in C_i$. Therefore, the "diagonal" subsequence $\{\{x_i(n_{k,k})\}_{i\geq 1}\}_{k\geq 1}$ converges to $\{x_i\}_{\geq 1} \in \Pi$. Indeed, for all i, $x_i(n_{k,k}) \to x_i$ as $k \to \infty$, since $\{n_{k,k}\}_{k\geq i}$ is a subsequence of $\{n_{i,k}\}_{k\geq 1}$ and $x_i(n_{i,k}) \to x_i$ as $k \to \infty$. Thus we have built the desired convergent subsequence of $\{\{x_i(n)\}_{i\geq 1}\}_{n\geq 1}$.

i.e. μ_{n_k} satisfies a LDP with rate s_{n_k} and good rate function I. The properties (i), (ii), (iii) and (v) are satisfied by $\|\cdot\|_{s_{n_k},\mu_{n_k}}$, hence also by Λ (it suffices to pass to the limit as $k \to \infty$). To prove (iv), let $f, g \in C(\mathbf{X})$, $f, g \geq 0$: if we apply (1.20) with $\alpha_k = \|f\|_{s_{n_k},\mu_{n_k}}^{s_{n_k}}$ and $\beta_k = \|g\|_{s_{n_k},\mu_{n_k}}^{s_{n_k}}$, we obtain

$$\begin{aligned}
\Lambda(f \vee g) &= \lim_{k\to\infty} \|f \vee g\|_{s_{n_k},\mu_{n_k}} \\
&= \lim_{k\to\infty} \left(\int_{\{f\geq g\}} f^{s_{n_k}}\, d\mu_{n_k} + \int_{\{f<g\}} g^{s_{n_k}}\, d\mu_{n_k} \right)^{1/s_{n_k}} \\
&\leq \limsup_{k\to\infty} \left(\|f\|_{s_{n_k},\mu_{n_k}}^{s_{n_k}} + \|g\|_{s_{n_k},\mu_{n_k}}^{s_{n_k}} \right)^{1/s_{n_k}} \\
&= \limsup_{k\to\infty} \left(\|f\|_{s_{n_k},\mu_{n_k}} \vee \|g\|_{s_{n_k},\mu_{n_k}} \right)(1 + o(1)) \\
&= \left(\limsup_{k\to\infty} \|f\|_{s_{n_k},\mu_{n_k}} \right) \vee \left(\limsup_{k\to\infty} \|g\|_{s_{n_k},\mu_{n_k}} \right) \\
&= \Lambda(f) \vee \Lambda(g).
\end{aligned}$$

□

Since we have proved the second statement of Theorem 2.58 for compact Polish spaces, for the extension to generic Polish spaces we will use a topological technique called *compactification*. We now discuss this notion, give some examples and state the result we will use about compactification of Polish spaces.

Definition 2.64. Let $(\mathbf{X}, \tau)$ be a topological space. A *compactification* of $\mathbf{X}$ is a topological space $\overline{\mathbf{X}}$ such that:

(i) $\mathbf{X}$ is a dense subset of $\overline{\mathbf{X}}$.

(ii) The induced topology from $\overline{\mathbf{X}}$ on $\mathbf{X}$ is τ.

(iii) $\overline{\mathbf{X}}$ is compact.

Remark 2.65. (iii) is obviously the fundamental property of a compactification. Property (i) is needed to guarantee that the compactification is not "too big"; it justifies the notation, since it turns out that the closure of $\mathbf{X}$ in the topology of the compactification is precisely $\overline{\mathbf{X}}$. Property (ii) is needed to guarantee that we are just extending the topology on $\mathbf{X}$; it is equivalent to require that the embedding $\mathbf{X} \hookrightarrow \overline{\mathbf{X}}$ is a homeomorphism onto its image (indeed, this means that a subset of $\mathbf{X}$ is open if and only if it is of the form $O \cap \mathbf{X}$, where O is an open subset of $\overline{\mathbf{X}}$, i.e. the open sets of $\mathbf{X}$ are the open sets of the induced topology from $\overline{\mathbf{X}}$). Finally, it is important to observe that compactifications are not usually unique. ♢

Example 2.66. A compactification of a bounded set $A \subseteq \mathbb{R}^d$ (with the euclidean topology) is obviously the closure $\overline{A}$, since a closed and bounded subset of $\mathbb{R}^d$ is compact. A compactification of $\mathbb{R}$ is $\mathbb{R} \cup \{-\infty, +\infty\}$, with the "euclidean" topology we already introduced in §2.1. ♢

Example 2.67. For any topological space $\mathbf{X}$, one can obtain a compactification in a standard way by adding only one extra point ∞, so that $\overline{\mathbf{X}} = \mathbf{X} \cup \{\infty\}$. This is called *(Alexandroff) one-point compactification*, and it turns out to be Hausdorff if and only if $\mathbf{X}$ is Hausdorff and locally compact† (for details, see [Man, §4.7 and Prop.4.62]). In particular, it is useful to note that, if $\mathbf{X}$ is not locally compact, its one-point compactification is not metrizable. ♢

We are in particular interested in metrizable compactifications, since they are Polish spaces by Example 2.18.

Theorem 2.68. Any Polish space has a metrizable compactification.

Proof. See [Cho, Th.6.3]. □

Remark 2.69. If $\mathbf{X}$ is a Polish space and $\overline{\mathbf{X}}$ is a metrizable compactification of $\mathbf{X}$, by definition the topology on $\overline{\mathbf{X}}$ is an extension of the topology on $\mathbf{X}$. One might also think that the metric on $\overline{\mathbf{X}}$ is an extension of a complete metric on $\mathbf{X}$; however, this is *always* false, unless $\mathbf{X}$ is compact. Indeed, let d be a metric on $\overline{\mathbf{X}}$ such that $(\overline{\mathbf{X}}, d)$ and $(\mathbf{X}, d|_{\mathbf{X}\times\mathbf{X}})$ are complete metric spaces, and let $x \in \overline{\mathbf{X}}$. Since $\mathbf{X}$ is dense in $\overline{\mathbf{X}}$, there exists a sequence $x_n \in \mathbf{X}$ such that $d(x_n, x) \to 0$; in particular, x_n is a Cauchy sequence in $(\mathbf{X}, d|_{\mathbf{X}\times\mathbf{X}})$, hence it converges in $\mathbf{X}$, i.e. $x \in \mathbf{X}$. This proves that $\overline{\mathbf{X}} = \mathbf{X}$, i.e. $\mathbf{X}$ is compact.

In particular, a metrizable compactification of $\mathbb{R}^d$ cannot be obtained by the extension of a metric on $\mathbb{R}^d$ that is induced by any norm, since all these metrics on $\mathbb{R}^d$ are complete. For example, the compactification $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ of $\mathbb{R}$ we have already discussed is metrizable, since a metric on $\overline{\mathbb{R}}$ is given by

$$d(x,y) \doteq |g(x) - g(y)| \quad \forall x, y \in \overline{\mathbb{R}}, \qquad g(x) \doteq \begin{cases} -1 & x = -\infty, \\ \dfrac{x}{1+|x|} & x \in \mathbb{R}, \\ 1 & x = +\infty \end{cases}$$

(see [Soa, §3.9]). Coherently with what we observed, this metric is not induced by any norm. ♢

Let $\mathbf{X}$ be a Polish space and *assume* that its one-point compactification $\mathbf{X} \cup \{\infty\}$ (see Example 2.67) is metrizable; then, $\mathbf{X} \cup \{\infty\}$ is Hausdorff and $\{\infty\}$ is closed in $\mathbf{X} \cup \{\infty\}$, hence $\mathbf{X}$ is an open subset (in particular, a Borel subset) of $\mathbf{X} \cup \{\infty\}$. Unfortunately, as we illustrated in Example 2.67, the one-point compactification is not metrizable in many interesting cases (for example, when $\mathbf{X}$ is not locally compact). However, Theorem 2.68 guarantees in any case that $\mathbf{X}$ has a metrizable compactification $\overline{\mathbf{X}}$, so that $\mathbf{X}$ and $\overline{\mathbf{X}}$ are both Polish; by Theorem 2.21, $\mathbf{X}$ is a G_δ-subset of $\overline{\mathbf{X}}$ (i.e. a countable intersection of open subsets), in particular we are able

†We recall that a topological space is said to be locally compact if any point has a compact neighborhood.

to conclude that $\mathbf{X}$ *is a Borel subset of* $\overline{\mathbf{X}}$ again. As the following lemma rigorously states, this means that any probability measure μ on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ has an extension to $(\overline{\mathbf{X}}, \mathcal{B}(\overline{\mathbf{X}}))$ that "behaves like" μ, and vice versa.

Lemma 2.70. Let $\mathbf{X}$ be a topological space and $\mathbf{Y} \in \mathcal{B}(\mathbf{X})$, equipped with the induced topology.

(i) If $\mu \in \mathfrak{M}(\mathbf{X})$ and $\mu(\mathbf{X} \setminus \mathbf{Y}) = 0$, then the restriction $\mu|_{\mathcal{B}(\mathbf{Y})} \in \mathfrak{M}(\mathbf{Y})$.

(ii) If $\mu \in \mathfrak{M}(\mathbf{Y})$, there exists $\overline{\mu} \in \mathfrak{M}(\mathbf{X})$ that extends μ and such that $\overline{\mu}(\mathbf{X} \setminus \mathbf{Y}) = 0$.

Proof. We first remark that, since $\mathbf{Y}$ is equipped with the induced topology and $\mathbf{Y} \in \mathcal{B}(\mathbf{X})$,

$$\mathcal{B}(\mathbf{Y}) = \{A \cap \mathbf{Y} : \quad A \in \mathcal{B}(\mathbf{X})\} \subseteq \mathcal{B}(\mathbf{X}).$$

(i) If μ is a measure on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$, then $\mu|_{\mathcal{B}(\mathbf{Y})}$ is a measure on $(\mathbf{Y}, \mathcal{B}(\mathbf{Y}))$, since $\mathcal{B}(\mathbf{Y}) \subseteq \mathcal{B}(\mathbf{X})$. If μ is a probability measure such that $\mu(\mathbf{X} \setminus \mathbf{Y}) = 0$, then

$$\mu|_{\mathcal{B}(\mathbf{Y})}(\mathbf{Y}) = \mu(\mathbf{Y}) + \mu(\mathbf{X} \setminus \mathbf{Y}) = \mu(\mathbf{X}) = 1,$$

hence $\mu|_{\mathcal{B}(\mathbf{Y})}$ is also a probability measure.

(ii) Let μ be a probability measure on $(\mathbf{Y}, \mathcal{B}(\mathbf{Y}))$. We define the function

$$\overline{\mu} \colon \mathcal{B}(\mathbf{X}) \to [0,1], \qquad \overline{\mu}(A) = \mu(A \cap \mathbf{Y}) \quad \forall A \in \mathcal{B}(\mathbf{X}).$$

It turns out that $\overline{\mu}$ is a probability measure on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$, since:

- $\overline{\mu}(\varnothing) = \mu(\varnothing) = 0$.
- If $\{A_k\}_{k \geq 1} \subseteq \mathcal{B}(\mathbf{X})$ is a collection of disjoint sets,

$$\overline{\mu}\left(\bigcup_{k=1}^{\infty} A_k\right) = \mu\left(\left(\bigcup_{k=1}^{\infty} A_k\right) \cap \mathbf{Y}\right) = \mu\left(\bigcup_{k=1}^{\infty} (A_k \cap \mathbf{Y})\right) = \sum_{k=1}^{\infty} \mu(A_k \cap \mathbf{Y}) = \sum_{k=1}^{\infty} \overline{\mu}(A_k),$$

  by the countable additivity of μ.
- $\overline{\mu}(\mathbf{X}) = \mu(\mathbf{X} \cap \mathbf{Y}) = \mu(\mathbf{Y}) = 1$, since μ is a *probability* measure.

Finally, $\overline{\mu}(\mathbf{X} \setminus \mathbf{Y}) = \mu((\mathbf{X} \setminus \mathbf{Y}) \cap \mathbf{Y}) = \mu(\varnothing) = 0$. □

We will use the latter theorem for the embedding of Polish spaces $\mathbf{X} \hookrightarrow \overline{\mathbf{X}}$, where $\overline{\mathbf{X}}$ is a metrizable compactification of $\mathbf{X}$. We note that, in case that $\overline{\mathbf{X}} = \mathbf{X} \cup \{\infty\}$ is the one-point compactification, the extension of a measure $\mu \in \mathfrak{M}(\mathbf{X})$ to $\overline{\mu} \in \mathfrak{M}(\overline{\mathbf{X}})$ is trivial: it suffices to set $\overline{\mu}(\{\infty\}) \doteq 0$.

If $\mathbf{X}$ is a Polish space and $\mathbf{Y}$ is a Polish subspace of $\mathbf{X}$, by Lemma 2.70 we may see "$\mathfrak{M}(\mathbf{Y}) \subseteq \mathfrak{M}(\mathbf{X})$": any $\nu \in \mathfrak{M}(\mathbf{Y})$ can be extended to a $\overline{\nu} \in \mathfrak{M}(\mathbf{X})$ such that $\overline{\nu}(\mathbf{X} \setminus \mathbf{Y}) = 0$, and vice versa. It is natural to wonder if the weak topology of $\mathfrak{M}(\mathbf{Y})$ coincides with the topology that $\mathfrak{M}(\mathbf{X})$ induces on $\mathfrak{M}(\mathbf{Y})$. We see in the following proposition that the answer is positive.

Lemma 2.71. Let $\mathbf{X}$ be a Polish space and let $\mathbf{Y}$ be a Polish subspace of $\mathbf{X}$. Then the inclusion $\mathfrak{M}(\mathbf{Y}) \hookrightarrow \mathfrak{M}(\mathbf{X})$ (in the sense of Lemma 2.70) is a topological embedding, i.e. the weak topology of $\mathfrak{M}(\mathbf{X})$ induces on $\mathfrak{M}(\mathbf{Y})$ the weak topology of $\mathfrak{M}(\mathbf{Y})$.

Proof. Since $\mathfrak{M}(\mathbf{X})$ and $\mathfrak{M}(\mathbf{Y})$ are metrizable (see Example 2.19), their topologies are entirely determined by convergence of sequences, so it suffices to prove that, for any $\mu_n, \mu \in \mathfrak{M}(\mathbf{Y})$, $\overline{\mu}_n \rightharpoonup \overline{\mu}$ in $\mathfrak{M}(\mathbf{X})$ if and only if $\mu_n \rightharpoonup \mu$ in $\mathfrak{M}(\mathbf{Y})$. By Theorem 2.23, $\overline{\mu}_n \rightharpoonup \overline{\mu}$ in the weak topology of $\mathfrak{M}(\mathbf{X})$ if and only if

$$\liminf_{n\to\infty} \overline{\mu}_n(O) \geq \overline{\mu}(O) \qquad \forall O \subseteq \mathbf{X} \text{ open.} \tag{2.28}$$

Since $\mu_n, \mu \in \mathfrak{M}(\mathbf{Y})$, $\overline{\mu}_n(\mathbf{X} \setminus \mathbf{Y}) = 0$ and $\overline{\mu}(\mathbf{X} \setminus \mathbf{Y}) = 0$; in particular, $\overline{\mu}_n(O) = \mu_n(O \cap \mathbf{Y})$ and $\overline{\mu}(O) = \mu(O \cap \mathbf{Y})$ for all $O \subseteq \mathbf{X}$ open, hence (2.28) is equivalent to

$$\liminf_{n\to\infty} \mu_n(O \cap \mathbf{Y}) \geq \mu(O \cap \mathbf{Y}) \qquad \forall O \subseteq \mathbf{X} \text{ open.} \tag{2.29}$$

Each open subset of $\mathbf{Y}$ is of the form $O \cap \mathbf{Y}$ for some $O \subset \mathbf{X}$ open, hence (2.29) holds if and only if $\mu_n \rightharpoonup \mu$ in the weak topology of $\mathfrak{M}(\mathbf{Y})$, by Theorem 2.23 again. □

The latter lemma essentially states that it is equivalent to require that a sequence of probability measures, supported in a Polish subspace of a Polish space, converges weakly in the subspace or in the whole space. Of course, we have a large deviations analogue: the following "restriction principle" states that, under certain conditions, it is equivalent to require that a sequence of probability measures, supported in a Polish subspace of a Polish space, satisfies a LDP in the subspace or in the whole space.

Lemma 2.72 (Restriction principle)**.** Let $\mathbf{X}$ be a Polish space and let $\mathbf{Y}$ be a Polish subspace. Let $\mu_n \in \mathfrak{M}(\mathbf{X})$ such that $\mu_n(\mathbf{X} \setminus \mathbf{Y}) = 0$ for all $n \geq 1$ and $1 \leq s_n \to \infty$. Let I be a good rate function on $\mathbf{X}$ such that $I(x) = +\infty$ for all $x \in \mathbf{X} \setminus \mathbf{Y}$. Then, μ_n satisfies the LDP with rate s_n and good rate function I if and only if $\mu_n|_{\mathcal{B}(\mathbf{Y})}$ satisfies the LDP with rate s_n and good rate function $I|_{\mathbf{Y}}$.

Proof. By Theorem 2.21, $\mathbf{Y}$ is a G_δ-subset of $\mathbf{X}$, hence a Borel subset. By Lemma 2.70 (i), $\mu_n|_{\mathcal{B}(\mathbf{Y})} \in \mathfrak{M}(\mathbf{Y})$.

Let us verify that $I|_{\mathbf{Y}}$ is a good rate function on $\mathbf{Y}$. $I|_{\mathbf{Y}} \geq 0$, since $I \geq 0$. Since $I(x) = +\infty$ for all $x \in \mathbf{X} \setminus \mathbf{Y}$ by hypothesis, if $I|_{\mathbf{Y}} \equiv +\infty$ then $I \equiv +\infty$, contradicting the definition of rate function; hence, $I|_{\mathbf{Y}} \not\equiv +\infty$. Since $I(x) = +\infty$ for all $x \in \mathbf{X} \setminus \mathbf{Y}$, the level sets of $I|_{\mathbf{Y}}$

$$\{x \in \mathbf{Y} : \ I|_{\mathbf{Y}}(x) \leq a\} = \{x \in \mathbf{X} : \ I(x) \leq a\} \qquad (a \in \mathbb{R})$$

are actually level sets of I: in particular, they are compact subsets of $\mathbf{X}$ that are contained in $\mathbf{Y}$, hence they are compact in $\mathbf{Y}$. We conclude that $I|_{\mathbf{Y}}$ is a good rate function on $\mathbf{Y}$.

We now prove the equivalence of the LDP's. For all $n \geq 1$, $A \in \mathcal{B}(\mathbf{X})$,

$$\mu_n(A) = \mu_n(A \cap \mathbf{Y}) + \mu_n(A \cap (\mathbf{X} \setminus \mathbf{Y})) = \mu_n|_{\mathcal{B}(\mathbf{Y})}(A \cap \mathbf{Y}),$$

since $\mu_n(A \cap (\mathbf{X} \setminus \mathbf{Y})) \leq \mu_n(\mathbf{X} \setminus \mathbf{Y}) = 0$. Moreover, for all $A \subseteq \mathbf{X}$

$$\inf_{x \in A} I(x) = \inf_{x \in A \cap \mathbf{Y}} I|_{\mathbf{Y}}(x),$$

since $I(x) = +\infty$ for all $x \in \mathbf{X} \setminus \mathbf{Y}$. Therefore, the following two conditions are equivalent:

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

$$\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n|_{\mathcal{B}(\mathbf{Y})}(C \cap \mathbf{Y}) \leq -\inf_{x \in C \cap \mathbf{Y}} I|_{\mathbf{Y}}(x) \qquad \forall C \subseteq \mathbf{X} \text{ closed.}$$

Since a subset of $\mathbf{Y}$ is closed (with respect to the induced topology from $\mathbf{X}$) if and only if it is of the form $C \cap \mathbf{Y}$, $C \subseteq \mathbf{X}$ closed, this proves the equivalence of the large deviations upper bounds. The proof of the equivalence of the lower bounds is similar. Thus the claim follows from Theorem 2.31. □

We are now able to prove the second statement of Theorem 2.58 in the general case.

Theorem 2.73. Let $\mathbf{X}$ be a Polish space. If $\mu_n \in \mathfrak{M}(\mathbf{X})$ is exponentially tight with rate s_n, then there exist a good rate function I and a subsequence μ_{n_k} such that μ_{n_k} satisfies the LDP with rate s_{n_k} and rate function I.

Proof. By Theorem 2.68, there exists a metrizable compactification $\overline{\mathbf{X}}$ of $\mathbf{X}$. By Lemma 2.70 (ii), the probability measures $\mu_n \in \mathfrak{M}(\mathbf{X})$ can be extended to probability measures $\overline{\mu}_n \in \mathfrak{M}(\overline{\mathbf{X}})$ such that $\overline{\mu}_n(\overline{\mathbf{X}} \setminus \mathbf{X}) = 0$. By Theorem 2.63, there exist a subsequence $\overline{\mu}_{n_k}$ and a good rate function I on $\overline{\mathbf{X}}$ such that $\overline{\mu}_{n_k}$ satisfies the LDP with rate s_{n_k} and rate function I. To apply the restriction principle to the embedding of Polish spaces $\mathbf{X} \hookrightarrow \overline{\mathbf{X}}$, we only have to show that $I(x) = +\infty$ for all $x \in \overline{\mathbf{X}} \setminus \mathbf{X}$. If $\overline{x} \in \overline{\mathbf{X}}$ satisfies $I(\overline{x}) < +\infty$, then there exists $M \in \mathbb{R}$ such that $I(\overline{x}) < M$. Since μ_n is exponentially tight with rate s_n, there exists a compact subset K of $\mathbf{X}$ such that

$$\limsup_{k\to\infty} \frac{1}{s_{n_k}} \log \mu_{n_k}(\mathbf{X} \setminus K) \leq -M. \tag{2.30}$$

We claim that $\overline{x} \in K$: assume conversely that $\overline{x} \in \overline{\mathbf{X}} \setminus K$. Since $\overline{\mathbf{X}}$ is a compactification of $\mathbf{X}$, $\overline{\mathbf{X}}$ and $\mathbf{X}$ induce the same topology on K, hence K is also compact as a subset of $\overline{\mathbf{X}}$. Since $\overline{\mathbf{X}}$, as a metrizable space, is Hausdorff, K is a closed subset of $\overline{\mathbf{X}}$, hence

$\overline{\mathbf{X}} \setminus K$ is an *open* neighborhood of $\overline{x}$. Using the large deviations lower bound for $\overline{\mu}_{n_k}$, the fact that $I(\overline{x}) < M$ and inequality (2.30), we obtain

$$\liminf_{k\to\infty} \frac{1}{s_{n_k}} \log \overline{\mu}_{n_k}(\overline{\mathbf{X}} \setminus K) \geq -\inf_{x\in\overline{\mathbf{X}}\setminus K} I(x) \geq -I(\overline{x}) > -M \geq \limsup_{k\to\infty} \frac{1}{s_{n_k}} \log \mu_{n_k}(\mathbf{X} \setminus K).$$

However, $\overline{\mu}_{n_k}(\overline{\mathbf{X}} \setminus K) = \mu_{n_k}((\overline{\mathbf{X}} \setminus K) \cap \mathbf{X}) = \mu_{n_k}(\mathbf{X} \setminus K)$, hence the inequality above leads to a contradiction, and $\overline{x} \in K \subseteq \mathbf{X}$. In particular, this argument proves that for all $\overline{x} \in \overline{\mathbf{X}}$ such that $I(\overline{x}) < +\infty$, one has $\overline{x} \in \mathbf{X}$; in other words, $I(\overline{x}) = +\infty$ for all $\overline{x} \in \overline{\mathbf{X}} \setminus \mathbf{X}$. Thus, we can apply the restriction principle: since $\overline{\mu}_{n_k}$ satisfies the LDP with rate s_{n_k} and rate function I, $\overline{\mu}_{n_k}|_{\mathcal{B}(\mathbf{X})} = \mu_{n_k}$ satisfies the LDP with rate s_{n_k} and good rate function $I|_{\mathbf{X}}$. □

By using the latter theorem, we can finally prove a useful criterion to show that a sequence of probability measures μ_n satisfies a LDP: in the presence of exponential tightness, it suffices to check that $\lim_{n\to\infty}\|f\|_{s_n,\mu_n}$ exists (and not necessarily for each $f \in C_{b,+}(\mathbf{X})$, it is sufficient for each f in a rate function determining set).

Proposition 2.74. Let $\mathbf{X}$ be a Polish space, let $\mu_n \in \mathfrak{M}(\mathbf{X})$ and $1 \leq s_n \to \infty$. Assume that

(i) μ_n is exponentially tight with rate s_n.

(ii) The limit $\Lambda(f) \doteq \lim_{n\to\infty}\|f\|_{s_n,\mu_n}$ exists for all $f \in D$, where $D \subseteq C_{b,+}(\mathbf{X})$ is a rate function determining set.

Then, there exists a unique good rate function I such that $\Lambda(f) = \|f\|_{\infty,I}$ for all $f \in D$. Moreover, μ_n satisfies the LDP with rate s_n and rate function I.

Proof. By exponential tightness, there exist a subsequence μ_{n_k} and a good rate function I such that μ_{n_k} satisfies the LDP with rate s_{n_k} and rate function I (Theorem 2.73). By (ii),

$$\Lambda(f) = \lim_{k\to\infty} \|f\|_{s_{n_k},\mu_{n_k}} = \|f\|_{\infty,I} \qquad \forall f \in D.$$

For any other good rate function I' such that $\Lambda(f) = \|f\|_{\infty,I'}$ for all $f \in D$, it turns out that

$$\|f\|_{\infty,I} = \|f\|_{\infty,I'} \qquad \forall f \in D,$$

hence $I = I'$, since D is rate function determining.

Assume now that μ_n does not satisfy the LDP with rate s_n and rate function I, i.e. there exist $g \in C_{b,+}(\mathbf{X})$, $\varepsilon > 0$ and a subsequence μ_{m_k} such that

$$\Big| \|g\|_{s_{m_k},\mu_{m_k}} - \|g\|_{\infty,I} \Big| \geq \varepsilon \qquad \forall k \geq 1. \tag{2.31}$$

By exponentially tightness again, there exist a subsequence $\mu_{m_{k_h}}$ and a good rate function I'' such that $\mu_{m_{k_h}}$ satisfies the LDP with rate $s_{m_{k_h}}$ and rate function I''.

In particular, by (ii) $\Lambda(f) = \|f\|_{\infty,I''}$ for all $f \in D$; since I is the unique good rate function satisfying this property, $I = I''$. Therefore, $\mu_{m_{k_h}}$ satisfies the LDP with rate $s_{m_{k_h}}$ and rate function I, which contradicts (2.31). We conclude that μ_n satisfies the LDP with rate s_n and rate function I. □

2.6 Contraction principle and projective limit

In this section we deal with two important tools of the large deviations theory, the contraction principle and the projective limit theorem: both of them regard *image measures* under a *continuous* function, and they may be considered in a certain way inverses of each other.

Consider two Polish spaces $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ and $(\mathbf{Y}, \mathcal{B}(\mathbf{Y}))$ with their Borel σ-algebras. We note that, if X is an $\mathbf{X}$-valued random variable with law μ and $\varphi\colon \mathbf{X} \to \mathbf{Y}$ is measurable, then $\varphi(X)$ is an $\mathbf{Y}$-valued random variable with law $\mu \circ \varphi^{-1}$ (the *image measure* under φ of the measure μ). Indeed, for all $A \in \mathcal{B}(\mathbf{Y})$

$$\mathbb{P}(\varphi(X) \in A) = \mathbb{P}(X \in \varphi^{-1}(A)) = \mu(\varphi^{-1}(A))$$

(observe that $\varphi^{-1}(A) \in \mathcal{B}(\mathbf{X})$, since φ is measurable).

If now φ is also continuous and $\{X_n\}_{n\geq 1}$ is a sequence of $\mathbf{X}$-valued random variables whose laws converge weakly to the law of X, then the laws of the $\varphi(X_n)$ also converge weakly to the law of $\varphi(X)$. This fact follows from the following classical theorem.

Theorem 2.75 (Continuous Mapping Theorem)**.** Let $\mathbf{X}$ and $\mathbf{Y}$ be Polish spaces and $\mu_n, \mu \in \mathfrak{M}(\mathbf{X})$. If $\varphi\colon \mathbf{X} \to \mathbf{Y}$ is continuous and $\mu_n \rightharpoonup \mu$, then also $\mu_n \circ \varphi^{-1} \rightharpoonup \mu \circ \varphi^{-1}$.

Proof. Let $C \subseteq \mathbf{Y}$ be closed. Since φ is continuous, $\varphi^{-1}(C)$ is closed, hence by Theorem 2.23 (i)

$$\limsup_{n\to\infty} \mu_n(\varphi^{-1}(C)) \leq \mu(\varphi^{-1}(C)).$$

By Theorem 2.23 (i) again, we conclude that $\mu_n \circ \varphi^{-1} \rightharpoonup \mu \circ \varphi^{-1}$. □

In the perspective of the analogy between weak convergence of probability measures and LDP's, we now want to prove an analogous result for large deviations theory: if φ is continuous and $\{X_n\}_{n\geq 1}$ is a sequence of $\mathbf{X}$-valued random variables whose laws satisfy a LDP, then the laws of the $\varphi(X_n)$ also satisfy a LDP, as the following theorem will show. It is natural to think that the rate of LDP is the same, but we have to determine the new rate function.

Theorem 2.76 (Contraction principle)**.** Let $\mathbf{X}$ and $\mathbf{Y}$ be Polish spaces, let I be a good rate function on $\mathbf{X}$ and let $\varphi\colon \mathbf{X} \to \mathbf{Y}$ be continuous. Then

$$J(y) \doteq \inf_{x\in\varphi^{-1}(y)} I(x), \qquad y \in \mathbf{Y}$$

is a good rate function on $\mathbf{Y}$. Moreover, if a sequence $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the LDP in $\mathbf{X}$ with rate s_n and good rate function I, then $\mu_n \circ \varphi^{-1}$ satisfies the LDP in $\mathbf{Y}$ with rate s_n and good rate function J.

Proof. We start noting that for all $A \in \mathcal{B}(\mathbf{Y})$

$$\inf_{x\in\varphi^{-1}(A)} I(x) = \inf_{y\in A} \inf_{x\in\varphi^{-1}(y)} I(x) = \inf_{y\in A} J(y). \tag{2.32}$$

Let us prove that J is a good rate function: we have $J \geq 0$ since $I \geq 0$, and by (2.32)

$$\inf_{y\in\mathbf{Y}} J(y) = \inf_{x\in\varphi^{-1}(\mathbf{Y})} I(x) = \inf_{x\in\mathbf{X}} I(x) < +\infty$$

since $I \not\equiv +\infty$, hence also $J \not\equiv +\infty$. Moreover, if $a \in \mathbb{R}$,

$$\begin{aligned}
\{J \leq a\} &= \left\{y \in \mathbf{Y} : \quad \inf_{x\in\varphi^{-1}(y)} I(x) \leq a\right\} \\
&= \bigcap_{\varepsilon>0} \left\{y \in \mathbf{Y} : \quad \inf_{\varphi(x)=y} I(x) < a+\varepsilon\right\} \\
&= \bigcap_{\varepsilon>0} \{y \in \mathbf{Y} : \quad \exists x \in \mathbf{X} : \quad \varphi(x) = y, \quad I(x) < a+\varepsilon\} \\
&= \bigcap_{\varepsilon>0} \{\varphi(x) : \quad x \in \mathbf{X}, \quad I(x) < a+\varepsilon\} \\
&= \{\varphi(x) : \quad x \in \mathbf{X}, \quad I(x) \leq a\} \\
&= \varphi(\{I \leq a\}).
\end{aligned}$$

Since I is a good rate function, the level set $\{I \leq a\}$ is compact, and since φ is continuous, also $\{J \leq a\} = \varphi(\{I \leq a\})$ is compact. This proves that J has compact level sets.

Assume now that $\mu_n \in \mathfrak{M}(\mathbf{X})$ satisfies the LDP in $\mathbf{X}$ with rate s_n and good rate function I. If $C \subseteq \mathbf{Y}$ is closed and $O \subseteq \mathbf{Y}$ is open, $\varphi^{-1}(C) \subseteq \mathbf{X}$ is also closed and $\varphi^{-1}(O) \subseteq \mathbf{X}$ is also open, since φ is continuous. Therefore, by Theorem 2.31 and (2.32)

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{s_n} \log \mu_n(\varphi^{-1}(C)) &\leq -\inf_{x\in\varphi^{-1}(C)} I(x) = -\inf_{y\in C} J(y); \\
\liminf_{n\to\infty} \frac{1}{s_n} \log \mu_n(\varphi^{-1}(O)) &\geq -\inf_{x\in\varphi^{-1}(O)} I(x) = -\inf_{y\in O} J(y).
\end{aligned}$$

By Theorem 2.31 again, $\mu_n \circ \varphi^{-1}$ satisfies the LDP in $\mathbf{Y}$ with rate s_n and good rate function J. □

Remark 2.77. From the proof of the contraction principle, it turns out that the level sets of J are given by

$$\{J \le a\} = \varphi(\{I \le a\}).$$

This proves that J is a good rate function, since the continuous image of a compact set is compact. If the rate function I is not good, $\{I \le a\}$ is closed (since I is lower semi-continuous) but not necessarily compact. Thus, $\{J \le a\}$ is in general non-compact, and it might even be non-closed, since the continuous image of a closed set does not need to be closed. Therefore, in this case J does not need to be lower semi-continuous, hence the contraction principle may fail. ◊

The contraction principle states that, if a sequence of measures satisfies a LDP, the image measures under a continuous function also satisfy a LDP. The projective limit theorem is considered as a sort of inverse, because it states that if a *collection* of sequences of image measures (under continuous functions) satisfies a LDP, then the initial measures also do, under certain conditions.

Definition 2.78. Let $\mathbf{X}$ be a set and let $\{\mathbf{Y}_\alpha\}_{\alpha\in A}$ be a collection of sets. We say that a collection $\{\varphi_\alpha\}_{\alpha\in A}$ of functions $\varphi_\alpha\colon \mathbf{X}\to\mathbf{Y}_\alpha$ *separates points* if

$$\forall x_1, x_2 \in \mathbf{X}, \quad x_1 \neq x_2, \quad \exists \alpha \in A: \quad f_\alpha(x_1) \neq f_\alpha(x_2).$$

Theorem 2.79 (Projective limit). Let $\mathbf{X}$ be a Polish space and let $\{\mathbf{Y}_i\}_{i\in\mathbb{N}}$ be a countable collection of Polish spaces. Assume that:

(i) The sequence $\mu_n \in \mathfrak{M}(\mathbf{X})$ is exponentially tight with rate s_n.

(ii) $\{\varphi_i\}_{i\in\mathbb{N}}$ is a collection of continuous functions $\varphi_i\colon \mathbf{X}\to\mathbf{Y}_i$ that separates points.

(iii) For all $F \subseteq \mathbb{N}$ finite, there exists a good rate function J_F on $\times_{i\in F} Y_i$ such that the sequence $\mu_n \circ \varphi_F^{-1} \in \mathfrak{M}(\times_{i\in F}\mathbf{Y}_i)$ satisfies the LDP with rate s_n and rate function J_F (where $\varphi_F\colon \mathbf{X}\to\times_{i\in F}\mathbf{Y}_i$ denotes the function $\varphi_F(x) = (\varphi_i(x))_{i\in F}$).

Then, there exists a unique good rate function I on $\mathbf{X}$ such that

$$J_F(y) = \inf_{x\in\varphi_F^{-1}(y)} I(x) \qquad \forall F\subseteq\mathbb{N} \text{ finite}, \quad \forall y \in \bigtimes_{i\in F}\mathbf{Y}_i. \tag{2.33}$$

Moreover, μ_n satisfies the LDP with rate s_n and rate function I.

Proof. If $F\subseteq\mathbb{N}$ is finite, since $\mu_n\circ\varphi_F^{-1}$ satisfies the LDP with rate s_n and rate function J_F, for all $f\in C_{b,+}(\times_{i\in F}\mathbf{Y}_i)$

$$\begin{aligned}\|f\|_{\infty,J_F} &= \lim_{n\to\infty}\|f\|_{s_n,\mu_n\circ\varphi_F^{-1}} = \lim_{n\to\infty}\left(\int_{\times_{i\in F}\mathbf{Y}_i} f(y)^{s_n}\mu_n\circ\varphi_F^{-1}(dy)\right)^{1/s_n} \\ &= \lim_{n\to\infty}\left(\int_{\mathbf{X}} f(\varphi_F(x))^{s_n}\mu_n(dx)\right)^{1/s_n} = \lim_{n\to\infty}\|f\circ\varphi_F\|_{s_n,\mu_n},\end{aligned}$$

by the change of variables formula. This proves that the limit of the (s_n, μ_n)-norms of $f \circ \varphi_F$ exists: the strategy of the proof consists of showing that these functions are rate function determining, so that Proposition 2.74 can be applied. We first note that $f \circ \varphi_F : \mathbf{X} \to \mathbb{R}$ is continuous (since φ_F is continuous), bounded and nonnegative, so that

$$D \doteq \left\{ f \circ \varphi_F : \quad F \subseteq \mathbb{N}, \quad f \in C_{b,+}\left(\bigtimes_{i \in F} \mathbf{Y}_i \right) \right\} \subseteq C_{b,+}(\mathbf{X}).$$

Let us prove that D is rate function determining. Let $\overline{x} \in \mathbf{X}$. For all $k \geq 1$ we define $f_k : \bigtimes_{i=1}^k \mathbf{Y}_i \to \mathbb{R}$ by

$$f_k(y_1, \dots, y_k) \doteq 0 \vee \left(1 - k \max_{i=1,\dots,k} d_i(\varphi_i(\overline{x}), y_i) \right),$$

where d_i is any metric generating the topology on $\mathbf{Y}_i$. Obviously we have that $f_k \in C_{b,+}(\bigtimes_{i=1}^k \mathbf{Y}_i)$, hence $f_k \circ \varphi_{F_k} \in D$, where $F_k \doteq \{1, \dots, k\}$. It turns out that

$$f_k \circ \varphi_{F_k}(x) = 0 \vee \left(1 - k \max_{i=1,\dots,k} d_i(\varphi_i(\overline{x}), \varphi_i(x)) \right),$$

hence

- The functions $f_k \circ \varphi_{F_k}$ are decreasing.
- $f_k \circ \varphi_{F_k}(\overline{x}) = 1$ for all $k \geq 1$, hence $f_k \circ \varphi_{F_k}(\overline{x}) \xrightarrow{k \to \infty} 1$.
- If $x \neq \overline{x}$, since $\{\varphi_i\}_{i \in \mathbb{N}}$ separates points, there exists $j \in \mathbb{N}$ such that $\varphi_j(x) \neq \varphi_j(\overline{x})$, hence $d_i(\varphi_j(\overline{x}), \varphi_j(x)) = c > 0$. It follows that $\forall k \geq j$

$$0 \leq f_k \circ \varphi_{F_k}(x) = 0 \vee \left(1 - k \max_{i=1,\dots,k} d_i(\varphi_i(\overline{x}), \varphi_i(x)) \right) \leq 0 \vee (1 - kc) \xrightarrow{k \to \infty} 0,$$

  hence $\forall x \neq \overline{x}$ $f_k \circ \varphi_{F_k}(x) \xrightarrow{k \to \infty} 0$.

This proves that $f_k \circ \varphi_{F_k} \searrow \mathbb{1}_{\{\overline{x}\}}$. By Lemma 2.38, D is rate function determining. Therefore, we have:

(i) μ_n is exponentially tight with rate s_n.

(ii) The limit $\lim_{n \to \infty} \|f \circ \varphi_F\|_{s_n, \mu_n} = \|f\|_{\infty, J_F}$ exists for all $f \circ \varphi_F \in D$, and $D \subseteq C_{b,+}(\mathbf{X})$ is rate function determining.

By Proposition 2.74, there exists a unique good rate function I such that

$$\|f\|_{\infty, J_F} = \|f \circ \varphi_F\|_{\infty, I} \qquad \forall F \subseteq \mathbb{N} \text{ finite}, \quad \forall f \in C_{b,+}\left(\bigtimes_{i \in F} \mathbf{Y}_i \right), \tag{2.34}$$

and μ_n satisfies the LDP with rate s_n and rate function I. For any $F \subseteq \mathbb{N}$ finite, by the contraction principle (which can be applied because the function $\varphi_F : \mathbf{X} \to \bigtimes_{i\in F} \mathbf{Y}_i$ is continuous), $\mu_n \circ \varphi_F^{-1}$ satisfies the LDP with rate s_n and good rate function

$$J'_F(y) = \inf_{x\in\varphi_F^{-1}(y)} I(x) \qquad\qquad \forall y \in \bigtimes_{i\in F} \mathbf{Y}_i.$$

By the uniqueness of the rate function (Proposition 2.39), $J'_F = J_F$ and (2.33) holds. It only remains to prove that there exists a *unique* good rate function satisfying (2.33): if I satisfies (2.33), for any $F \subseteq \mathbb{N}$ finite and $f \in C_{b,+}(\bigtimes_{i\in F} \mathbf{Y}_i)$ we have

$$\begin{aligned}
\|f\|_{\infty,J_F} &= \sup_{y\in\bigtimes_{i\in F}\mathbf{Y}_i} \exp(-J_F(y))f(y) = \sup_{y\in\bigtimes_{i\in F}\mathbf{Y}_i} \exp\Big(-\inf_{x\in\varphi_F^{-1}(y)} I(x)\Big)f(y)\\
&= \sup_{y\in\bigtimes_{i\in F}\mathbf{Y}_i} \sup_{x\in\varphi_F^{-1}(y)} \exp(-I(x))f(\varphi_F(x)) = \sup_{x\in\mathbf{X}} \exp(-I(x))f(\varphi_F(x))\\
&= \|f\circ\varphi_F\|_{\infty,I}.
\end{aligned}$$

Therefore (2.34) holds, and we have already remarked, by Proposition 2.74, that this identifies uniquely I. □

Chapter 3

Empirical averages and distributions

The most important cases of large deviations principles concern sequences of i.i.d. random variables. Let $(\Omega, \mathcal{E}, \mathbb{P})$ denote a probability space. Let $\mathbf{X}$ be a Polish space and let $\mathfrak{M}(\mathbf{X})$ be the space of probability measures on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$, equipped with the weak topology (that is Polish again, as we stated in Example 2.19). Let

$$X_i \colon (\Omega, \mathcal{E}, \mathbb{P}) \to (\mathbf{X}, \mathcal{B}(\mathbf{X})), \qquad i \geq 1,$$

be a sequence of i.i.d. $\mathbf{X}$-valued random variables. It turns out that:

(i) If $\mathbf{X} = \mathbb{R}^d$ for any $d \geq 1$, the laws of the *empirical averages*

$$\frac{S_n}{n} \doteq \frac{1}{n} \sum_{i=1}^{n} X_i$$

of the $\{X_i\}_{i \geq 1}$ satisfy a LDP in $\mathfrak{M}(\mathbf{X})$. We will analyze the case $\mathbf{X} = \mathbb{R}$ in section 3.1 (generalized Cramér's Theorem in $\mathbb{R}$) and we will allude to the case $\mathbf{X} = \mathbb{R}^d$ in section 3.2 (Cramér's Theorem in $\mathbb{R}^d$). This first part is mainly based on [Swa, §2.3].

(ii) For a generic Polish space $\mathbf{X}$, the laws of the *empirical distributions*

$$L_n \doteq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

of the $\{X_i\}_{i \geq 1}$ satisfy a LDP in $\mathfrak{M}(\mathfrak{M}(\mathbf{X}))$. This result is known as Sanov's Theorem and it will be analyzed in section 3.4. The properties of the corresponding good rate function, which is called *relative entropy*, will be studied in section 3.3. This second part is mainly based on [Swa, §2.4].

3.1 Generalized Cramér's Theorem in $\mathbb{R}$

We recall the notation of chapter 1:

- $(\Omega, \mathcal{E}, \mathbb{P})$ is a probability space.
- X is a real-valued random variable defined on $(\Omega, \mathcal{E}, \mathbb{P})$.
- $M(t) \doteq \mathbb{E}(e^{tX}) \in (0, +\infty]$ is the moment generating function of X.
- If t is such that $M(t) < +\infty$, $\mathbb{P}_t$ is the probability measure on $(\Omega, \mathcal{E})$, absolutely continuous with respect to $\mathbb{P}$, defined by the Radon-Nikodym derivative

$$\frac{d\mathbb{P}_t}{d\mathbb{P}} = \frac{e^{tX}}{M(t)}.$$

  Moreover, $\mathbb{E}_t$ denotes the expectation with respect to $\mathbb{P}_t$.

The version of Cramér's Theorem we saw in chapter 1 is a particular case of the following more general theorem. We will prove it by using Theorem 1.20, because an "independent" proof would be very similar to the proof of Cramér's Theorem itself.

Theorem 3.1 (Generalized Cramér's Theorem)**.** Let X be a real-valued random variable such that its logarithmic moment generating function $\log M$ is finite in a neighborhood of 0 and steep. Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. copies of X. Then the laws μ_n of the empirical averages

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

$$I(x) \doteq (\log M)^*(x) = \sup_{t\in\mathbb{R}}[tx - \log M(t)] \qquad \forall x \in \mathbb{R}.$$

Proof. By Theorem 1.19 (iii), (viii) and (x), I is a good rate function. Thus, by Theorem 2.31, it suffices to prove that

$$\limsup_{n\to\infty} \frac{1}{n}\log \mu_n(C) \leq -\inf_{x\in C} I(x) \qquad \forall C \subseteq \mathbb{R} \text{ closed}, \tag{3.1}$$

$$\liminf_{n\to\infty} \frac{1}{n}\log \mu_n(O) \geq -\inf_{x\in O} I(x) \qquad \forall O \subseteq \mathbb{R} \text{ open}. \tag{3.2}$$

The case that $X = \overline{x}$ almost surely is obvious, since for all $A \in \mathcal{B}(\mathbb{R})$ $\mu_n(A) = \mathbb{1}_A(\overline{x})$ and

$$I(x) = \begin{cases} 0 & x = \overline{x}, \\ +\infty & x \neq \overline{x}. \end{cases}$$

Therefore, we may suppose that X is not almost surely constant.

We start by proving the first inequality. Let $C \subseteq \mathbf{X}$ be closed. Assume first that $\overline{x} \in C$. By Theorem 1.19 (x) $I(\overline{x}) = 0$, hence

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(C) \le \limsup_{n\to\infty} \frac{1}{n} \log 1 = 0 = -I(\overline{x}) \le -\inf_{x\in C} I(x).$$

Assume now that $\overline{x} \notin C$, and let (a,b) the connected component of C^c that contains $\overline{x}$, so that $C \subseteq (-\infty, a] \cup [b, +\infty)$ and $a, b \in C$. By recalling Remark 1.24 about two-sided deviations from $\frac{S_n}{n}$, we obtain

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(C) &\le \limsup_{n\to\infty} \frac{1}{n} \log\left(\mathbb{P}\left(\frac{S_n}{n} \le a\right) + \mathbb{P}\left(\frac{S_n}{n} \ge b\right)\right)\\ &= -(I(a) \wedge I(b)) = -\inf_{x\in C} I(x),\end{aligned}$$

where the latter equality follows from the fact that I is decreasing in $(-\infty, \overline{x})$ and increasing in $(\overline{x}, +\infty)$ by Theorem 1.19 (ix). This proves (3.1).

To prove (3.2), let $O \subseteq \mathbb{R}$ be open. For all $x \in O$, there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq O$, hence

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(O) \ge \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\varepsilon(x)).$$

If we prove that $\forall x \in \mathbb{R}$ and $\forall \varepsilon > 0$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\varepsilon(x)) \ge -I(x), \tag{3.3}$$

it will follow that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(O) \ge \sup_{x\in O}(-I(x)) = -\inf_{x\in O} I(x),$$

i.e. (3.2). To prove (3.3), we fix $x \in \mathbb{R}$ and $\varepsilon > 0$. If $x = \overline{x}$,

$$\mu_n(B_\varepsilon(\overline{x})) = \mathbb{P}\left(\left|\frac{S_n}{n} - \overline{x}\right| < \varepsilon\right) \xrightarrow{n\to\infty} 1$$

by the weak law of large numbers, hence

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(B_\varepsilon(\overline{x})) = 0 = -I(\overline{x}).$$

Let now $\mathring{\mathcal{D}}_I = (x_-, x_+)$, with

$$x_- \doteq \inf(\operatorname{supp} X), \qquad x_+ \doteq \sup(\operatorname{supp} X),$$

by Theorem 1.19 (xiii). If $x \in (\overline{x}, x_+)$, there exists $\delta \in (0,\varepsilon)$ such that $x+\delta \in (\overline{x}, x_+)$, and by applying (1.16) to x and $x+\delta$ we obtain

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n}\log \mu_n(B_\varepsilon(x)) &\geq \liminf_{n\to\infty} \frac{1}{n}\log \mu_n([x, x+\delta)) \\
&= \liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}\Big(\frac{S_n}{n} \in [x, x+\delta)\Big) \\
&= \liminf_{n\to\infty} \frac{1}{n}\log \Big(\mathbb{P}\Big(\frac{S_n}{n} \geq x\Big) - \mathbb{P}\Big(\frac{S_n}{n} \geq x+\delta\Big)\Big) \\
&= \liminf_{n\to\infty} \frac{1}{n}\log \big(e^{-nI(x)+o(n)} - e^{-nI(x+\delta)+o(n)}\big) \\
&= \liminf_{n\to\infty} \frac{1}{n}\log \big(e^{-nI(x)+o(n)}(1 - e^{-n(I(x+\delta)-I(x))+o(n)})\big) \\
&= \liminf_{n\to\infty} \Big(\frac{-nI(x)+o(n)}{n} + \frac{\log(1 - e^{-n(I(x+\delta)-I(x))+o(n)})}{n}\Big) \\
&= -I(x),
\end{aligned}$$

since $I(x+\delta) - I(x) > 0$ (I is strictly increasing in $(\overline{x}, x_+)$). If $x = x_+$, by applying (1.13) of Cramér's Theorem to x_+ we obtain

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n}\log \mu_n(B_\varepsilon(x_+)) &\geq \liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}\Big(\frac{S_n}{n} = x_+\Big) \\
&= \liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}\Big(\frac{S_n}{n} \geq x_+\Big) = -I(x_+). \qquad \square
\end{aligned}$$

If $x \in (x_+, +\infty)$, $I(x) = +\infty$ and (3.3) does not need to be proved. In a similar way one can prove (3.3) when $x \in (-\infty, \overline{x})$.

Remark 3.2. We have just derived the generalized Cramér's Theorem 3.1 from Cramér's Theorem 1.20, but it is possible to do the opposite. Indeed, item (iii) of the proof of Theorem 1.20, that is its actual core, is extremely simplified by applying Theorem 3.1: let $x \in (\overline{x}, x_+)$. Since $[x, +\infty)$ is closed and $(x, +\infty)$ is open,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n}\log \mu_n([x, +\infty)) &\leq -\inf_{[x,+\infty)} I \\
\liminf_{n\to\infty} \frac{1}{n}\log \mu_n((x, +\infty)) &\geq -\inf_{(x,+\infty)} I.
\end{aligned}$$

Since I is continuous on (x_-, x_+) and increasing on $[\overline{x}, +\infty)$, it turns out that

$$\inf_{(x,+\infty)} I = \inf_{[x,+\infty)} I = I(x).$$

Therefore

$$\begin{aligned}
-I(x) = -\inf_{(x,+\infty)} I &\leq \liminf_{n\to\infty} \frac{1}{n}\log \mu_n((x, +\infty)) \leq \liminf_{n\to\infty} \frac{1}{n}\log \mu_n([x, +\infty)) \\
&\leq \limsup_{n\to\infty} \frac{1}{n}\log \mu_n([x, +\infty)) \leq -\inf_{[x,+\infty)} I = -I(x),
\end{aligned}$$

which proves that

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\frac{S_n}{n}\geq x\Big)=\lim_{n\to\infty}\frac{1}{n}\log\mu_n([x,+\infty))=-I(x). \qquad \lozenge$$

Remark 3.3. In generalized Cramér's Theorem, the large deviations upper and lower bounds remain true even without the hypothesis that $\log M$ is steep and finite in a neighborhood of 0, although in these cases I loses some properties, as we already observed in Remark 1.21. In particular, it is important to stress that, if M is not finite in a neighborhood of 0, then I is not a *good* rate function, because it no longer has compact level sets. ◊

3.2 Cramér's Theorem in $\mathbb{R}^d$

In this section we state Cramér's Theorem in $\mathbb{R}^d$, whose proof is more complicated but somewhat similar to the case $d = 1$, and we see the properties of the associated good rate function. We refer to [Swa, §2.3] for details and proofs; however, one can identify the analogous steps in the path we followed to prove the properties of the rate function and Cramér's Theorem in $\mathbb{R}$ in chapter 1.

We start with the generalization of the concept of Legendre transform for functions defined on $\mathbb{R}^d$.

Definition 3.4. The *Legendre transform* of a function $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is the function $f^*\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ defined by

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

Lemma 3.5. The Legendre transform f^* of any function $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$, $f \not\equiv +\infty$, is convex and lower semi-continuous.

We recall that $\mathcal{D}_f$ denotes the domain of a $\overline{\mathbb{R}}$-valued function f, i.e. the set where f is finite. If $f\colon \mathbb{R}^d \to (-\infty,+\infty]$ is convex, then $\mathcal{D}_f$ and $\mathring{\mathcal{D}}_f$ are convex subsets of $\mathbb{R}^d$ by Remark 1.9.

Lemma 3.6. Let $f\colon \mathbb{R}^d \to (-\infty,+\infty]$ be convex and lower semi-continuous. Then f is continuous on $\overline{\mathcal{D}}_f$.

Definition 3.7. A function $f\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is said to be *steep* if f is differentiable on $\mathring{\mathcal{D}}_f$ and for all $t \in \partial\mathring{\mathcal{D}}_f$ and any sequence $\{t_k\}_{k\geq 1} \subseteq \mathring{\mathcal{D}}_f$ such that $t_k \to t$,

$$\lim_{k\to\infty}|\nabla f(t_k)| = +\infty.$$

Definition 3.8. We denote by $\mathrm{Conv}_\infty(\mathbb{R}^d)$ the space of the functions $f\colon \mathbb{R}^d \to (-\infty,+\infty]$ such that

(i) $\mathring{\mathcal{D}}_f \neq \varnothing$.

(ii) f is lower semi-continuous.

(iii) $f \in C^\infty(\mathring{\mathcal{D}}_f)$.

(iv) f is convex and for all $t \in \mathring{\mathcal{D}}_f$ the hessian matrix $D^2 f(t)$ is positive definite, i.e. $\langle x, D^2 f(t) \cdot x \rangle > 0$ for any $x \in \mathbb{R}^d \setminus \{0\}$.

(v) f is steep.

Proposition 3.9. If $f \in \mathrm{Conv}_\infty(\mathbb{R}^d)$, then:

(i) $f^*(x) = \sup_{t \in \mathring{\mathcal{D}}_f} [\langle t, x \rangle - f(t)] \qquad \forall x \in \mathbb{R}$.

(ii) $\nabla f : \mathring{\mathcal{D}}_f \to \mathring{\mathcal{D}}_{f^*}$ is a bijection and

$$(\nabla f)^{-1} = \nabla f^*.$$

Moreover, for any $x \in \mathring{\mathcal{D}}_{f^*}$ the supremum in the definition of f^* is attained at $t = (\nabla f)^{-1}(x)$, that is

$$f^*(x) = \langle (\nabla f)^{-1}(x), x \rangle - f\Big((\nabla f)^{-1}(x)\Big) \qquad \forall x \in \mathring{\mathcal{D}}_{f^*}. \tag{3.4}$$

(iii) $f^* \in \mathrm{Conv}_\infty$.

(iv) $f^{**} = f$.

From now on, we let $X = (X_1, \dots, X_d)$ denote a fixed multivariate random variable defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ and taking values in $\mathbb{R}^d$, with mean vector $\mathbb{E}(X)$ and covariance matrix $\mathrm{Cov}(X)$. Similarly to the 1-dimensional case, we define the moment generating function of X and we study its properties.

Definition 3.10. We call *moment generating function* of the random variable X the function $M : \mathbb{R}^d \to (0, +\infty]$ given by

$$M(t) \doteq \mathbb{E}\Big(e^{\langle t, X \rangle}\Big).$$

We call *logarithmic moment generating function* of X the function $\log M$, taking values in $(-\infty, +\infty]$.

If $M(t) < +\infty$, we can define a new probability measure $\mathbb{P}_t$ on $(\Omega, \mathcal{E})$, absolutely continuous with respect to $\mathbb{P}$, by the following Radon-Nikodym derivative:

$$\frac{d\mathbb{P}_t}{d\mathbb{P}} = \frac{e^{\langle t, X \rangle}}{M(t)}.$$

We let $\mathbb{E}_t$ and Cov_t denote mean vector and covariance matrix respectively, with respect to $\mathbb{P}_t$.

Theorem 3.11. M is analytic (hence, C^∞) on $\mathring{\mathcal{D}}_M$, with derivatives

$$\frac{\partial^{\alpha_1+\dots+\alpha_d} M}{\partial t_1^{\alpha_1}\dots\partial t_d^{\alpha_d}}(t) = M(t)\mathbb{E}_t(X_1^{\alpha_1}\dots X_d^{\alpha_d}) = \mathbb{E}\Big(X_1^{\alpha_1}\dots X_d^{\alpha_d} e^{\langle t,X\rangle}\Big)$$

for all $t \in \mathring{\mathcal{D}}_M$, $\alpha_1,\dots,\alpha_d \in \mathbb{N}_0$. In particular, if M is finite in a neighborhood of 0, then $X \in L^p(\Omega,\mathcal{E},\mathbb{P})$ for all $p \in [1,+\infty)$ and its moments are given by

$$\mathbb{E}(X_1^{\alpha_1}\dots X_d^{\alpha_d}) = \frac{\partial^{\alpha_1+\dots+\alpha_d} M}{\partial t_1^{\alpha_1}\dots\partial t_d^{\alpha_d}}(0) \qquad \forall \alpha_1,\dots\alpha_d \in \mathbb{N}_0.$$

Lemma 3.12. The function $\log M$ satisfies:

(i) For all $t \in \mathring{\mathcal{D}}_M$

$$\nabla \log M(t) = \mathbb{E}_t(X), \qquad D^2 \log M(t) = \mathrm{Cov}_t(X).$$

(ii) If $\mathcal{D}_M = \mathbb{R}^d$ (i.e. M is finite everywhere) and $\mathrm{Cov}(X)$ is positive definite, then $\log M \in \mathrm{Conv}_\infty$.

The next theorem summarizes the main properties of the Legendre transform of the logarithmic moment generating function, defined by

$$I(x) \doteq (\log M)^*(x) = \sup_{t\in\mathbb{R}^d}[\langle t,x\rangle - \log M(t)] \qquad \forall x \in \mathbb{R}^d.$$

We recall that the *convex hull* of a set $A \subseteq \mathbb{R}^d$ is the intersection of all convex sets containing A, and the closed convex hull is its closure. We proved in Theorem 1.19 (xiii) that, for a real-valued random variable X, $\mathring{\mathcal{D}}_I = (\inf(\operatorname{supp} X), \sup(\operatorname{supp} X))$; we may reformulate this by saying that the $\overline{\mathcal{D}}_I$ is the closed convex hull of $\operatorname{supp}(X)^\dagger$. The latter statement also holds for a multivariate random variable.

Theorem 3.13. Let X be a multivariate random variable such that $M(t)$ is finite for all $t \in \mathbb{R}^d$ (in particular, $X \in L^p(\Omega,\mathcal{E},\mathbb{P})$ for all $p \in [1,+\infty)$) and $\mathrm{Cov}(X)$ is a positive definite matrix. Then the Legendre transform I of $\log M$ satisfies:

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

(ii) $I \geq 0$ and $I(x) = 0$ if and only if $x = \mathbb{E}(X)$.

†In general, the closed convex hull of a set $A \subseteq \mathbb{R}$ is the closure of $(\inf A, \sup A)$.

(iii) I is a good rate function.

(iv) $\overline{\mathcal{D}}_I$ is the closed convex hull of $\operatorname{supp}(X)$.

Similarly to the case $d = 1$, I turns out to be the good rate function for the LDP satisfied by the laws of the empirical averages of a sequence of i.i.d. copies of X.

Theorem 3.14 (Cramér's Theorem in $\mathbb{R}^d$). Let X be a multivariate random variable such that its moment generating function M is finite everywhere. Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. copies of X. Then the laws of the empirical averages

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

$$I(x) \doteq (\log M)^*(x) = \sup_{t\in\mathbb{R}^d}[\langle t,x\rangle - \log M(t)] \qquad \forall x \in \mathbb{R}^d.$$

3.3 Relative entropy

Let $\mathbf{X}$ be a Polish space and let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. $\mathbf{X}$-valued random variables. As we announced in the introduction of this chapter, the laws of the empirical distributions of the $\{X_i\}_{i\geq 1}$ satisfy a LDP in $\mathfrak{M}(\mathfrak{M}(\mathbf{X}))$: it will turn out that the corresponding good rate function $\mathfrak{M}(\mathbf{X}) \to [0,+\infty]$ is the so called *relative entropy*, which is the subject of this section.

We start by recalling some concepts and results, which will be often useful from now on, about the space $\mathfrak{M}(\mathbf{X})$ of probability measures on the Polish space $\mathbf{X}$.

We first note that $\mathfrak{M}(\mathbf{X})$ is *convex*: for any $\mu,\nu \in \mathfrak{M}(\mathbf{X})$ and $p \in (0,1)$, the function

$$\mathcal{B}(\mathbf{X}) \to [0,1], \qquad A \to p\mu(A) + (1-p)\nu(A) \qquad \forall A \in \mathcal{B}(\mathbf{X})$$

is a probability measure on $(\mathbf{X},\mathcal{B}(\mathbf{X}))$ again, i.e. $p\mu + (1-p)\nu \in \mathfrak{M}(\mathbf{X})$.

Remark 3.15 (Radon-Nikodym Theorem). If $\mu,\nu \in \mathfrak{M}(\mathbf{X})$, ν is said to be *absolutely continuous* with respect to μ if $\nu(A) = 0$ for all $A \in \mathcal{B}(\mathbf{X})$ such that $\mu(A) = 0$; in this case, we write $\nu \ll \mu$. By the Radon-Nikodym Theorem, $\nu \ll \mu$ if and only if ν has a density with respect to μ, i.e. if and only if there exists a measurable function $f : (\mathbf{X},\mathcal{B}(\mathbf{X})) \to [0,+\infty)$, satisfying $\int_{\mathbf{X}} f\,d\mu = 1$, such that

$$\nu(A) = \int_A f\,d\mu \qquad \forall A \in \mathcal{B}(\mathbf{X}).$$

Moreover, f is μ-a.s. unique, i.e. if g satisfies the same properties then μ-a.s. $f = g$. The density f of ν with respect to μ, which we call *Radon-Nikodym derivative* and we denote by $f = \frac{d\nu}{d\mu}$ (we will often also write $d\nu = f\,d\mu$), satisfies the following simple properties:

(i) If $\mu, \nu \in \mathfrak{M}(\mathbf{X})$ with $\nu \ll \mu$ and $g \colon \mathbf{X} \to \mathbb{R}$ is a measurable function, then the integral of g with respect to ν is well-defined (possibly infinite) if and only if the integral of $g\frac{d\nu}{d\mu}$ with respect to μ is so. In this case, the two integrals are equal:

$$\int_{\mathbf{X}} g\,d\nu = \int_{\mathbf{X}} g\frac{d\nu}{d\mu}\,d\mu.$$

In particular, the equality holds if g is bounded from above or bounded from below.

(ii) If $\mu, \nu, \lambda \in \mathfrak{M}(\mathbf{X})$ with $\nu \ll \mu$ and $\lambda \ll \mu$ and $p \in (0,1)$, then $p\nu + (1-p)\lambda \ll \mu$ and

$$\frac{d(p\nu + (1-p)\lambda)}{d\mu} = p\frac{d\nu}{d\mu} + (1-p)\frac{d\lambda}{d\mu} \qquad \mu\text{ -a.s..}$$

(iii) If $\mu, \nu, \lambda \in \mathfrak{M}(\mathbf{X})$ with $\nu \ll \mu$ and $\lambda \ll \nu$, then $\lambda \ll \mu$ and

$$\frac{d\lambda}{d\mu} = \frac{d\lambda}{d\nu} \cdot \frac{d\nu}{d\mu} \qquad \mu\text{ -a.s..}$$

◊

Remark 3.16 (Lebesgue decomposition theorem). Let $\mu, \nu \in \mathfrak{M}(\mathbf{X})$. If $\nu \ll \mu$, by the Radon-Nikodym Theorem we can write $d\nu = f\,d\mu$, where f is the density of ν with respect to μ. However, even in case that $\nu \not\ll \mu$, by the Lebesgue decomposition theorem (see [Kle, Th.7.33]) ν can be expressed as

$$d\nu = f\,d\mu + \mathbb{1}_A\,d\nu,$$

where $f : (\mathbf{X}, \mathcal{B}(\mathbf{X})) \to [0, +\infty)$ is measurable and $A \in \mathcal{B}(\mathbf{X})$ is such that $\mu(A) = 0$. In other words,

$$\nu(B) = \int_B f\,d\mu + \int_B \mathbb{1}_A\,d\nu, \qquad \forall B \in \mathcal{B}(\mathbf{X}).$$

◊

Definition 3.17. Let $\mu \in \mathfrak{M}(\mathbf{X})$. We call *relative entropy* with respect to μ the function $H(\cdot|\mu) \colon \mathfrak{M}(\mathbf{X}) \to \overline{\mathbb{R}}$ defined by

$$H(\nu|\mu) \doteq \begin{cases} \displaystyle\int_{\mathbf{X}} \log\left(\frac{d\nu}{d\mu}\right) d\nu = \int_{\mathbf{X}} \frac{d\nu}{d\mu}\log\left(\frac{d\nu}{d\mu}\right) d\mu & \text{if } \nu \ll \mu, \\ +\infty & \text{otherwise} \end{cases}$$

for all $\nu \in \mathfrak{M}(\mathbf{X})$.

Remark 3.18. First of all, we show that $H(\cdot|\mu)$ is well-defined. If $\nu \ll \mu$, the first integral in the definition of $H(\nu|\mu)$ is well-defined if and only if the second one is so, and in this case they are actually equal; moreover, the second integral is

$$\int_{\mathbf{X}} \frac{d\nu}{d\mu} \log\Big(\frac{d\nu}{d\mu}\Big) d\mu = \int_{\mathbf{X}} \psi\Big(\frac{d\nu}{d\mu}\Big) d\mu,$$

where $\psi\colon [0,+\infty) \to \mathbb{R}$ is defined by

$$\psi(t) \doteq \begin{cases} t\log t & t > 0, \\ 0 & t = 0. \end{cases} \tag{3.5}$$

Since $\psi(t) \to 0$ as $t \to 0$ and $\psi(t) \to +\infty$ as $t \to +\infty$, ψ is continuous and bounded from below. This proves that the negative part of $\psi\big(\frac{d\nu}{d\mu}\big)$ is μ-integrable, hence the integral is well-defined. We conclude that $H(\cdot|\mu)$ is well-defined and takes values in $(-\infty,+\infty]$. ♢

The following proposition deals with the main properties of the relative entropy.

Proposition 3.19. Let $\mu \in \mathfrak{M}(\mathbf{X})$. The relative entropy $H(\cdot|\mu)$ satisfies the following properties:

(i) $H(\cdot|\mu) \geq 0$ and $H(\nu|\mu) = 0$ if and only if $\nu = \mu$.

(ii) $H(\cdot|\mu)$ is convex.

(iii) $H(\cdot|\mu)$ is a good rate function.

The proof that $H(\cdot|\mu)$ has compact level sets is based on functional analysis arguments: in particular, some knowledge about weak topologies of Banach spaces is required. We recall here the concepts and the results we need to use:

- Let $(V, \|\cdot\|)$ be a Banach space. The dual V^* of V is the Banach space

  $$V^* \doteq \{\alpha\colon V \to \mathbb{R}, \quad \alpha \text{ linear and continuous}\},$$

  equipped with the norm

  $$\|\alpha\|_{V^*} \doteq \sup_{v \in V\setminus\{0\}} \frac{|\alpha(v)|}{\|v\|}.$$

  We can equip V with another topology, the *weak topology*, i.e. the coarsest topology on V such that each functional $\alpha \in V^*$ is continuous. Since each $\alpha \in V^*$ is continuous for the norm topology on V by definition, the weak topology is actually coarser than the norm topology (i.e. each weakly open subset of V is also norm-open); it turns out that the weak topology is strictly coarser if and only if V is infinite dimensional.

- Let C be a *convex* subset of the Banach space V. Then C is weakly closed if and only if it is norm-closed (see [Bre, Th.3.7]).

- If $\mu \in \mathfrak{M}(\mathbf{X})$, the dual space of $L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ is isometric to $L^\infty(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$: indeed, the map

$$L^\infty(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu) \to L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)^*, \qquad g \to \alpha_g$$

is a bijective linear isometry of Banach spaces, if for any $g \in L^\infty(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ $\alpha_g \colon L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu) \to \mathbb{R}$ is the linear form defined by

$$\alpha_g(f) = \int_{\mathbf{X}} f g \, d\mu \qquad \forall f \in L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$$

(see [Bre, Th.4.14]).

- We will also use a criterion of weak pre-compactness in $L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$, which is known as Dunford-Pettis Theorem (see [Ale, Th.1.3]): $C \subseteq L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ is weakly pre-compact if and only if C is uniformly integrable, i.e.

$$\forall \varepsilon > 0 \qquad \exists c \in \mathbb{R} : \qquad \int_{\mathbf{X}} |f| \mathbb{1}_{\{|f| \geq c\}} \, d\mu < \varepsilon \qquad \forall f \in C.$$

Proof of Proposition 3.19. In the following, ψ will denote the function defined by (3.5), which is continuous on $[0, +\infty)$ and bounded from below, as we observed in Remark 3.18.

(i) For all $t > 0$

$$\psi'(t) = \log t + 1, \qquad \psi''(t) = \frac{1}{t},$$

hence ψ is strictly convex on $[0, +\infty)$. By Jensen's inequality,

$$H(\nu|\mu) = \int_{\mathbf{X}} \psi\Big(\frac{d\nu}{d\mu}\Big) d\mu \geq \psi\Big(\int_{\mathbf{X}} \frac{d\nu}{d\mu} \, d\mu\Big) = \psi(1) = 1 \log 1 = 0,$$

and the equality holds if and only if $\frac{d\nu}{d\mu}$ is μ-a.s. constant; in this case, since ν and μ are probability measures, such a constant is 1, i.e. $\nu = \mu$. This proves that $H(\nu|\mu) \geq 0$ for all $\nu \in \mathfrak{M}(\mathbf{X})$, with equality if and only if $\nu = \mu$.

(ii) Since $\mathfrak{M}(\mathbf{X})$ is a convex set, it makes sense to wonder if $H(\cdot|\mu)$ is a convex function. Let $\nu_1, \nu_2 \in \mathfrak{M}(\mathbf{X})$ and $p \in (0,1)$. We have to show that

$$H(p\nu_1 + (1-p)\nu_2|\mu) \leq pH(\nu_1|\mu) + (1-p)H(\nu_2|\mu).$$

If $\nu_1 \not\ll \mu$ or $\nu_2 \not\ll \mu$, then $H(\nu_1|\mu) = +\infty$ or $H(\nu_2|\mu) = +\infty$, hence the inequality holds because the right-hand side is $+\infty$. If $\nu_1 \ll \mu$ and $\nu_2 \ll \mu$, then

$$\begin{aligned} H(p\nu_1 + (1-p)\nu_2|\mu) &= \int_{\mathbf{X}} \psi\Big(\frac{d(p\nu_1 + (1-p)\nu_2)}{d\mu}\Big) d\mu \\ &= \int_{\mathbf{X}} \psi\Big(p\frac{d\nu_1}{d\mu} + (1-p)\frac{d\nu_2}{d\mu}\Big) d\mu \\ &\le \int_{\mathbf{X}} \Big(p\psi\Big(\frac{d\nu_1}{d\mu}\Big) + (1-p)\psi\Big(\frac{d\nu_2}{d\mu}\Big)\Big) d\mu \\ &= pH(\nu_1|\mu) + (1-p)H(\nu_2|\mu), \end{aligned}$$

by the convexity of ψ, and the inequality holds again.

(iii) By (i), $H(\cdot|\mu) \ge 0$ and $H(\cdot|\mu) \not\equiv +\infty$. Thus, we only have to prove that $H(\cdot|\mu)$ has compact level sets. Fix $a \in \mathbb{R}$, and let

$$L \doteq \{\nu \in \mathfrak{M}(\mathbf{X}) : \quad H(\nu|\mu) \le a\}$$

be an arbitrary level set. By definition, the weak topology on $\mathfrak{M}(\mathbf{X})$ is the coarsest topology that makes continuous all maps $\varphi_g \colon \mathfrak{M}(\mathbf{X}) \to \mathbb{R}$ defined by $\varphi_g(\nu) \doteq \int_{\mathbf{X}} g\, d\nu$, where $g \in C_{b,+}(\mathbf{X})$. Therefore, the weak topology that $\mathfrak{M}(\mathbf{X})$ induces on L is the coarsest topology that makes continuous all maps $\varphi_g|_L \colon L \to \mathbb{R}$, where $g \in C_{b,+}(\mathbf{X})$. For any $\nu \in L$, $H(\nu|\mu) < +\infty$, hence $\nu \ll \mu$ and

$$\varphi_g|_L(\nu) = \int_{\mathbf{X}} g \frac{d\nu}{d\mu}\, d\mu.$$

We note that there is a bijection

$$L \longleftrightarrow C \doteq \Big\{ f \in L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu) : \quad f \ge 0 \ \mu\text{-a.s.}, \quad \|f\|_{1,\mu} = 1, \quad \int_{\mathbf{X}} \psi \circ f\, d\mu \le a \Big\}$$

defined by identifying a measure with its density:

- If $\nu \in L$, the corresponding $f \in C$ is $f = \frac{d\nu}{d\mu}$.
- If $f \in C$, the corresponding $\nu \in \mathfrak{M}(\mathbf{X})$ is the probability measure ν that is absolutely continuous with respect to μ with density f, so that $f = \frac{d\nu}{d\mu}$.

By the identification between L and C, $\mathfrak{M}(\mathbf{X})$ induces on C the coarsest topology that makes continuous all maps

$$C \to \mathbb{R}, \qquad f \to \int_{\mathbf{X}} f g\, d\mu, \qquad g \in C_{b,+}(\mathbf{X}).$$

We wish to prove that C is compact in this topology. The dual of $L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$ is composed of all maps

$$L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu) \to \mathbb{R}, \qquad f \to \int_{\mathbf{X}} f g\, d\mu, \qquad g \in L^\infty(\mathbf{X}, \mathcal{B}(\mathbf{X}), \mu)$$

and $C_{b,+}(\mathbf{X}) \subseteq L^\infty(\mathbf{X},\mathcal{B}(\mathbf{X}),\mu)$, hence the weak topology of $L^1(\mathbf{X},\mathcal{B}(\mathbf{X}),\mu)$ is stronger than the weak topology of $\mathfrak{M}(\mathbf{X})$, if they are induced on $C \leftrightarrow L$. It follows that it suffices[†] (although it is not necessary) to prove that C is compact in the weak topology of $L^1(\mathbf{X},\mathcal{B}(\mathbf{X}),\mu)$. In fact, we will prove that C is closed and pre-compact in the weak topology of $L^1(\mathbf{X},\mathcal{B}(\mathbf{X}),\mu)$:

- To prove that C is weakly closed, it suffices to show that it is convex and norm-closed. Let $f_1, f_2 \in C$ and $p \in (0,1)$: then $pf_1 + (1-p)f_2 \geq 0$,

$$\|pf_1 + (1-p)f_2\|_{1,\mu} = p\int_{\mathbf{X}} f_1\,d\mu \;+\; (1-p)\int_{\mathbf{X}} f_2\,d\mu = p + 1 - p = 1,$$

  and since ψ is convex

$$\begin{aligned}\int_{\mathbf{X}} \psi(pf_1 + (1-p)f_2)\,d\mu &\leq p\int_{\mathbf{X}} \psi(f_1)\,d\mu \;+\; (1-p)\int_{\mathbf{X}} \psi(f_2)\,d\mu \\ &\leq pa + (1-p)a = a,\end{aligned}$$

  hence $pf_1 + (1-p)f_2 \in C$. This proves that C is convex. To prove that it is norm-closed, since $L^1(\mathbf{X},\mathcal{B}(\mathbf{X}),\mu)$ is first-countable as Banach space, it suffices to prove that, if f_n is a sequence in C and $\|f_n - f\|_{1,\mu} \to 0$ for some $f \in L^1(\mathbf{X},\mathcal{B}(\mathbf{X}),\mu)$, then $f \in C$. Since the convergence *in norm* implies the convergence *of the norms*, $1 = \|f_n\|_{1,\mu} \to \|f\|_{1,\mu}$ and hence $\|f\|_{1,\mu} = 1$. Let f_{n_k} be a subsequence that converges to f μ-a.s.. Since $f_{n_k} \geq 0$ for all k and μ-a.s. $f_{n_k} \to f$, μ-a.s. $f \geq 0$. Moreover, the functions $\psi \circ f_{n_k}$ are bounded from below (since ψ is so), hence Fatou's Lemma can be applied to obtain

$$\int_{\mathbf{X}} \psi \circ f\,d\mu = \int_{\mathbf{X}} \liminf_{k\to\infty} \psi \circ f_{n_k}\,d\mu \leq \liminf_{k\to\infty} \int_{\mathbf{X}} \psi \circ f_{n_k}\,d\mu \leq a.$$

  We conclude that $f \in C$.

- By the Dunford-Pettis Theorem, C is a weakly pre-compact subset of $L^1(\mathbf{X},\mathcal{B}(\mathbf{X}),\mu)$ if and only if it is uniformly integrable. Let $\varepsilon > 0$; since $\log t \to +\infty$, there exists $c > 0$ such that $\log t \geq (a \wedge 1)/\varepsilon$ for all $t \geq c$, hence for all $x \in C$ such that $f(x) \geq c$

$$\psi \circ f(x) = f(x)\log f(x) \geq f(x)\frac{a \wedge 1}{\varepsilon}, \qquad \text{i.e.} \qquad f(x) \leq \psi \circ f(x)\frac{\varepsilon}{a \wedge 1}.$$

  It follows that

$$\int_{\mathbf{X}} f\,\mathbb{1}_{\{f\geq c\}}\,d\mu \leq \int_{\mathbf{X}} \psi \circ f\,\frac{\varepsilon}{a \wedge 1}\mathbb{1}_{\{f\geq c\}}\,d\mu \leq \frac{\varepsilon}{a \wedge 1}\int_{\mathbf{X}} \psi \circ f\,d\mu \leq \frac{\varepsilon}{a \wedge 1}a \leq \varepsilon,$$

  which proves the uniform integrability of C. □

[†]In general, if $\mathbf{Y}$ is a set and τ_w, τ_s are two topology on $\mathbf{Y}$ such that τ_s is stronger than τ_w (i.e. $\tau_w \subseteq \tau_s$) and $(\mathbf{Y}, \tau_s)$ is compact, then $(\mathbf{Y}, \tau_w)$ is also compact: indeed, any open cover of $(\mathbf{Y}, \tau_w)$ is also an open cover of $(\mathbf{Y}, \tau_s)$, hence it has a finite subcover. This is just the main reason why one can consider weakening a topology.

Remark 3.20 (Finite spaces). In case $\mathbf{X} = \{x_1, \dots, x_d\}$ is *finite* with d elements, we may identify $\mathfrak{M}(\mathbf{X})$ with the following convex (and compact) subset of $\mathbb{R}^d$:

$$\Big\{(\nu_1, \dots, \nu_d) \in \mathbb{R}^d : \quad \nu_i \geq 0 \ \forall i, \quad \nu_1 + \cdots + \nu_d = 1\Big\},$$

identifying each $\nu \in \mathfrak{M}(\mathbf{X})$ with the vector $(\nu_1, \dots, \nu_d)$ such that $\nu_i \doteq \nu(\{x_i\})$ for all $i = 1, \dots, d$. We also note that the weak topology on $\mathfrak{M}(\mathbf{X})$ corresponds to the Euclidean topology on $\mathbb{R}^d$. Assume now that $\mu \in \mathfrak{M}(\mathbf{X})$ is such that $\mu_i > 0$ for all i; the relative entropy of ν with respect to μ is given by

$$H(\nu|\mu) = \sum_{i=1}^{d} \log\Big(\frac{\nu_i}{\mu_i}\Big)\nu_i = \sum_{i=1}^{d} \frac{\nu_i}{\mu_i} \log\Big(\frac{\nu_i}{\mu_i}\Big)\mu_i,$$

with the convention that $0 \log 0 = 0$. It is easy to see that in this case $\nu \to H(\nu|\mu)$ is *continuous* on $\mathfrak{M}(\mathbf{X})$ (not only lower semi-continuous): it suffices to note that $t \to t \log t$ is a continuous function on $(0, 1]$ and $t \log t \to 0$ as $t \to 0$. ♢

We conclude this section with some examples of computation of the relative entropy for specific distributions.

Example 3.21. Let $\mathbf{X} = \mathbb{N}$, let μ be the geometric distribution with parameter p and let ν be the discrete uniform distribution on $\{1, \dots, r\}$ for some $r \geq 1$, i.e.

$$\begin{aligned} \mu(\{x\}) &= (1-p)^{x-1} p && \forall x \in \mathbb{N}, \\ \nu(\{x\}) &= \frac{1}{r} && \forall x = 1, \dots r. \end{aligned}$$

Since $\{1, \dots, r\} \subseteq \mathbb{N}$, ν is absolutely continuous with respect to μ, and

$$\begin{aligned} H(\nu|\mu) &= \sum_{x=1}^{r} \log\Big(\frac{\nu(\{x\})}{\mu(\{x\})}\Big)\nu(\{x\}) = \sum_{x=1}^{r} \log\Big(\frac{r^{-1}}{(1-p)^{x-1}p}\Big)\frac{1}{r} \\ &= \sum_{x=1}^{r} \Big(-\log(pr) - (x-1)\log(1-p)\Big)\frac{1}{r} = -\log(pr) \sum_{x=1}^{r} \frac{1}{r} - \frac{\log(1-p)}{r} \sum_{x=1}^{r-1} x \\ &= -\log(pr) - \frac{\log(1-p)}{r} \frac{r(r-1)}{2} = -\log(pr) - \log(1-p)\frac{r-1}{2} \end{aligned}$$

♢

Example 3.22. Let $\mathbf{X} = \mathbb{N}_0$ and let μ and ν be Poisson distributions, with (positive) parameters λ_μ and λ_ν respectively, i.e.

$$\begin{aligned} \mu(\{x\}) &= \frac{\lambda_\mu^x}{x!} e^{-\lambda_\mu} && \forall x \in \mathbb{N}_0, \\ \nu(\{x\}) &= \frac{\lambda_\nu^x}{x!} e^{-\lambda_\nu} && \forall x \in \mathbb{N}_0. \end{aligned}$$

Then μ and ν are equivalent, i.e. each is absolutely continuous with respect to the other, and

$$\begin{aligned}
H(\nu|\mu) &= \sum_{x=0}^{\infty} \log\left(\frac{\nu(\{x\})}{\mu(\{x\})}\right)\nu(\{x\}) = \sum_{x=0}^{\infty} \log\left(\frac{(x!)^{-1}\lambda_\nu^x e^{-\lambda_\nu}}{(x!)^{-1}\lambda_\mu^x e^{-\lambda_\mu}}\right)\frac{\lambda_\nu^x}{x!}e^{-\lambda_\nu} \\
&= \sum_{x=0}^{\infty}\left(x\log\left(\frac{\lambda_\nu}{\lambda_\mu}\right) - (\lambda_\nu - \lambda_\mu)\right)\frac{\lambda_\nu^x}{x!}e^{-\lambda_\nu} = \log\left(\frac{\lambda_\nu}{\lambda_\mu}\right)\sum_{x=1}^{\infty}\frac{\lambda_\nu^x}{(x-1)!}e^{-\lambda_\nu} - (\lambda_\nu - \lambda_\mu) \\
&= \log\left(\frac{\lambda_\nu}{\lambda_\mu}\right)\lambda_\nu\sum_{x=0}^{\infty}\frac{\lambda_\nu^x}{x!}e^{-\lambda_\nu} - (\lambda_\nu - \lambda_\mu) = \lambda_\nu(\log\lambda_\nu - \log\lambda_\mu) - (\lambda_\nu - \lambda_\mu)
\end{aligned}$$

◊

Example 3.23. Let $\mathbf{X} = \mathbb{R}$, let μ be the standard normal distribution and let ν be the exponential distribution with parameter $\lambda > 0$, i.e.

$$\mu(dx) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx, \qquad \nu(dx) = \lambda e^{-\lambda x}\mathbb{1}_{[0,+\infty)}(x)\,dx,$$

so that ν is absolutely continuous with respect to μ and

$$\frac{d\nu}{d\mu} = \left(\frac{d\nu}{dx}\right)\left(\frac{d\mu}{dx}\right)^{-1} = \lambda\sqrt{2\pi}\exp\left(-\lambda x + \frac{x^2}{2}\right)\mathbb{1}_{[0,+\infty)}(x).$$

Therefore, the relative entropy of ν with respect to μ is

$$\begin{aligned}
H(\nu|\mu) &= \int_{\mathbb{R}} \log\left(\frac{d\nu}{d\mu}\right)d\nu = \int_0^{\infty}\log\left(\lambda\sqrt{2\pi}\exp\left(-\lambda x + \frac{x^2}{2}\right)\right)e^{-\lambda x}\,dx \\
&= \int_0^{\infty}\left(\log\left(\lambda\sqrt{2\pi}\right) - \lambda x + \frac{x^2}{2}\right)e^{-\lambda x}\,dx \\
&= \frac{\log\left(\lambda\sqrt{2\pi}\right)}{\lambda}\int_{\mathbb{R}}\nu(dx) - \int_{\mathbb{R}} x\,\nu(dx) + \frac{1}{2\lambda}\int_{\mathbb{R}} x^2\,\nu(dx) \\
&= \frac{\log\left(\lambda\sqrt{2\pi}\right)}{\lambda} - \frac{1}{\lambda} + \frac{1}{2\lambda}\frac{2}{\lambda^2} = \frac{1}{\lambda}\left(\log\left(\lambda\sqrt{2\pi}\right) - 1 + \frac{1}{\lambda^2}\right),
\end{aligned}$$

using the following formula† for the kth moment of the exponential distribution of parameter λ:

$$\int_{\mathbb{R}} x^k\,\nu(dx) = \frac{k!}{\lambda^k}.$$

◊

†For instance, this formula can be obtained by computing the moment generating function M of such a distribution. As we showed in Example 1.28, $M(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$, hence $M^{(k)}(t) = \frac{\lambda k!}{(\lambda-t)^{k+1}}$. By Theorem 1.16, $M^{(k)}(0) = \frac{k!}{\lambda^k}$ is the kth moment.

3.4 Sanov's Theorem

Let $(\Omega, \mathcal{E}, \mathbb{P})$ be a probability space and let $\mathbf{X}$ be a Polish space. We consider a sequence $\{X_i\}_{i\geq 1}$ of i.i.d. copies of a random variable $X\colon (\Omega, \mathcal{E}, \mathbb{P}) \to (\mathbf{X}, \mathcal{B}(\mathbf{X}))$ with law μ. We call *empirical distributions* the random probability measures

$$L_n \doteq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \tag{3.6}$$

viewed as random variables $(\Omega, \mathcal{E}, \mathbb{P}) \to (\mathfrak{M}(\mathbf{X}), \mathcal{B}(\mathfrak{M}(\mathbf{X})))$, where $\mathfrak{M}(\mathbf{X})$ is equipped with the weak topology. In other words, for all $\omega \in \Omega$,

$$L_n^{\omega} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}$$

is the convex combination, with coefficients $1/n$, of the Dirac probability measures $\delta_{X_i(\omega)}$, defined by

$$\delta_{X_i(\omega)}(A) = \begin{cases} 1 & \text{if } X_i(\omega) \in A, \\ 0 & \text{if } X_i(\omega) \notin A, \end{cases} \qquad \forall A \in \mathcal{B}(\mathbf{X}).$$

Note that, if there are no repetitions in $X_1(\omega), \ldots, X_n(\omega)$, then L_n^{ω} is the discrete uniform distribution on $\{X_1(\omega), \ldots, X_n(\omega)\}$.

Remark 3.24. We first observe that the empirical distributions L_n are actually measurable functions $(\Omega, \mathcal{E}) \to (\mathfrak{M}(\mathbf{X}), \mathcal{B}(\mathfrak{M}(\mathbf{X})))$. If for all $n \in \mathbb{N}$ we set

$$\psi_n\colon \mathbf{X}^n \to \mathfrak{M}(\mathbf{X}), \qquad \psi_n(x_1, \ldots, x_n) \doteq \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i},$$

we have that $L_n = \psi_n(X_1, \ldots X_n)$. Since $(X_1, \ldots, X_n)$ is a random vector (its projections are measurable), it suffices to show that ψ_n is measurable. We can prove a stronger condition: we claim that ψ_n is continuous. By definition, the weak topology on $\mathfrak{M}(\mathbf{X})$ is the coarsest topology that makes continuous all maps of the form

$$\varphi_f\colon \mathfrak{M}(\mathbf{X}) \to \mathbb{R}, \qquad \varphi_f(\nu) \doteq \int_{\mathbf{X}} f\, d\nu$$

for $f \in C_b(\mathbf{X})$; therefore, ψ_n is continuous if and only if $\varphi_f \circ \psi_n\colon \mathbf{X}^n \to \mathbb{R}$ is continuous for all $f \in C_b(\mathbf{X})$ (see [Bre, Prop. 3.2]). For all $(x_1, \ldots, x_n) \in \mathbf{X}^n$

$$\varphi_f \circ \psi_n(x_1, \ldots, x_n) = \int_{\mathbf{X}} f\, d\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}\right) = \frac{1}{n}\sum_{i=1}^{n} f(x_i),$$

which proves the continuity of $\varphi_f \circ \psi_n$, since $f\colon \mathbf{X} \to \mathbb{R}$ is continuous. $\lozenge$

We note that, for any $A \in \mathcal{B}(\mathbf{X})$, $L_n(A)$ is a real-valued random variable:

$$L_n(A) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}.$$

Since $\{\mathbb{1}_{\{X_i \in A\}}\}_{i\geq 1}$ is a sequence of i.i.d. real-valued random variable in $L^1(\Omega,\mathcal{E},\mathbb{P})$ with mean $\mathbb{E}(\mathbb{1}_{\{X\in A\}}) = \mathbb{P}(X \in A) = \mu(A)$, the strong law of large numbers guarantees that $L_n(A) \to \mu(A)$ $\mathbb{P}$-a.s.†.

Let us try to go a bit beyond this. The strong law of large numbers states that the empirical averages converge to the mean almost surely. Similarly, is it possible to state that the empirical distributions converge weakly to the actual distribution almost surely? The answer lies in the following theorem.

Theorem 3.25 (Strong law of large numbers for empirical distributions)**.** Let $(\Omega,\mathcal{E},\mathbb{P})$ be a probability space and let $\mathbf{X}$ be a Polish space. Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. copies of a random variable $X\colon (\Omega,\mathcal{E},\mathbb{P}) \to (\mathbf{X},\mathcal{B}(\mathbf{X}))$, with law μ. Let

$$L_n \doteq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

be the empirical distributions, considered as a sequence of random variables $(\Omega,\mathcal{E},\mathbb{P}) \to (\mathfrak{M}(\mathbf{X}),\mathcal{B}(\mathfrak{M}(\mathbf{X})))$, where $\mathfrak{M}(\mathbf{X})$ is equipped with the weak topology. Then, $L_n \rightharpoonup \mu$ $\mathbb{P}$-almost surely.

We will first prove the proposition for a compact Polish space $\mathbf{X}$. In this case we have a handier characterization of weak topology, as we show in the next lemma.

Lemma 3.26. Let $\mathbf{X}$ be a compact Polish space. Then the weak topology of $\mathfrak{M}(\mathbf{X})$ coincides with the coarsest topology that makes continuous all maps

$$\varphi_f\colon \mathfrak{M}(\mathbf{X}) \to \mathbb{R}, \qquad \varphi_f(\nu) \doteq \int_{\mathbf{X}} f\, d\nu$$

for $f \in D$, where D is a countable dense subset of $C(\mathbf{X})$.

†Several researches concern Glivenko-Cantelli classes, i.e. the classes $\mathcal{C} \subseteq \mathcal{B}(\mathbf{X})$ such that the almost sure convergence of $L_n(A)$ to $\mu(A)$ is also uniform for $A \in \mathcal{C}$, meaning that

$$\sup_{A\in\mathcal{C}} |L_n(A) - \mu(A)| \to 0 \qquad \mathbb{P}\text{-a.s..}$$

For instance, the classical Glivenko-Cantelli Theorem (see [Kle, Th.5.23]) states that, for real-valued random variables, the class $\mathcal{C} \doteq \{(-\infty,x] : x \in \mathbb{R}\} \subseteq \mathcal{B}(\mathbb{R})$ is Glivenko-Cantelli. In other words, if $F_n(x) \doteq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\leq x\}}$ is the empirical distribution function (which is a random variable, for any fixed $x \in \mathbb{R}$) of the sequence $\{X_i\}_{i\geq 1}$, and $F(x) \doteq \mathbb{P}(X \leq x)$ is the distribution function of $\mathbf{X}$, then

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \|F_n - F\|_\infty \to 0 \qquad \mathbb{P}\text{-a.s..}$$

Proof. Since $\mathbf{X}$ is a compact metrizable space, by Example 2.17 (ii) $C(\mathbf{X}) = C_b(\mathbf{X})$ is Polish, hence there exists a countable dense subset $D \subseteq C(\mathbf{X})$. Let τ_w be the weak topology of $\mathfrak{M}(\mathbf{X})$, i.e. the coarsest topology that makes continuous all maps φ_f for $f \in C(\mathbf{X})$, and let τ be the coarsest topology on $\mathfrak{M}(\mathbf{X})$ that makes continuous all maps φ_f for $f \in D$. Since $D \subseteq C(\mathbf{X})$, it is obvious that $\tau \subseteq \tau_w$. Conversely, to prove that $\tau_w \subseteq \tau$, it suffices to show that φ_f is continuous with respect to τ, for any $f \in C(\mathbf{X})$ (not necessarily $\in D$). Let $f \in C(\mathbf{X})$ and $A \subseteq \mathbb{R}$ open: we wish to prove that $\varphi_f^{-1}(A) \in \tau$, i.e. any $\nu \in \varphi_f^{-1}(A)$ has an open (for the topology τ) neighborhood $O \subseteq \varphi_f^{-1}(A)$. Since $\varphi_f(\nu) \in A$ and A is open, there exists $\varepsilon > 0$ such that $(\varphi_f(\nu) - \varepsilon, \varphi_f(\nu) + \varepsilon) \subseteq A$. Since D is dense in $C(\mathbf{X})$, there exists $g \in D$ such that $\|f - g\|_\infty < \varepsilon/4$. The set

$$O \doteq \left\{ \nu' \in \mathfrak{M}(\mathbf{X}) : \quad |\varphi_g(\nu) - \varphi_g(\nu')| < \frac{\varepsilon}{2} \right\}$$

contains ν and it is open with respect to τ, since $g \in D$. Moreover, for all $\nu' \in O$ we have that

$$\begin{aligned}
&\left| \int_{\mathbf{X}} f\, d\nu - \int_{\mathbf{X}} f\, d\nu' \right| \\
\leq &\left| \int_{\mathbf{X}} f\, d\nu - \int_{\mathbf{X}} g\, d\nu \right| + \left| \int_{\mathbf{X}} g\, d\nu - \int_{\mathbf{X}} g\, d\nu' \right| + \left| \int_{\mathbf{X}} g\, d\nu' - \int_{\mathbf{X}} f\, d\nu' \right| \\
\leq &\|f - g\|_\infty + |\varphi_g(\nu) - \varphi_g(\nu')| + \|g - f\|_\infty < \frac{\varepsilon}{4} + \frac{\varepsilon}{2} + \frac{\varepsilon}{4} = \varepsilon.
\end{aligned}$$

In other words, $|\varphi_f(\nu) - \varphi_f(\nu')| < \varepsilon$, i.e. $\varphi_f(\nu') \in (\varphi_f(\nu) - \varepsilon, \varphi_f(\nu) + \varepsilon) \subseteq A$. Since $\nu' \in O$ is arbitrary, $O \subseteq \varphi_f^{-1}(A)$. Hence, with respect to τ, O is an open neighborhood of ν such that $O \subseteq \varphi_f^{-1}(A)$. □

Proof of Theorem 3.25. We should prove that

$$\mathbb{P}\text{-a.s.}, \qquad \forall f \in C_b(\mathbf{X}) \qquad \int_{\mathbf{X}} f\, dL_n \xrightarrow{n\to\infty} \int_{\mathbf{X}} f\, d\mu.$$

We stress that a priori this is *not* equivalent to require that "$\forall f \in C_b(\mathbf{X})$, $\mathbb{P}$-a.s. the convergence of the integrals holds", since $C_b(\mathbf{X})$ is not countable. For this reason, we will first assume that $\mathbf{X}$ is compact, to apply Lemma 3.26; then, we will extend the result by using compactifications.

Let $\mathbf{X}$ be a compact Polish space. By Lemma 3.26, the weak topology of $\mathfrak{M}(\mathbf{X})$ coincides with the coarsest topology that makes continuous all maps

$$\mathfrak{M}(\mathbf{X}) \to \mathbb{R}, \qquad \nu \to \int_{\mathbf{X}} f\, d\nu$$

for $f \in D$, where D is a countable dense subset of $C(\mathbf{X})$. If we *fix* $f \in D$, we have

$$\int_{\mathbf{X}} f\, dL_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \xrightarrow{n\to\infty} \mathbb{E}(f(X)) = \int_{\mathbf{X}} f\, d\mu \qquad \mathbb{P}\text{-a.s.},$$

by the strong law of large numbers. In other words, for all $f \in D$, there exists $\Omega_f \in \mathcal{E}$ such that $\mathbb{P}(\Omega_f) = 1$ and $\int_{\mathbf{X}} f\, dL_n^\omega \xrightarrow{n\to\infty} \int_{\mathbf{X}} f\, d\mu$ for all $\omega \in \Omega_f$. Since D is countable, $\Omega' \doteq \bigcap_{f\in D} \Omega_f$ is in $\mathcal{E}$, $\mathbb{P}(\Omega') = 1$ and

$$\forall \omega \in \Omega', \qquad \forall f \in D \qquad \int_{\mathbf{X}} f\, dL_n^\omega \xrightarrow{n\to\infty} \int_{\mathbf{X}} f\, d\mu.$$

This proves that $\mathbb{P}$-a.s. $L_n \rightharpoonup \mu$ in the weak topology of $\mathfrak{M}(\mathbf{X})$.

Let now $\mathbf{X}$ be a generic Polish space and let $\overline{\mathbf{X}}$ be a metrizable compactification of $\mathbf{X}$ (see Theorem 2.68). We see "$\mathfrak{M}(\mathbf{X}) \subseteq \mathfrak{M}(\overline{\mathbf{X}})$", in the sense of Lemma 2.70: any $\nu \in \mathfrak{M}(\mathbf{X})$ can be extended to a $\overline{\nu} \in \mathfrak{M}(\overline{\mathbf{X}})$ such that $\overline{\nu}(\overline{\mathbf{X}} \setminus \mathbf{X}) = 0$, and vice versa. We may consider the $\overline{X}_i$ (extensions of the X_i, viewed as $\overline{\mathbf{X}}$-random variables), with common law $\overline{\mu}$ (extension of μ to $\mathfrak{M}(\overline{\mathbf{X}})$). Their empirical distributions $\overline{L}_n$ are $\mathfrak{M}(\overline{\mathbf{X}})$-valued random variables that converge weakly to $\overline{\mu}$ a.s., since the claim of the theorem holds for compact Polish spaces. By Lemma 2.71, the weak topology of $\mathfrak{M}(\mathbf{X})$ coincides with the topology that $\mathfrak{M}(\overline{\mathbf{X}})$ induces on $\mathfrak{M}(\mathbf{X})$, hence also $L_n \rightharpoonup \mu$ a.s. in the weak topology of $\mathfrak{M}(\mathbf{X})$. □

In the light of the almost sure convergence of L_n to μ, it is reasonable to wonder if the laws of the empirical distributions satisfy a LDP in $\mathfrak{M}(\mathfrak{M}(\mathbf{X}))$: the answer is positive and it lies in Sanov's Theorem. A discrete version of this theorem for a finite space $\mathbf{X}$ was proved in 1877 by the statistical physicist Boltzmann, who was studying the dynamics of a discrete ideal gas (see the paper [Ell], which contains various interesting applications of large deviations theory to statistical mechanics). However, the theorem owes its name to the russian mathematician Ivan Nikolaevich Sanov, who proved the result in the significantly more general case $\mathbf{X} = \mathbb{R}$, in 1957 (see [San]). We start with a preliminary lemma.

Lemma 3.27. Let $\mathbf{X}$ be a Polish space, let $f : \mathbf{X} \to \mathbb{R}^d$ be continuous and bounded and let X be a $\mathbf{X}$-valued random variable with law μ. Let

$$M(t) \doteq \mathbb{E}\Big(e^{\langle t, f(X)\rangle}\Big), \qquad t \in \mathbb{R}^d$$

be the moment generating function of the $\mathbb{R}^d$-valued random variable $f(X)$. Let $H(\cdot|\mu)$ be the relative entropy with respect to μ. Then

$$\sup_{t\in\mathbb{R}^d} [\langle t, y\rangle - \log M(t)] = \inf_{\substack{\nu\in\mathfrak{M}(\mathbf{X}) \\ \int_{\mathbf{X}} f\, d\nu = y}} H(\nu|\mu), \qquad \forall y \in \mathbb{R}^d. \tag{3.7}$$

Proof. Let

$$J(y) \doteq \sup_{t\in\mathbb{R}^d} [\langle t, y\rangle - \log M(t)], \qquad y \in \mathbb{R}^d$$

be the Legendre transform of $\log M$ and let

$$G(y) \doteq \inf_{\substack{\nu\in\mathfrak{M}(\mathbf{X})\\ \int_{\mathbf{X}} f\,d\nu=y}} H(\nu|\mu), \qquad y\in\mathbb{R}^d.$$

To prove that $J = G$ on $\mathbb{R}^d$, we distinguish the following three cases, according to whether y is an element of $\mathbb{R}^d \setminus \overline{\mathcal{D}}_J$, $\mathring{\mathcal{D}}_J$ or $\partial\mathcal{D}_J$.

- Let $y \in \mathbb{R}^d \setminus \overline{\mathcal{D}}_J$. We claim that for all $\nu \in \mathfrak{M}(\mathbf{X})$, $\nu \ll \mu$, $\int_{\mathbf{X}} f(x)\,\nu(dx)$ is in the closed convex hull of $\operatorname{supp}(\mu\circ f^{-1})$. Indeed, since f is bounded and $\nu \ll \mu$, there exists $M > 0$ such that $\operatorname{supp}(\nu\circ f^{-1}) \subseteq \operatorname{supp}(\mu\circ f^{-1}) \subseteq [-M,M)^d$. For all $n \in \mathbb{N}$, we consider the partition $\left\{Q_i^{(n)}\right\}_{i=1}^{n^d}$ of $[-M,M)^d$ into n^d "equal" squares $Q_i^{(n)}$ of the form $[a_1,b_1)\times\cdots\times[a_d,b_d)$, where $b_j - a_j = 2M/n$ for all $j = 1,\ldots d$. We then define a simple function $h_n\colon [-M,M)^d \to \mathbb{R}^d$ by setting

$$h_n \doteq \sum_{i=1}^{n^d} y_i^{(n)} \mathbb{1}_{Q_i^{(n)}},$$

where $y_i^{(n)}$ is either any fixed point of $Q_i^{(n)}$ such that $y_i^{(n)} \in \operatorname{supp}(\mu\circ f^{-1})$ if it exists, or any fixed point of $Q_i^{(n)}$ if $Q_i^{(n)} \cap \operatorname{supp}(\mu\circ f^{-1}) = \varnothing$. It is easy to see that $h_n(y) \to y$ uniformly in $[-M,M)^d$, hence

$$\begin{aligned}\sum_{i=1}^{n^d} y_i^{(n)} (\nu\circ f^{-1})\big(Q_i^{(n)}\big) &= \int_{[-M,M)^d} h_n(y)\,(\nu\circ f^{-1})(dy)\\ &\xrightarrow{n\to\infty} \int_{[-M,M)^d} y\,(\nu\circ f^{-1})(dy) = \int_{\mathbf{X}} f(x)\,\nu(dx).\end{aligned} \tag{3.8}$$

We note that the sums above are convex combinations of points of $\operatorname{supp}(\mu\circ f^{-1})$. Indeed, if some $y_i^{(n)} \notin \operatorname{supp}(\mu\circ f^{-1})$, then $Q_i^{(n)} \cap \operatorname{supp}(\mu\circ f^{-1}) = \varnothing$, hence also $Q_i^{(n)} \cap \operatorname{supp}(\nu\circ f^{-1}) = \varnothing$ since $\nu \ll \mu$, and $(\nu\circ f^{-1})\big(Q_i^{(n)}\big) = 0$; it follows that, for any non-zero term in the sum, $y_i^{(n)} \in \operatorname{supp}(\mu\circ f^{-1})$; moreover, since $\operatorname{supp}(\nu\circ f^{-1}) \subseteq [-M,M)^d$,

$$\sum_{i=1}^{n^d} \nu\circ f^{-1}\big(Q_i^{(n)}\big) = \sum_{i=1}^{n^d} \nu\Big(x\in\mathbf{X}:\ f(x)\in Q_i^{(n)}\Big) = \nu\Big(x\in\mathbf{X}:\ f(x)\in[-M,M)^d\Big) = 1.$$

We deduce that the sums in (3.8) are in the convex hull of $\operatorname{supp}(\mu\circ f^{-1})$, therefore their limit $\int_{\mathbf{X}} f(x)\,\nu(dx)$ is an element of the *closed* convex hull of $\operatorname{supp}(\mu\circ f^{-1}) = \operatorname{supp}(f(X))$: this concludes the proof of the claim. For all

$\nu \in \mathfrak{M}(\mathbf{X})$, $\nu \ll \mu$, by Theorem 3.13 (iv) we have that $\int_{\mathbf{X}} f(x)\,\nu(dx) \in \overline{\mathcal{D}}_J$, hence $\int_{\mathbf{X}} f\,d\nu \neq y \in \mathbb{R}^d \setminus \overline{\mathcal{D}}_J$. In other words, for all $\nu \in \mathfrak{M}(\mathbf{X})$ such that $\int_{\mathbf{X}} f\,d\nu = y$, it turns out that $\nu \not\ll \mu$, hence $H(\nu|\mu) = +\infty$. This proves that $G(y) = +\infty$; since $y \notin \mathcal{D}_J$, also $J(y) = +\infty$, and (3.7) holds.

- Let $y \in \mathring{\mathcal{D}}_J$. Since by Theorem 3.13 (i) $J \in \mathrm{Conv}_\infty(\mathbb{R}^d)$, by Theorem 3.9 (ii)

$$J(y) = \langle t, y\rangle - \log M(t),$$

where t is the unique point in $\mathbb{R}^d$ such that $y = \nabla \log M(t)$. If as usual $\mathbb{P}_t$ denotes the probability measure on $(\Omega, \mathcal{E})$ defined by

$$\frac{d\mathbb{P}_t}{d\mathbb{P}} = \frac{e^{\langle t, f(X)\rangle}}{M(t)},$$

$\mathbb{E}_t$ denotes the expectation with respect to $\mathbb{P}_t$ and μ_t denotes the law of X with respect to $\mathbb{P}_t$, by Theorem 3.12 (i)

$$\nabla \log M(t) = \mathbb{E}_t(f(X)) = \int_{\mathbf{X}} f(x)\,\mu_t(dx).$$

It follows that $y = \int_{\mathbf{X}} f\,d\mu_t$. Thus, in order to obtain the equality $J(y) = G(y)$, we shall prove that $J(y) \leq H(\nu|\mu)$ for any $\nu \in \mathfrak{M}(\mathbf{X})$ such that $\int_{\mathbf{X}} f\,d\nu = y$, and the equality holds if and only if $\nu = \mu_t$. We obviously may assume that $H(\nu|\mu) < +\infty$ (in particular, $\nu \ll \mu$). Even in case that $\mu \not\ll \nu$, by the Lebesgue decomposition theorem (see Remark 3.16), we can write $d\mu = h\,d\nu + \mathbb{1}_A\,d\mu$, where $h\colon \mathbf{X} \to \mathbb{R}$ is a nonnegative measurable function and $A \in \mathcal{B}(\mathbf{X})$ is such that $\nu(A) = 0$. In particular,

$$\frac{d\nu}{d\mu}\,d\mu = \frac{d\nu}{d\mu} h\,d\nu + \frac{d\nu}{d\mu}\mathbb{1}_A\,d\mu.$$

Since $0 = \nu(A) = \int_{\mathbf{X}} \mathbb{1}_A \frac{d\nu}{d\mu}\,d\mu$, it follows that μ-a.s. $\mathbb{1}_A \frac{d\nu}{d\mu} = 0$, hence $d\nu = \frac{d\nu}{d\mu} h\,d\nu$: this means that ν-a.s. $h = \left(\frac{d\nu}{d\mu}\right)^{-1}$. Thus, for any $\nu \in \mathfrak{M}(\mathbf{X})$ such that $\int_{\mathbf{X}} f\,d\nu = y$,

$$\begin{aligned}
&\langle t, y\rangle - H(\nu|\mu) = \left\langle t, \int_{\mathbf{X}} f(x)\,\nu(dx)\right\rangle - \int_{\mathbf{X}} \log\left(\frac{d\nu}{d\mu}(x)\right)\nu(dx)\\
&= \int_{\mathbf{X}} \left(\langle t, f(x)\rangle - \log\left(\frac{d\nu}{d\mu}(x)\right)\right)\nu(dx) = \int_{\mathbf{X}} \log\left[e^{\langle t, f(x)\rangle}\left(\frac{d\nu}{d\mu}(x)\right)^{-1}\right]\nu(dx)\\
&\leq \log\left[\int_{\mathbf{X}} e^{\langle t, f(x)\rangle}\left(\frac{d\nu}{d\mu}(x)\right)^{-1}\nu(dx)\right] = \log\left[\int_{\mathbf{X}} e^{\langle t, f(x)\rangle} h(x)\,\nu(dx)\right]\\
&= \log\left[\int_{\mathbf{X}} e^{\langle t, f(x)\rangle}\,\mu(dx) - \int_{\mathbf{X}} e^{\langle t, f(x)\rangle}\mathbb{1}_A\,\mu(dx)\right] \leq \log\left[\int_{\mathbf{X}} e^{\langle t, f(x)\rangle}\,\mu(dx)\right]\\
&= \log \mathbb{E}\left(e^{\langle t, f(X)\rangle}\right) = \log M(t),
\end{aligned}$$

where the first inequality is an application of Jensen's inequality to the concave function log. Since log is also *strictly* concave, this inequality is an equality if and only if ν-a.s. $e^{\langle t,f(x)\rangle}\left(\frac{d\nu}{d\mu}(x)\right)^{-1} = c$ is constant, i.e. $\frac{d\nu}{d\mu}(x) = e^{\langle t,f(x)\rangle}c^{-1}$, where the constant satisfies

$$c = \int_{\mathbf{X}} e^{\langle t,f(x)\rangle}\mu(dx) = M(t),$$

so that $\frac{d\nu}{d\mu}$ is a density; in particular, $\nu = \mu_t$. In this case, $\mu \ll \nu$ (since $\frac{d\nu}{d\mu} > 0$), hence $\nu(A) = 0$ implies that $\mu(A) = 0$, i.e. μ-a.s. $\mathbb{1}_A = 0$, and also the second inequality is an equality. We conclude that $\langle t,y\rangle - H(\nu|\mu) \leq \log M(t)$, i.e. $J(y) \leq H(\nu|\mu)$, for any $\nu \in \mathfrak{M}(\mathbf{X})$ such that $\int_{\mathbf{X}} f\,d\nu = y$, and the equality holds if and only if $\nu = \mu_t$, as we claimed.

- The last case is $y \in \partial\mathcal{D}_J$. Since we have already proved that G and J are equal on $\mathring{\mathcal{D}}_J$ and $\mathbb{R}^d \setminus \overline{\mathcal{D}}_J$, their domains have the same boundary $\partial\mathcal{D}_J = \overline{\mathcal{D}}_J \setminus \mathring{\mathcal{D}}_J$. If we are able to prove that J and G are both lower semi-continuous and convex, by Lemma 3.6 they are continuous on the *closure* of their domains, hence equality on $\mathring{\mathcal{D}}_J$ implies equality on $\partial\mathcal{D}_J$. Thus it only remains to prove the two properties for both J and G. As a Legendre transform, J is lower semi-continuous and convex by Lemma 3.5. By Proposition 3.19 (iii), we know that $H(\cdot|\mu)$ is a good rate function; moreover, $\nu \to \int_{\mathbf{X}} f\,d\nu$ is a continuous function $\mathfrak{M}(\mathbf{X}) \to \mathbb{R}^d$ if $\mathfrak{M}(\mathbf{X})$ is equipped with the weak topology, since f is continuous and bounded; by the contraction principle (Theorem 2.76), G is a good rate function, in particular it is lower semi-continuous. To prove that G is convex, let $y_1, y_2 \in \mathbb{R}^d$ and $p \in (0,1)$: we wish to show that

$$\inf_{\int_{\mathbf{X}} f\,d\nu = py_1+(1-p)y_2} H(\nu|\mu) \leq p \inf_{\int_{\mathbf{X}} f\,d\nu = y_1} H(\nu|\mu) + (1-p) \inf_{\int_{\mathbf{X}} f\,d\nu = y_2} H(\nu|\mu). \tag{3.9}$$

For any $\nu_1, \nu_2 \in \mathfrak{M}(\mathbf{X})$ such that $\int_{\mathbf{X}} f\,d\nu_1 = y_1$ and $\int_{\mathbf{X}} f\,d\nu_2 = y_2$, we have

$$py_1 + (1-p)y_2 = p\int_{\mathbf{X}} f\,d\nu_1 + (1-p)\int_{\mathbf{X}} f\,d\nu_2 = \int_{\mathbf{X}} f\,d(p\nu_1 + (1-p)\nu_2),$$

hence

$$\begin{aligned} G(py_1 + (1-p)y_2) &= \inf_{\int_{\mathbf{X}} f\,d\nu = \int_{\mathbf{X}} f\,d(p\nu_1+(1-p)\nu_2)} H(\nu|\mu) \\ &\leq H(p\nu_1 + (1-p)\nu_2|\mu) \leq pH(\nu_1|\mu) + (1-p)H(\nu_2|\mu), \end{aligned}$$

since $H(\cdot|\mu)$ is convex by Proposition 3.19 (ii). Taking the infima over the sets $\{\nu_1 \in \mathfrak{M}(\mathbf{X}) : \int_{\mathbf{X}} f\,d\nu_1 = y_1\}$ and $\{\nu_2 \in \mathfrak{M}(\mathbf{X}) : \int_{\mathbf{X}} f\,d\nu_2 = y_2\}$, we obtain

$$\begin{aligned} G(py_1 + (1-p)y_2) &\leq p \inf_{\int_{\mathbf{X}} f\,d\nu_1 = y_1} H(\nu_1|\mu) + (1-p) \inf_{\int_{\mathbf{X}} f\,d\nu_2 = y_2} H(\nu_2|\mu) \\ &= pG(y_1) + (1-p)G(y_2), \end{aligned}$$

which proves the convexity of G. □

Theorem 3.28 (Sanov). Let $(\Omega,\mathcal{E},\mathbb{P})$ be a probability space and let $\mathbf{X}$ be a Polish space. Let $X\colon(\Omega,\mathcal{E},\mathbb{P})\to(\mathbf{X},\mathcal{B}(\mathbf{X}))$ be a random variable with law μ, and let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. copies of X. Let

$$L_n \doteq \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

be the empirical distributions, considered as a sequence of random variables $(\Omega,\mathcal{E},\mathbb{P})\to(\mathfrak{M}(\mathbf{X}),\mathcal{B}(\mathfrak{M}(\mathbf{X})))$, where $\mathfrak{M}(\mathbf{X})$ is equipped with the weak topology. Then, the laws $\mathbb{P}(L_n\in\cdot)$ of the L_n satisfy the large deviations principle with rate n and good rate function the relative entropy $H(\cdot|\mu)$.

Proof. We first assume that $\mathbf{X}$ is *compact*. The strategy of our proof basically consists in applying Cramér's Theorem in $\mathbb{R}^d$ and the projective limit theorem. We now see in what sense the three hypotheses of the projective limit theorem are satisfied (in the notation of that theorem, $\mathbf{X}$ is replaced with $\mathfrak{M}(\mathbf{X})$ and $\mathbf{Y}_i$ with $\mathbb{R}$ for all $i\in\mathbb{N}$):

(i) Since $\mathbf{X}$ is compact, $\mathfrak{M}(\mathbf{X})$ is tight (see Example 2.45), hence $\mathfrak{M}(\mathbf{X})$ is compact in the weak topology by Prohorov's Theorem (Theorem 2.48). It follows that any sequence of probability measures on the space $(\mathfrak{M}(\mathbf{X}),\mathcal{B}(\mathfrak{M}(\mathbf{X})))$, i.e. any sequence in $\mathfrak{M}(\mathfrak{M}(\mathbf{X}))$, is exponentially tight with any rate (see Example 2.54); in particular, $\mathbb{P}(L_n\in\cdot)$ is exponentially tight with any rate.

(ii) Since $\mathbf{X}$ is compact, by Lemma 3.26 the weak topology of $\mathfrak{M}(\mathbf{X})$ coincides with the coarsest topology that makes continuous all maps

$$\varphi_j\colon\mathfrak{M}(\mathbf{X})\to\mathbb{R},\qquad \varphi_j(\nu)\doteq\int_{\mathbf{X}} f_j\,d\nu$$

for $j\in\mathbb{N}$, where $\{f_j\}_{j\in\mathbb{N}}\subseteq C(\mathbf{X})$ is a dense sequence. We claim that $\{\varphi_j\}_{j\in\mathbb{N}}$ separates points: let $\nu_1,\nu_2\in\mathfrak{M}(\mathbf{X})$ be such that for all $j\in\mathbb{N}$ $\varphi_j(\nu_1)=\varphi_j(\nu_2)$, i.e. $\int_{\mathbf{X}} f_j\,d\nu_1=\int_{\mathbf{X}} f_j\,d\nu_2$. Since $\{f_j\}_{j\in\mathbb{N}}$ is dense in $C(\mathbf{X})$, for each $f\in C(\mathbf{X})$ and $\varepsilon>0$ there exists j such that $\|f_j-f\|_\infty<\varepsilon/2$, hence

$$\begin{aligned}&\left|\int_{\mathbf{X}} f\,d\nu_1-\int_{\mathbf{X}} f\,d\nu_2\right|\\ \leq&\left|\int_{\mathbf{X}} f\,d\nu_1-\int_{\mathbf{X}} f_j\,d\nu_1\right|+\left|\int_{\mathbf{X}} f_j\,d\nu_1-\int_{\mathbf{X}} f_j\,d\nu_2\right|+\left|\int_{\mathbf{X}} f_j\,d\nu_2-\int_{\mathbf{X}} f\,d\nu_2\right|\\ \leq&\|f-f_j\|_\infty+0+\|f_j-f\|_\infty<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon.\end{aligned}$$

Since ε is arbitrary, $\int_{\mathbf{X}} f\,d\nu_1=\int_{\mathbf{X}} f\,d\nu_2$. Since the integrals of the real-valued bounded continuous functions determine a measure on $(\mathbf{X},\mathcal{B}(\mathbf{X}))^{\dagger}$, it follows that $\nu_1 = \nu_2$. This proves that $\{\varphi_j\}_{j\in\mathbb{N}}$ separates points.

(iii) For any $F \subseteq \mathbb{N}$ finite, we define

$$\varphi_F : \mathfrak{M}(\mathbf{X}) \to \mathbb{R}^F, \qquad \varphi_F(\nu) \doteq \big(\varphi_j(\nu)\big)_{j\in F},$$
$$f_F : \mathbf{X} \to \mathbb{R}^F, \qquad f_F(x) \doteq \big(f_j(x)\big)_{j\in F}.$$

Both φ_F and f_F are continuous, since their components are so. The image measures of $\mathbb{P}(L_n \in \cdot)$ under φ_F are $\mathbb{P}(\varphi(L_n) \in \cdot)$, i.e. the laws of the $\varphi(L_n)$ as probability measures on $\mathbb{R}^F$. It turns out that

$$\varphi_F(L_n) = \big(\varphi_j(L_n)\big)_{j\in F} = \left(\int_{\mathbf{X}} f_j \, dL_n\right)_{j\in F} = \int_{\mathbf{X}} f_F \, dL_n = \frac{1}{n}\sum_{i=1}^{n} f_F(X_i)$$

is the empirical average of $\{f_F(X_i)\}_{i\in\mathbb{N}}$, which is a sequence of $\mathbb{R}^F$-valued i.i.d. random variables, whose moment generating function

$$M_F(t) = \mathbb{E}(\exp\langle t, f_F(X)\rangle)$$

is finite for all $t \in \mathbb{R}^F$ (since f_F is bounded). By Cramér's Theorem in $\mathbb{R}^F$ (Theorem 3.14), the laws of the $\varphi_F(L_n)$ satisfy the LDP with rate n and good rate function

$$J_F(y) = \sup_{t\in\mathbb{R}^F}[\langle t, y\rangle - \log M_F(t)] \qquad \forall y \in \mathbb{R}^F.$$

Therefore, the three hypotheses of the projective limit theorem (Theorem 2.79) are satisfied: there exists a unique good rate function I such that

$$J_F(y) = \inf_{\substack{\nu\in\mathfrak{M}(\mathbf{X})\\ \varphi_F(\nu)=y}} I(\nu) \qquad \forall F \subseteq \mathbb{N} \text{ finite}, \qquad \forall y \in \mathbb{R}^F,$$

and the sequence $\mathbb{P}(L_n \in \cdot)$ satisfies the LDP with rate n and good rate function I. Since $\varphi_F(\nu) = \int_{\mathbf{X}} f_F \, d\nu$ and $f_F : \mathbf{X} \to \mathbb{R}^d$ is continuous and bounded, by Lemma 3.27 $I = H(\cdot|\mu)$. This proves the theorem for compact Polish spaces.

Let now $\mathbf{X}$ be a generic Polish space, and let $\overline{\mathbf{X}}$ be a metrizable compactification of $\mathbf{X}$ (see Theorem 2.68). Once again, we see "$\mathfrak{M}(\mathbf{X}) \subseteq \mathfrak{M}(\overline{\mathbf{X}})$" in the sense of Lemma 2.70: any $\nu \in \mathfrak{M}(\mathbf{X})$ can be extended to a $\overline{\nu} \in \mathfrak{M}(\overline{\mathbf{X}})$ such that $\overline{\nu}(\overline{\mathbf{X}}\setminus\mathbf{X}) = 0$, and

†This is a well-known fact, which can be easily proved. Let $O \subseteq \mathbf{X}$ be open. By Proposition 2.12 (ii), there exists a sequence $g_k \in C_{b,+}(\mathbf{X})$ such that $g_k \nearrow \mathbb{1}_O$. By hypothesis, $\int_{\mathbf{X}} g_k \, d\nu_1 = \int_{\mathbf{X}} g_k \, d\nu_2$ for all k. By dominated convergence (the g_k are bounded), also $\int_{\mathbf{X}} \mathbb{1}_O \, d\nu_1 = \int_{\mathbf{X}} \mathbb{1}_O \, d\nu_2$, i.e. $\nu_1(O) = \nu_2(O)$. Since the open sets form a system of generators for $\mathcal{B}(\mathbf{X})$ that is closed under finite intersection, it follows that $\nu_1 = \nu_2$.

vice versa. We consider the $\overline{X}_i$ (extensions of the X_i, viewed as $\overline{\mathbf{X}}$-random variables), with common law $\overline{\mu}$ (extension of μ to $\mathfrak{M}(\overline{\mathbf{X}})$). Their empirical distributions $\overline{L}_n$ are $\mathfrak{M}(\overline{\mathbf{X}})$-valued random variables, whose laws $\mathbb{P}(\overline{L}_n \in \cdot) \in \mathfrak{M}(\mathfrak{M}(\overline{\mathbf{X}}))$ satisfy the LDP with rate n and good rate function $H(\cdot|\overline{\mu})$, since the claim of the theorem holds for compact Polish spaces. To apply the restriction principle (Lemma 2.72), we have to show that $H(\nu|\overline{\mu}) = +\infty$ for all $\nu \in \mathfrak{M}(\overline{\mathbf{X}}) \setminus \mathfrak{M}(\mathbf{X})$, i.e. $\nu(\overline{\mathbf{X}} \setminus \mathbf{X}) > 0$. Since $\overline{\mu}(\overline{\mathbf{X}} \setminus \mathbf{X}) = 0$, if $\nu \ll \overline{\mu}$ then $\nu(\overline{\mathbf{X}} \setminus \mathbf{X}) = 0$; it follows that, if $\nu(\overline{\mathbf{X}} \setminus \mathbf{X}) > 0$, then $\nu \not\ll \mu$ and $H(\nu|\overline{\mu}) = +\infty$, as we claimed. By the restriction principle, the probability measures $\mathbb{P}(L_n \in \cdot) \in \mathfrak{M}(\mathfrak{M}(\mathbf{X}))$ satisfy the LDP with rate n and good rate function $H(\cdot|\mu)$. □

Remark 3.29. We saw in section 2.6 that the projective limit theorem is a sort of inverse of the contraction principle. Since we basically deduced Sanov's Theorem from the projective limit theorem and Cramér's Theorem in $\mathbb{R}^d$, it is natural to think that Cramér's Theorem in $\mathbb{R}^d$ may be deduced from the contraction principle and Sanov's Theorem. We briefly show the argument, assuming that the $\mathbb{R}^d$-valued i.i.d. random variables X_i, copies of a random variable X with law μ, are *bounded* for the sake of simplicity (in this case, obviously, the moment generating function is finite everywhere). Let $M > 0$ such that $|X| \leq M$ a.s.. We define the function

$$f\colon \mathbb{R}^d \to \mathbb{R}^d, \qquad f(x) = \begin{cases} x & |x| \leq M, \\ M\dfrac{x}{|x|} & |x| > M. \end{cases}$$

The continuity of f is obvious at all points x_0 such that $|x_0| < M$ or $|x_0| > M$; if $|x_0| = M$, then $f(x) \to x_0 = f(x_0)$ as $x \to x_0$, $|x| \leq M$, and $f(x) \to Mx_0|x_0|^{-1} = x_0$ as $x \to x_0$, $|x| > M$. Hence, f is continuous everywhere; moreover, it is bounded since $|f| \leq M$. Since $\mathfrak{M}(\mathbb{R}^d)$ is equipped with the weak topology, the function

$$\varphi\colon \mathfrak{M}(\mathbb{R}^d) \to \mathbb{R}^d, \qquad \varphi(\nu) = \int_{\mathbb{R}^d} f\, d\nu$$

is continuous, as an integral of a bounded continuous function. By Sanov's Theorem, the laws of the empirical distributions L_n satisfy the LDP with rate n and rate function $H(\cdot|\mu)$. By the contraction principle, the laws of the $\varphi(L_n)$ satisfy the LDP with rate n and rate function

$$J(y) \doteq \inf_{\substack{\nu \in \mathfrak{M}(\mathbb{R}^d) \\ \int_{\mathbb{R}^d} f\, d\nu = y}} H(\nu|\mu).$$

Since $|X| \leq M$ a.s., $f(X_i) = X_i$ a.s. for all $i \geq 1$, hence

$$\varphi(L_n) = \int_{\mathbb{R}^d} f\, dL_n = \frac{1}{n}\sum_{i=1}^n f(X_i) = \frac{1}{n}\sum_{i=1}^n X_i = \frac{S_n}{n} \qquad \text{a.s..}$$

Moreover, by Lemma 3.27, J is the Legendre transform of the logarithmic moment generating function of $f(X) = X$. This proves the claim of Cramér's Theorem in $\mathbb{R}^d$ (Theorem 3.14) for bounded random variables. ♢

Chapter 4

Applications

Large deviations theory is a very active research topic also because it finds applications in a lot of different contexts, both inside and outside of mathematics. The applications we will present here are only given by way of example. Section 4.1 is about ruin probabilities in actuarial risk theory, and it is primarily based on §3.1 of the lecture notes [Pha], concerning applications of large deviations theory in finance and insurance; we also refer to [EKM, §1.1 and §1.2]. For the statistical problem we consider in section 4.2, about the optimization of certain decision tests, we refer to [Hol, ch.VI], [DeZe, §3.4] and [CoTh, §12.7]. Finally, section 4.3 concerns the Curie-Weiss model of ferromagnetism in statistical mechanics, and it is based on the article [Ell] and on the bachelor's thesis [Fav].

4.1 Actuarial risk theory: ruin probabilities

In this section we see how to deduce the classical Cramér-Lundberg inequality for the ruin probability of an insurance company from the properties of moment generating functions we proved in chapter 1 and large deviations techniques we have often used, such as exponential changes of probability measures.

The classical insurance risk model we present here goes back to the early work by Lundberg (1903, see [Lun]), and considers an insurance company with the following data (all random variables are defined on a probability space $(\Omega, \mathcal{E}, \mathbb{P})$):

- The constant rate c of *premium earnings*: the amount of total premiums collected over the time interval $[0,t]$ is ct.
- The *claim sizes* $\{Y_i\}_{i\in\mathbb{N}}$: the size of the ith claim is Y_i. $\{Y_i\}_{i\in\mathbb{N}}$ is a sequence of i.i.d. copies of a random variable Y having finite mean $\overline{y}$.
- The *claim inter-arrival times* $\{\xi_i\}_{i\in\mathbb{N}}$: the first claim occurs after a time interval of length ξ_1, and the time interval between the $(i-1)$th claim and the ith claim

is ξ_i, for any $i > 1$. $\{\xi_i\}_{i\in\mathbb{N}}$ is supposed to be a sequence of i.i.d. copies of a random variable ξ having exponential law[†] with parameter λ, and independent of the sequence $\{Y_i\}_{i\in\mathbb{N}}$.

- The *claim arrival times* $\{T_n\}_{n\geq 0}$: the nth claim occurs at the random instant of time T_n, so that $T_n = \xi_1 + \cdots + \xi_n$ for all $i \in \mathbb{N}$ and $0 < T_1 < \cdots < T_n < \dots$ a.s.. We also set $T_0 \doteq 0$.
- The *claim arrivals* $\{N_t\}_{t\in[0,\infty)}$: the number of claims arriving in $[0,t]$ is N_t. $\{N_t\}_{t\in[0,\infty)}$ is a continuous-time process defined by
$$N_t \doteq \sup\{n \geq 1 : \ T_n \leq t\} = \sum_{n=1}^{\infty} \mathbb{1}_{\{T_n\leq t\}}.$$
- The *total claim amounts* $\{S_t\}_{t\in[0,\infty)}$: $S_t = \sum_{i=1}^{N_t} Y_i$ is the sum of all claim sizes arriving in $[0,t]$ (if $N_t = 0$, we set $S_t \doteq 0$).
- The *risk process* $\{X_t^x\}_{t\in[0,\infty)}$: if $x > 0$ represents the initial capital of the company, $X_t^x \doteq x + ct - S_t$ is the capital of the company at time t.
- The *time of ruin* $\tau_x \doteq \inf\{t \geq 0 : \ X_t^x < 0\}$, that is the first random instant when the capital of the insurance company becomes negative. If we denote the natural filtration of the process $\{X_t^x\}_{t\in[0,\infty)}$ by $\mathcal{F}_t^x \doteq \sigma\{X_s^x, \ 0 \leq s \leq t\}$, τ_x turns to be a $\{\mathcal{F}_t\}_{t\in[0,\infty)}$-stopping time.
- The *probability of ruin* with infinite horizon $\psi(x) \doteq \mathbb{P}(\tau_x < +\infty)$.

We are interested in the estimation of $\psi(x)$; in particular, we wish to compute the smallest initial capital x such that the ruin probability $\psi(x)$ of the insurance company remains below a certain level.

Remark 4.1. By definition, $\{N_t\}_{t\in[0,\infty)}$ turns out to be a Poisson process with parameter λ. In particular, $N_0 = 0$ a.s. and, for any $t > 0$, N_t is a discrete random variable having Poisson distribution with parameter λt, i.e.
$$\mathbb{P}(N_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \qquad \forall k \in \mathbb{N}_0. \qquad \diamondsuit$$

Remark 4.2. We first study the convergence of $\frac{X_t^x}{t}$. Even if S_t is a *random* sum, we can compute the limit of $\frac{S_t}{t}$ by the strong law of large numbers in this case. Since $N_t = \sup\{n \geq 1 : \ T_n \leq t\}$, $N_t = k$ if and only if $T_k \leq t < T_{k+1}$, hence $T_{N_t} \leq t < T_{N_t+1}$ and
$$\frac{1}{N_t}\sum_{i=1}^{N_t} \xi_i \leq \frac{t}{N_t} \leq \frac{1}{N_t}\sum_{i=1}^{N_t+1} \xi_i.$$

[†]Assuming that the claim inter-arrival times are exponentially distributed is very reasonable, since they describe "memoryless" waiting times.

Since $N_t \to \infty$ a.s. as $t \to \infty$, both the left-hand and the right-hand sides converge to $\mathbb{E}(\xi) = \frac{1}{\lambda}$ a.s. by the strong law of large numbers; hence $\frac{t}{N_t} \to \frac{1}{\lambda}$, i.e. $\frac{N_t}{t} \to \lambda$ a.s. as $t \to \infty$. Therefore,

$$\frac{S_t}{t} = \frac{N_t}{t}\frac{1}{N_t}\sum_{i=1}^{N_t} Y_i \xrightarrow{t\to\infty} \lambda\overline{y} \qquad \text{a.s.,}$$

since $\frac{1}{N_t}\sum_{i=1}^{N_t} Y_i \to \overline{y}$ a.s. by the strong law of large numbers. We conclude that

$$\frac{X_t^x}{t} = \frac{x + ct - S_t}{t} \xrightarrow{t\to\infty} c - \lambda\overline{y} \qquad \text{a.s..}$$

◊

Since $\frac{X_t^x}{t} \to c - \lambda\overline{y}$ a.s., if $c - \lambda\overline{y} < 0$ then $X_t^x \to -\infty$ a.s., hence $\psi(x) = 1$ for all x. If $c - \lambda\overline{y} = 0$, it can be shown that $\liminf_{t\to\infty} X_t^x = -\infty$ a.s., and $\psi(x) = 1$ again. It follows that, in order to avoid the sure ruin of the insurance company (with infinite horizon), we have to choose the "premium" c so that $c - \lambda\overline{y} > 0$.

Hypothesis 4.3 (Net profit assumption)**.** We assume that

$$\rho \doteq \frac{c - \lambda\overline{y}}{\lambda\overline{y}} > 0$$

(ρ is called *safety loading*).

By the definition of the risk process, ruin can occur only at the claim times T_n, hence $\psi(x) = \mathbb{P}(\sigma_x < +\infty)$, where σ_x is the discrete stopping time

$$\begin{aligned}\sigma_x &\doteq \inf\{n \geq 1 : \ X_{T_n}^x < 0\} = \inf\{n \geq 1 : \ x + cT_n - S_{T_n} < 0\} \\ &= \inf\left\{n \geq 1 : \ \sum_{i=1}^n Y_i - c\sum_{i=1}^n \xi_i > x\right\} = \inf\{n \geq 1 : \ R_n > x\},\end{aligned}$$

where $R_n \doteq Z_1 + \cdots + Z_n$ is the random walk associated with the sequence $Z_i \doteq Y_i - c\xi_i$, $i \in \mathbb{N}$ (i.i.d. copies of $Z \doteq Y - c\xi$). We note that, under the net profit assumption,

$$\mathbb{E}(Z) = \overline{y} - \frac{c}{\lambda} = \frac{\lambda\overline{y} - c}{\lambda} < 0.$$

Moreover, since Y and ξ are independent, the moment generating function of Z is

$$M_Z(t) = \mathbb{E}(e^{t(Y-c\xi)}) = \mathbb{E}(e^{tY})\mathbb{E}(e^{-ct\xi}) = M_Y(t)M_\xi(-ct) = \begin{cases} M_Y(t)\dfrac{\lambda}{\lambda + ct} & \forall t > -\dfrac{\lambda}{c}, \\ +\infty & \forall t \leq -\dfrac{\lambda}{c}, \end{cases}$$

by the formula we computed in Example 1.28 for the moment generating function of the exponential distribution (M_Y and M_ξ are the moment generating functions of Y and ξ). Since $Y \geq 0$, $M_Y(t) = \mathbb{E}(e^{tY}) \leq 1$ for all $t \leq 0$, and the interior domain of M_Y is unbounded below: let us denote it by $(-\infty, t_+)$. We now make the following additional hypothesis:

Hypothesis 4.4. We assume that Y has a *light-tailed distribution*, i.e. $t_+ \in (0, +\infty]$ and $\lim_{t \nearrow t_+} M_Y(t) = +\infty$.

We remark that the estimates we are going to get holds only under the assumption that Y has a light-tailed distribution; the estimation of $\psi(x)$ for heavy-tailed distributions of Y changes completely (see for instance [EKM, §1.3]). If Hypothesis 4.4 holds, $\mathring{\mathcal{D}}_{M_Z} = \mathring{\mathcal{D}}_{\log M_Z} = \left(-\frac{\lambda}{c}, t_+\right)$ and

$$\log M_Z(t) = \log M_Y(t) + \log \frac{\lambda}{\lambda + ct} \qquad \forall t \in \left(-\frac{\lambda}{c}, t_+\right).$$

Therefore, we see that $\log M_Z$ satisfies the following properties:

- $\log M_Z(0) = \log 1 = 0$, and by Lemma 1.18 (i) $(\log M_Z)'(0) = \mathbb{E}(Z) < 0$ (by the net profit assumption).
- $\log M_Z(t) \to +\infty$ as $t \nearrow t_+$, since $M_Y(t) \to +\infty$ as $t \nearrow t_+$ by the light-tailed hypothesis.
- $\log M_Z$ is continuous and convex on $\mathring{\mathcal{D}}_{\log M_Z}$ by Lemma 1.18 (iii).

It follows that there exists a unique $t^* \in (0, t_+)$ such that $\log M_Z(t^*) = 0$, i.e. $M_Z(t^*) = 1$. In other words, $t = t^*$ is the unique solution of the equation

$$M_Y(t) = \frac{ct}{\lambda} + 1.$$

Definition 4.5. t^* is called *Lundberg exponent*.

Since $\log M_Z$ is increasing in a neighborhood of t^*, $(\log M_Z)'(t^*) > 0$. If as usual (see for example section 1.3) $\mathbb{P}_{t^*}$ denotes the probability measure on $(\Omega, \mathcal{E})$ defined by

$$\frac{d\mathbb{P}_{t^*}}{d\mathbb{P}} = \frac{e^{t^*Z}}{M_Z(t^*)} = e^{t^*Z}, \qquad \text{i.e.} \qquad \frac{d\mathbb{P}}{d\mathbb{P}_{t^*}} = e^{-t^*Z}$$

(recalling that $M_Z(t^*) = 1$), by Lemma 1.18 (i) $\mathbb{E}_{t^*}(Z) = (\log M_Z)'(t^*) > 0$. So we claim that for all x the exit time $\sigma_x = \inf\{n \geq 1 : R_n > x\}$ is $\mathbb{P}_{t^*}$-a.s. finite, i.e. $\mathbb{P}_{t^*}(\sigma_x < +\infty) = 1$, as an easy consequence of the strong law of large numbers. Indeed, if $\omega \in \Omega$ is such that $\frac{R_n(\omega)}{n} \to \mathbb{E}_{t^*}(Z)$, then there exists n such that $\frac{R_n(\omega)}{n} > \frac{x}{n}$, since $\mathbb{E}_{t^*}(Z) > 0$ and $\frac{x}{n} \to 0$. Hence by the strong law of large numbers

$$1 = \mathbb{P}_{t^*}\left(\frac{R_n}{n} \to \mathbb{E}_{t^*}(Z)\right) \leq \mathbb{P}_{t^*}\left(\exists n : \frac{R_n}{n} > \frac{x}{n}\right) = \mathbb{P}_{t^*}(\exists n : R_n > x) = \mathbb{P}_{t^*}(\sigma_x < +\infty),$$

and the claim follows. Thus, to estimate $\psi(x) = \mathbb{P}(\sigma_x < +\infty)$, it is reasonable to make a change of probability measure, passing from $\mathbb{P}$ to $\mathbb{P}_{t^*}$:

$$\begin{aligned}\psi(x) &= \mathbb{P}(\sigma_x < +\infty) = \mathbb{E}(\mathbb{1}_{\{\sigma_x<+\infty\}}) = \sum_{n=1}^{\infty} \mathbb{E}(\mathbb{1}_{\{\sigma_x=n\}}) \\ &= \sum_{n=1}^{\infty} \mathbb{E}_{t^*}\Big(\mathbb{1}_{\{\sigma_x=n\}} \prod_{i=1}^{n} e^{-t^* Z_i}\Big) = \sum_{n=1}^{\infty} \mathbb{E}_{t^*}\Big(\mathbb{1}_{\{\sigma_x=n\}} e^{-t^* R_n}\Big) \\ &= \sum_{n=1}^{\infty} \mathbb{E}_{t^*}\Big(\mathbb{1}_{\{\sigma_x=n\}} e^{-t^* R_{\sigma_x}}\Big) = \mathbb{E}_{t^*}\Big(\mathbb{1}_{\{\sigma_x<+\infty\}} e^{-t^* R_{\sigma_x}}\Big) = \mathbb{E}_{t^*}\Big(e^{-t^* R_{\sigma_x}}\Big) \le e^{-t^* x}.\end{aligned}$$

The third and the seventh equalities are justified by dominated convergence; the change of probability measure in the fourth equality follows from the fact that the event $\{\sigma_x = n\} \in \sigma\{Z_1, \dots, Z_n\}$ since σ_x is a stopping time; the last equality holds because $\mathbb{1}_{\{\sigma_x<+\infty\}} = 1$ $\mathbb{P}_{t^*}$-a.s.; the inequality holds because $t^* > 0$ and $R_{\sigma_x} \ge x$ by definition of σ_x. What we have just deduced is the Cramér-Lundberg inequality for the ruin probability of an insurance company: if t^* is the Lundberg exponent, then for all value x of the initial capital, the estimate

$$\psi(x) \le e^{-t^* x}$$

holds, under the net profit assumption and the hypothesis of light-tailed distribution of the claim sizes. In particular, if we want the ruin probability to remain below a certain level p, it suffices that the initial capital of the insurance company is at least

$$x = \frac{1}{t^*} \log \frac{1}{p}.$$

4.2 Statistics: hypothesis testing

One of the standard problems in statistics is to decide between two alternatives concerning the observed data from a scientific study. For example, in medical testing, one may wish to test whether a new drug is effective or not.

In the simplest case of hypothesis testing, we have to decide between two possible distributions of a sequence of i.i.d. random variables. In this section we will see which is the "optimal" test in such a case, in a suitable meaning, with the help of Cramér's Theorem.

We first explain precisely the statistical problem. Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. copies of a real-valued random variable X with *unknown* law μ; we only *know* that the law μ is either μ_0 or μ_1, where μ_0 and μ_1 are given, hence we have two hypotheses:

$$\begin{aligned}&\text{Hypothesis } H_0: \qquad \mu = \mu_0; \\ &\text{Hypothesis } H_1: \qquad \mu = \mu_1.\end{aligned}$$

From a Bayesian point of view, μ is a random distribution with an *a priori* law on $\{\mu_0, \mu_1\}$, given by the probabilities $p \doteq \mathbb{P}(\mu = \mu_0)$, $1 - p = \mathbb{P}(\mu = \mu_1)$.

Hypothesis 4.6. We make the following assumptions:

(i) $\mu_0 \neq \mu_1$;

(ii) $0 < p = \mathbb{P}(\mu = \mu_0) < 1$ (and hence $0 < \mathbb{P}(\mu = \mu_1) < 1$);

(iii) μ_0 and μ_1 are *equivalent*, i.e. mutually absolutely continuous with respect to each other, with densities $L = \frac{d\mu_1}{d\mu_0}$ and $L^{-1} = \frac{d\mu_0}{d\mu_1}$ (μ_0-a.s. or μ_1-a.s., which is the same).

The first two assumptions are just to avoid trivial cases; the third one is the actual hypothesis we make. For instance, μ_0 and μ_1 may be discrete distributions with probability mass on the same points, or continuous distributions on $\mathbb{R}^d$ with strictly positive densities.

The following definition states what is a decision test.

Definition 4.7. A *decision test* T is defined as a sequence of measurable functions $T_n : \mathbb{R}^n \to \{0,1\}$, with the interpretation that, when the data $X_1 = x_1, \dots, X_n = x_n$ are observed,

$$\begin{aligned} &H_0 \text{ is accepted } (H_1 \text{ is rejected}) \text{ when } T_n(x_1,\dots,x_n) = 0,\\ &H_1 \text{ is accepted } (H_0 \text{ is rejected}) \text{ when } T_n(x_1,\dots,x_n) = 1. \end{aligned}$$

We denote by $\mathcal{T}$ the set of all possible decision tests.

In other words, for any decision test T, T_n gives a method to decide whether to accept the hypothesis H_0 or the hypothesis H_1, based on the observation of the sample $(X_1,\dots,X_n)$. We also note that, by definition, each T_n is the indicator function of a suitable measurable subset of $\mathbb{R}^n$.

To evaluate how good a decision test T is, we have to consider the error probabilities

$$\begin{aligned} \alpha_n(T) &\doteq \mathbb{P}(T_n(X_1,\dots,X_n) = 1 \,|\, \mu = \mu_0),\\ \beta_n(T) &\doteq \mathbb{P}(T_n(X_1,\dots,X_n) = 0 \,|\, \mu = \mu_1). \end{aligned}$$

Namely, α_n is the probability that H_0 is rejected from the test T_n while it is true, and β_n is the probability that H_1 is rejected from the test T_n while it is true. We wish to find a decision test that minimizes the total error probability

$$\varepsilon_n(T) \doteq \alpha_n \mathbb{P}(\mu = \mu_0) + \beta_n \mathbb{P}(\mu = \mu_1) = \alpha_n(T)p + \beta_n(T)(1-p),$$

which is called *Bayesian error probability* of the decision test T. The very natural Neyman-Pearson approach to the identification of optimal decision tests tries to minimize β_n according to a pre-set constraint on α_n (or vice versa). In order to determine such tests, the following definitions will be convenient.

Definition 4.8. We call *likelihood ratios* the densities of the product measures $\mu_1^n = \mu_1 \otimes \cdots \otimes \mu_1$ (n times) with respect to the product measures $\mu_0^n = \mu_0 \otimes \cdots \otimes \mu_0$ (n times), i.e. the measurable functions $L_n \colon \mathbb{R}^n \to [0, +\infty)$ defined by

$$L_n(x_1, \ldots, x_n) \doteq L(x_1) \cdots L(x_n) = \frac{d\mu_1}{d\mu_0}(x_1) \ldots \frac{d\mu_1}{d\mu_0}(x_n)$$

(according to this definition, $L_1 = L$). We also call log-*likelihood ratios* the functions $\log L_n$.

Definition 4.9. A decision test T is called a *Neyman-Pearson test* if there exists a sequence $\gamma_n > 0$ such that

$$T_n(x_1, \ldots, x_n) = \mathbb{1}_{\{L_n(x_1,\ldots,x_n) > \gamma_n\}}.$$

We note that, for a Neyman-Pearson test, T_n is actually a measurable function $\mathbb{R}^n \to \{0, 1\}$, since the L_n are measurable functions.

Lemma 4.10 (Neyman-Pearson)**.** Neyman-Pearson tests are optimal in the following meanings.

(i) Let T be a Neyman-Pearson test; then, for any other decision test T' such that $\alpha_n(T') \le \alpha_n(T)$, one has $\beta_n(T') \ge \beta_n(T)$ (and vice versa, for any other decision test T' such that $\beta_n(T') \le \beta_n(T)$, one has $\alpha_n(T') \ge \alpha_n(T)$). In particular, in the class of all tests with a fixed value of α_n, the Neymann-Pearson tests have the smallest value of β_n (and vice versa).

(ii) The Neyman-Pearson test T given by $T_n \doteq \mathbb{1}_{\{L_n > p/(1-p)\}}$ minimizes the Bayesian error probability, i.e. $\varepsilon_n(T') \ge \varepsilon_n(T)$ for all $n \in \mathbb{N}$ and for any other decision test T'.

Proof. Let T be a Neyman-Pearson test, defined by $T_n = \mathbb{1}_{\{L_n > \gamma_n\}}$ for some $\gamma_n > 0$ and for all $n \in \mathbb{N}$. For any other $T' \in \mathcal{T}$, we have therefore that

$$(T_n - T_n')(L_n - \gamma_n) \ge 0 \qquad \forall n \in \mathbb{N}.$$

This inequality can be proved by observing that:

- On $\{L_n > \gamma_n\}$, $T_n - T_n' = 1 - T_n' \ge 0$ (since T_n' takes values in $\{0, 1\}$) and $L_n - \gamma_n > 0$.
- On $\{L_n \le \gamma_n\}$, $T_n - T_n' = -T_n' \le 0$ (since T_n' takes values in $\{0, 1\}$) and $L_n - \gamma_n \le 0$.

We recall that L_n is the density of μ_1^n with respect to μ_0^n, and that μ_1^n and μ_0^n are the laws of the random vector $(X_1, \ldots, X_n)$ given μ_0 and μ_1 respectively (since the X_i are independent). It follows that, if we integrate the function $(T_n - T'_n)(L_n - \gamma_n)$ over $\mathbb{R}^n$ with respect to the product measure μ_0^n, we obtain

$$\begin{aligned}
0 &\leq \int_{\mathbb{R}^n} (T_n - T'_n)(L_n - \gamma_n)\, d\mu_0^n \\
&= \int_{\mathbb{R}^n} (T_n - T'_n)\, d\mu_1^n - \gamma_n \int_{\mathbb{R}^n} (T_n - T'_n)\, d\mu_0^n \\
&= \mathbb{E}(T_n(X_1,\dots,X_n) \,|\, \mu = \mu_1) - \mathbb{E}(T'_n(X_1,\dots,X_n) \,|\, \mu = \mu_1) \\
&\quad - \gamma_n \Big[\mathbb{E}(T_n(X_1,\dots,X_n) \,|\, \mu = \mu_0) - \mathbb{E}(T'_n(X_1,\dots,X_n) \,|\, \mu = \mu_0)\Big] \\
&= \mathbb{P}(T_n(X_1,\dots,X_n) = 1 \,|\, \mu = \mu_1) - \mathbb{P}(T'_n(X_1,\dots,X_n) = 1 \,|\, \mu = \mu_1) \\
&\quad - \gamma_n \Big[\mathbb{P}(T_n(X_1,\dots,X_n) = 1 \,|\, \mu = \mu_0) - \mathbb{P}(T'_n(X_1,\dots,X_n) = 1 \,|\, \mu = \mu_0)\Big] \\
&= (1 - \beta_n(T)) - (1 - \beta_n(T')) - \gamma_n[\alpha_n(T) - \alpha_n(T')] \\
&= [\beta_n(T') - \beta_n(T)] + \gamma_n[\alpha_n(T') - \alpha_n(T)],
\end{aligned}$$

hence

$$\beta_n(T') - \beta_n(T) \geq -\gamma_n(\alpha_n(T') - \alpha_n(T)). \tag{4.1}$$

We have therefore:

(i) By (4.1), since $\gamma_n > 0$, if $\alpha_n(T') \leq \alpha_n(T)$ then $\beta_n(T') \geq \beta_n(T)$, and vice versa if $\beta_n(T') \leq \beta_n(T)$ then $\alpha_n(T') \geq \alpha_n(T)$.

(ii) By (4.1), the difference between the Bayesian error probabilities of T' and T is

$$\begin{aligned}
\varepsilon_n(T') - \varepsilon_n(T) &= [\alpha_n(T')p + \beta_n(T')(1-p)] - [\alpha_n(T)p + \beta_n(T)(1-p)] \\
&= [\alpha_n(T') - \alpha_n(T)]p + [\beta_n(T') - \beta_n(T)](1-p) \\
&\geq [\alpha_n(T') - \alpha_n(T)]p - [\alpha_n(T') - \alpha_n(T)]\gamma_n(1-p) \\
&= [p - \gamma_n(1-p)][\alpha_n(T') - \alpha_n(T)].
\end{aligned}$$

Choosing $\gamma_n \doteq p/(1-p)$, i.e. considering the Neyman-Pearson test T defined by $T_n \doteq \mathbb{1}_{\{L_n > p/(1-p)\}}$, we get $\varepsilon_n(T') - \varepsilon_n(T) \geq 0$ for all n. The claim follows since $T' \in \mathcal{T}$ is arbitrary. □

We now define the new random variables

$$Y_i \doteq \log L(X_i) = \log \frac{d\mu_1}{d\mu_0}(X_i) = -\log \frac{d\mu_0}{d\mu_1}(X_i),$$

which are i.i.d. copies of the random variable $Y \doteq \log L(X)$. According to whether the law of X is μ_0 or μ_1, the mean of Y is one of the following:

$$\begin{aligned}
\overline{y}_0 &\doteq \mathbb{E}(Y \,|\, \mu = \mu_0) = \int_{\mathbb{R}} \log\left(\frac{d\mu_1}{d\mu_0}\right) d\mu_0 = -\int_{\mathbb{R}} \log\left(\frac{d\mu_0}{d\mu_1}\right) d\mu_0 = -H(\mu_0|\mu_1), \\
\overline{y}_1 &\doteq \mathbb{E}(Y \,|\, \mu = \mu_1) = \int_{\mathbb{R}} \log\left(\frac{d\mu_1}{d\mu_0}\right) d\mu_1 = H(\mu_1|\mu_0),
\end{aligned}$$

where $H(\cdot|\cdot)$ denotes the relative entropy, according to Definition 3.17. Since $\mu_0 \neq \mu_1$, $H(\mu_0|\mu_1) > 0$ and $H(\mu_1|\mu_0) > 0$ by Proposition 3.19 (i), hence we have that

$$-\infty \leq \overline{y}_0 < 0 < \overline{y}_1 \leq +\infty.$$

We may give another simple characterization of Neyman-Pearson tests in terms of log-likelihood ratios.

Lemma 4.11. A decision test T is a Neyman-Pearson test if and only if there exists a real sequence δ_n such that

$$T_n(X_1,\dots,X_n) = \mathbb{1}_{\{\frac{1}{n}(Y_1+\dots+Y_n)>\delta_n\}}, \tag{4.2}$$

where $Y_i \doteq \log L(X_i)$ for all $i \geq 1$.

Proof. By definition, T is a Neyman-Pearson test if there exists a sequence of positive constants γ_n such that

$$T_n(X_1,\dots,X_n) = \mathbb{1}_{\{L_n(X_1,\dots,X_n)>\gamma_n\}}.$$

It now suffices to note that the following inequalities are equivalent:

$$\begin{aligned} L_n(X_1,\dots,X_n) &> \gamma_n, \\ \log(L(X_1)\cdots L(X_n)) &> \log\gamma_n, \\ \frac{1}{n}(\log L(X_1)+\cdots+\log L(X_n)) &> \frac{\log\gamma_n}{n}, \\ \frac{1}{n}(Y_1+\cdots+Y_n) &> \delta_n, \end{aligned}$$

having set $\delta_n \doteq \log\gamma_n/n$. □

If in (4.2) δ_n is constant, it turns out that T_n is the indicator function of a large deviations event, in the sense of Cramér's Theorem. As a consequence, in this case the error probabilities α_n and β_n decay exponentially with rate n, as we show in the following theorem.

Theorem 4.12. Let $\delta \in (\overline{y}_0, \overline{y}_1)$ and let T be the Neyman-Pearson test defined by

$$T_n(X_1,\dots,X_n) = \mathbb{1}_{\{\frac{1}{n}(Y_1+\dots+Y_n)>\delta\}}.$$

Then

$$\lim_{n\to\infty}\frac{1}{n}\log\alpha_n(T) = -I_0(\delta) \in (-\infty,0), \tag{4.3}$$

$$\lim_{n\to\infty}\frac{1}{n}\log\beta_n(T) = -[I_0(\delta)-\delta] \in (-\infty,0), \tag{4.4}$$

where I_0 is the Legendre transform of the logarithmic moment generating function of Y given $\mu = \mu_0$, i.e.

$$I_0(y) = \sup_{t\in\mathbb{R}}[ty - \log M_0(t)], \qquad M_0(t) = \mathbb{E}(e^{tY} \,|\, \mu = \mu_0).$$

Proof. We first prove that $\delta \in \mathring{\mathcal{D}}_{I_0}$. We have that

$$M_0(1) = \mathbb{E}(e^Y \,|\, \mu = \mu_0) = \int_{\mathbb{R}} \frac{d\mu_1}{d\mu_0}\, d\mu_0 = \int_{\mathbb{R}} d\mu_1 = 1 < +\infty.$$

Since also $0 \in \mathcal{D}_{M_0}$ and $\mathcal{D}_{M_0}$ is an interval, $[0,1] \subseteq \mathcal{D}_{M_0}$. Hence, by Lemma 1.18 (i)

$$(\log M_0)'(t) = \mathbb{E}\left(Y \frac{e^{tY}}{M_0(t)} \,\middle|\, \mu = \mu_0\right) \qquad \forall t \in (0,1).$$

Using the latter equality, the equality $M_0(1) = 1$, the continuity of M_0 on $[0,1] \subseteq \mathcal{D}_{M_0}$ and dominated and monotone convergence, we now write $\overline{y}_0$ and $\overline{y}_1$ as

$$\begin{aligned}
\overline{y}_0 &= \mathbb{E}(Y \,|\, \mu = \mu_0) = \lim_{t\searrow 0} \mathbb{E}\left(Y \frac{e^{tY}}{M_0(t)} \,\middle|\, \mu = \mu_0\right) = \lim_{t\searrow 0} (\log M_0)'(t),\\
\overline{y}_1 &= \mathbb{E}(Y \,|\, \mu = \mu_1) = \int_{\mathbb{R}} \log(L)\, d\mu_1 = \int_{\mathbb{R}} \log(L) \frac{d\mu_1}{d\mu_0}\, d\mu_0\\
&= \mathbb{E}(Ye^Y \,|\, \mu = \mu_0) = \lim_{t\nearrow 1} \mathbb{E}\left(Y \frac{e^{tY}}{M_0(t)} \,\middle|\, \mu = \mu_0\right) = \lim_{t\nearrow 1} (\log M_0)'(t).
\end{aligned}$$

By hypothesis $\overline{y}_0 < \delta < \overline{y}_1$, i.e.

$$\lim_{t\searrow 0} (\log M_0)'(t) < \delta < \lim_{t\nearrow 1} (\log M_0)'(t),$$

hence in particular

$$\lim_{t\searrow \inf \mathcal{D}_{M_0}} (\log M_0)'(t) < \delta < \lim_{t\nearrow \sup \mathcal{D}_{M_0}} (\log M_0)'(t),$$

since $(\log M_0)'$ is increasing on $\mathcal{D}_{M_0}$. Just as in the proof of Proposition 1.14 (ii), one can prove that

$$\left(\lim_{t\searrow \inf \mathcal{D}_{M_0}} (\log M_0)'(t), \lim_{t\nearrow \sup \mathcal{D}_{M_0}} (\log M_0)'(t)\right) \subseteq \mathring{\mathcal{D}}_{I_0}$$

(equality does not need to hold if $\log M_0$ is not steep). We conclude that $\delta \in \mathring{\mathcal{D}}_{I_0}$. In particular, I_0 is continuous at δ, hence $(\delta, +\infty)$ is I_0-continuous in the sense of Remark 2.32. We can therefore apply the generalized Cramér's Theorem 3.1 (see also Remark 3.3) to the sequence $\{Y_i\}_{i\in\mathbb{N}}$, without distinguishing between limit superior and limit inferior:

$$\lim_{n\to\infty}\frac{1}{n}\log\alpha_n(T) = \lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(T_n(X_1,\dots,X_n)=1\,|\,\mu=\mu_0)$$
$$=\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\frac{1}{n}(Y_1+\cdots+Y_n)>\delta\,\Big|\,\mu=\mu_0\Big) = -\inf_{y>\delta} I_0(y) = -I_0(\delta),$$

where the latter equality follows from the fact that I_0 is increasing on $(\overline{y}_0,+\infty)$ and continuous at δ. Since M_0 is finite in a right-neighborhood of 0 and $\delta>\overline{y}_0$, we have that $I_0(\delta)>0$ (see Remark 1.21). Moreover, $I_0(\delta)<+\infty$ since $\delta\in\mathring{\mathcal{D}}_{M_0}$. This concludes the proof of (4.3).

We now consider the moment generating function M_1 of Y given $\mu=\mu_1$, and its Legendre transform I_1.

$$M_1(t)=\mathbb{E}(e^{tY}\,|\,\mu=\mu_1)=\int_{\mathbb{R}} e^{t\log(L)}\,d\mu_1=\int_{\mathbb{R}} e^{t\log(L)}\frac{d\mu_1}{d\mu_0}\,d\mu_0$$
$$=\int_{\mathbb{R}} e^{t\log(L)+\log(L)}\,d\mu_0=\mathbb{E}(e^{(t+1)Y}\,|\,\mu=\mu_0)=M_0(t+1),$$

and in particular M_1 is finite in $[-1,0]$, since M_0 is finite in $[0,1]$. Moreover,

$$I_1(y)=\sup_{t\in\mathbb{R}}[ty-\log M_1(t)]=\sup_{t\in\mathbb{R}}[ty-\log M_0(t+1)]$$
$$=\sup_{t\in\mathbb{R}}[(t+1)y-\log M_0(t+1)]-y=I_0(y)-y.$$

Recalling that $\delta<\overline{y}_1$, we can apply Cramér's Theorem 1.20 (see also Remark 1.21) to the sequence $\{Y_i\}_{i\in\mathbb{N}}$:

$$\lim_{n\to\infty}\frac{1}{n}\log\beta_n(T)=\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}(T_n(X_1,\dots,X_n)=0\,|\,\mu=\mu_1)$$
$$=\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\Big(\frac{1}{n}(Y_1+\cdots+Y_n)\le\delta\,\Big|\,\mu=\mu_1\Big)=-I_1(\delta)=-[I_0(\delta)-\delta].$$

Since M_1 is finite in a left-neighborhood of 0 and $\delta<\overline{y}_1$, we have that $I_1(\delta)>0$ (see Remark 1.21). Moreover, $I_1(\delta)<+\infty$ since $I_0(\delta)<+\infty$. This concludes the proof of (4.4). □

In the light of the latter asymptotic result, we now wish to determine a decision test that minimizes the Bayesian error probability asymptotically (i.e. in the limit as $n\to\infty$). By Lemma 4.10, we already know that the choice $\gamma_n \doteq p/(1-p)$ yields a Neyman-Pearson test $T_n\doteq\mathbb{1}_{\{L_n>p/(1-p)\}}$ that is optimal, in the sense of minimizing ε_n, for any fixed n. Since for such a choice of γ_n (in general, for any choice of a constant sequence $\gamma_n>0$) $\delta_n\doteq\log\gamma_n/n\to 0$ as $n\to\infty$, it is natural to think that an asymptotically optimal choice is $\delta_n\doteq 0$, which yields the Neyman-Pearson test $\mathbb{1}_{\{\frac{1}{n}(Y_1+\cdots+Y_n)>0\}}$. The following theorem makes rigorous this heuristic argument.

Theorem 4.13. For any $p = \mathbb{P}(\mu = \mu_0) \in (0,1)$ one has

$$\inf_{T\in\mathcal{T}} \liminf_{n\to\infty} \frac{1}{n} \log \varepsilon_n(T) = -I_0(0) \in (-\infty, 0), \tag{4.5}$$

and the infimum is attained at the Neyman-Pearson test T^* defined by

$$T_n^*(X_1,\dots,X_n) \doteq \mathbb{1}_{\{\frac{1}{n}(Y_1+\dots Y_n)>0\}} = \mathbb{1}_{\{L_n(X_1,\dots,X_n)>1\}}. \tag{4.6}$$

Proof. We start by observing that the two definition of T^* in (4.6) are equivalent by Lemma 4.11, since $0 = (\log 1)/n$.

To compute the infimum, it suffices to consider Neyman-Pearson tests, since they minimize ε_n by Lemma 4.10. Let T be any other Neyman-Pearson test, defined by

$$T_n(X_1,\dots,X_n) \doteq \mathbb{1}_{\{\frac{1}{n}(Y_1+\dots+Y_n)>\delta_n\}}$$

for some real sequence δ_n. For all $n \geq 1$, either $\alpha_n(T^*) \leq \alpha_n(T)$ (if $\delta_n \leq 0$) or $\beta_n(T^*) \leq \beta_n(T)$ (if $\delta_n \geq 0$). It follows that

$$\varepsilon_n(T) = \alpha_n(T)p + \beta_n(T)(1-p) \geq [\alpha_n(T^*)p] \wedge [\beta_n(T^*)(1-p)].$$

Since $p > 0$ and $1 - p > 0$,

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n} \log \varepsilon_n(T) \\
\geq &\liminf_{n\to\infty} \frac{1}{n} \log[\alpha_n(T^*)p] \wedge \liminf_{n\to\infty} \frac{1}{n} \log[\beta_n(T^*)(1-p)] \\
= &\liminf_{n\to\infty} \left[\frac{1}{n} \log \alpha_n(T^*) + \frac{1}{n} \log p\right] \wedge \liminf_{n\to\infty} \left[\frac{1}{n} \log \beta_n(T^*) + \frac{1}{n} \log(1-p)\right] \\
= &\liminf_{n\to\infty} \left[\frac{1}{n} \log \alpha_n(T^*)\right] \wedge \liminf_{n\to\infty} \left[\frac{1}{n} \log \beta_n(T^*)\right] \\
= &[-I_0(0)] \wedge [-I_0(0)] = -I_0(0),
\end{aligned}$$

where the second to last equality follows from Theorem 4.12, with $\delta = 0 \in (\overline{y}_0, \overline{y}_1)$. Moreover, if $T = T^*$, using (1.19) we obtain

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n} \log \varepsilon_n(T^*) \\
= &\liminf_{n\to\infty} \frac{1}{n} \log[\alpha_n(T^*)p + \beta_n(T^*)(1-p)] \\
= &\liminf_{n\to\infty} \left[\frac{1}{n} \log[\alpha_n(T^*)p] \vee \frac{1}{n} \log[\beta_n(T^*)(1-p)] + o(1)\right] \\
= &\liminf_{n\to\infty} \frac{1}{n} \log[\alpha_n(T^*)p] \vee \liminf_{n\to\infty} \frac{1}{n} \log[\beta_n(T^*)(1-p)] \\
= &[-I_0(0)] \vee [-I_0(0)] = -I_0(0).
\end{aligned}$$

This proves that the infimum in (4.5) is attained at $T = T^*$, and it is equal to $-I_0(0)$. We also note that, by Theorem (4.12) again, $-I_0(0) \in (-\infty, 0)$. □

The latter theorem shows that the asymptotically best choice for a decision test is the so-called *zero threshold* Neyman-Pearson test, meaning that its Bayesian error probability decay is as fast as possible. We conclude this section with an example of computation of the optimal decay exponential rate $I_0(0)$.

Example 4.14. We want to compute $I_0(0)$ where μ_0 and μ_1 are two exponential distributions with parameter $\lambda_0 > 0$ and $\lambda_1 > 0$ respectively (they are equivalent since they are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, with the same support $[0, +\infty)$). We assume without loss of generality that $\lambda_0 < \lambda_1$ and we set $\lambda \doteq \lambda_1/\lambda_0 > 1$. The log-likelihood ratio $\log L$ is

$$\log L(x) = \log\Big(\frac{d\mu_1}{d\mu_0}(x)\Big) = \log\Big(\frac{\lambda_1 e^{-\lambda_1 x}}{\lambda_0 e^{-\lambda_0 x}}\Big) = \log\Big(\frac{\lambda_1}{\lambda_0}\Big) - (\lambda_1 - \lambda_0)x,$$

for all $x \in [0, +\infty)$; since $\operatorname{supp}(X) = [0, +\infty)$, $\operatorname{supp}(Y) = \operatorname{supp}(\log L(X)) = (-\infty, \log\lambda]$. The moment generating function of Y given $\mu = \mu_0$ is

$$\begin{aligned} M_0(t) &= \mathbb{E}(e^{tY} \,|\, \mu = \mu_0) = \mathbb{E}(e^{t\log L(X)} \,|\, \mu = \mu_0) = \Big(\frac{\lambda_1}{\lambda_0}\Big)^t \mathbb{E}(e^{-(\lambda_1-\lambda_0)tX} \,|\, \mu = \mu_0) \\ &= \begin{cases} \Big(\dfrac{\lambda_1}{\lambda_0}\Big)^t \dfrac{\lambda_0}{\lambda_0 + (\lambda_1 - \lambda_0)t} & \text{if } -(\lambda_1 - \lambda_0)t < \lambda_0, \\ +\infty & \text{otherwise} \end{cases} \\ &= \begin{cases} \dfrac{\lambda^t}{1 + (\lambda - 1)t} & \text{if } t > -\dfrac{1}{\lambda - 1}, \\ +\infty & \text{otherwise,} \end{cases} \end{aligned}$$

having used formula (1.22) for the moment generating function of an exponential distribution with parameter λ_0. For all $t \in \mathring{\mathcal{D}}_{M_0} = (-(\lambda - 1)^{-1}, +\infty)$,

$$\log M_0(t) = t\log\lambda - \log(1 + (\lambda - 1)t), \qquad (\log M_0)'(t) = \log\lambda - \frac{\lambda - 1}{1 + (\lambda - 1)t}.$$

We have $0 \in (-\infty, \log\lambda) = \mathring{\mathcal{D}}_{I_0}$ (since $\lambda > 1$), and $0 = (\log M_0)'(t)$ if and only if

$$0 = \log\lambda - \frac{\lambda - 1}{1 + (\lambda - 1)t} \qquad \Longleftrightarrow \qquad t = \frac{1}{\log\lambda} - \frac{1}{\lambda - 1}.$$

By (1.4), $I_0(0)$ is given by

$$\begin{aligned} I_0(0) &= \Big(\frac{1}{\log\lambda} - \frac{1}{\lambda - 1}\Big)0 - \log M_0\Big(\frac{1}{\log\lambda} - \frac{1}{\lambda - 1}\Big) \\ &= -\Big(\frac{1}{\log\lambda} - \frac{1}{\lambda - 1}\Big)\log\lambda + \log\Big(1 + (\lambda - 1)\Big(\frac{1}{\log\lambda} - \frac{1}{\lambda - 1}\Big)\Big) \\ &= \frac{\log\lambda}{\lambda - 1} - \log\Big(\frac{\log\lambda}{\lambda - 1}\Big) - 1. \end{aligned}$$

◊

4.3 Statistical mechanics: the Curie-Weiss model of ferromagnetism

In the so-called paramagnetic materials, a magnetic moment is induced by an externally applied magnetic field. On the other hand, ferromagnetism is a physical mechanism by which certain materials, such as iron, cobalt and nickel, form permanent magnets, or are permanently attracted to magnets. In other words, they show a spontaneous magnetization, even in the absence of an external magnetic field, given to the alignment of intrinsic magnetic moments. However, such an alignment only occurs below a certain critical temperature, called the Curie temperature. Above the material's specific Curie temperature, one can observe a *phase transition*: thermal motion obstructs the natural alignment of intrinsic magnetic moments, which therefore change direction and become randomly oriented; as a consequence, the material can no longer maintain a spontaneous magnetization, although it still has a paramagnetic behavior in the presence of an external field. In this section we will see how large deviations results (in particular, Cramér's Theorem and the Varadhan-Bryc Theorem) can explain the phase transition of a ferromagnetic material in the simplest example of interacting system: the Curie-Weiss probabilistic model, in the absence of an external magnetic field.

To describe the model, we consider the discrete probability space $\Omega \doteq \{-1,1\}$, equipped with the uniform probability $\mathbb{P}$ (i.e., $\mathbb{P}(\{-1\}) = \mathbb{P}(\{1\}) = 2^{-1}$): Ω is interpreted as the set of the two possible directions in which the *spin* (or magnetic moment) may align, with the same probability; the value -1 represents *spin-down* and the value 1 *spin-up*. For any $n \in \mathbb{N}$, we define the discrete configuration space

$$\Omega^n \doteq \{-1,1\}^n = \Big\{\omega = (\omega_1,\dots,\omega_n) : \quad \omega_i \in \{-1,1\} \ \forall i = 1,\dots,n\Big\},$$

where any element ω is interpreted as a possible configuration of an n-dimensional spin system; we consider on Ω^n the product measure $\mathbb{P}^n \doteq \mathbb{P}\otimes\cdots\otimes\mathbb{P}$ (n times), which corresponds to the uniform probability on Ω^n (i.e., $\mathbb{P}^n(\omega) = 2^{-n}$ for all $\omega \in \Omega^n$). Thus the projections X_i, defined by $X_i(\omega) \doteq \omega_i$ for all $i = 1,\dots,n$ and $\omega \in \Omega^n$, turn out to be i.i.d. Ω-valued random variables, since the law of the random vector $(X_1,\dots,X_n)$ is $\mathbb{P}^n$. If we define

$$S_n(\omega) \doteq \sum_{i=1}^{n} X_i(\omega) = \sum_{i=1}^{n} \omega_i \qquad\qquad \forall \omega \in \Omega^n,$$

the random variable $\frac{S_n}{n}$ represents the so-called *average spin per site*, that is the macroscopical average configuration of the spin system.

In a model of ferromagnetism, the probability of the spin configurations must depend on the temperature of the material; as usual in this context, we will consider the parameter $\beta > 0$, which is supposed to be proportional to the *inverse* temperature. The Curie-Weiss model consists in choosing the probability measure $\mathbb{P}_{n,\beta}$ on Ω^n, depending on β and equivalent to $\mathbb{P}^n$, defined by

$$\frac{d\mathbb{P}_{n,\beta}}{d\mathbb{P}^n} \doteq \frac{1}{Z_{n,\beta}} \exp(\beta H_n), \tag{4.7}$$

where for all $\omega \in \Omega^n$

$$H_n(\omega) \doteq \frac{1}{2n}\sum_{i,j=1}^n \omega_i\omega_j = \frac{1}{2n}\left(\sum_{i=1}^n \omega_i\right)^2 = \frac{1}{2n}S_n(\omega)^2 = \frac{n}{2}\left(\frac{S_n(\omega)}{n}\right)^2 \tag{4.8}$$

and $Z_{n,\beta}$ is the normalization factor that makes $\mathbb{P}_{n,\beta}$ a probability measure, i.e.

$$Z_{n,\beta} \doteq \int_{\Omega_n} \exp(\beta H_n)\, d\mathbb{P}^n = \sum_{\omega\in\Omega_n} \exp(\beta H_n(\omega))\frac{1}{2^n}. \tag{4.9}$$

The value $H_n(\omega)$ represents the energy of the configuration ω and the $\mathbb{R}^+$-valued random variable H_n is called *Hamiltonian* and represents the energy of the system. The probability $\mathbb{P}_{n,\beta}$ is called *Gibbs measure* relative to the Hamiltonian H_n, and $Z_{n,\beta}$ is known as *partition function*. The Gibbs measures, also describing much more complicated models, are standard in statistical mechanics.

Remark 4.15. The following simple observations are intended to comment and justify the choice of the Hamiltonian H_n for a model of ferromagnetism.

- By the definition of the Hamiltonian,

$$\begin{aligned} H_n &= \frac{1}{2n}\left(\sum_{i=1}^n \omega_i\right)^2 = \frac{1}{2n}\left(\sum_{i=1}^n \omega_i^2 + \sum_{1\le i\ne j\le n} \omega_i\omega_j\right) \\ &= \frac{1}{2n}\left(n + \sum_{i=1}^n \omega_i \sum_{j\ne i} \omega_j\right) = \frac{1}{2} + \frac{n-1}{n}\left[\frac{1}{2}\sum_{i=1}^n \omega_i \frac{\sum_{j\ne i}\omega_j}{n-1}\right], \end{aligned}$$

  having used that $\sum_{i^1}^n \omega_i^2 = n$ since $\omega_i = \pm 1$. If we assume that any spin ω_i only feels the effect of the average $(\sum_{j\ne i}\omega_j)/(n-1)$ of the other spins†, the number $\omega_i(\sum_{j\ne i}\omega_j)/(n-1)$ represents the interaction between ω_i and the other spins. Therefore, the number

$$\frac{1}{2}\sum_{i=1}^n \omega_i \frac{\sum_{j\ne i}\omega_j}{n-1}$$

  is the sum of all these interactions (the factor 1/2 is just to not count twice the mutual interactions), that is the total energy of the system. Since $(n-1)/n$ is approximately 1 when n is large, this argument shows how, up to an additive constant, H_n may be actually interpreted as the energy of the system (up to a sign), in a mean field model.

†This hypothesis is known as *mean field approximation*, and it is often used in mathematical modeling of physical systems. Although it does not precisely correspond to the physical reality, it may be useful to study some aspects of certain phenomena.

- Since $S_n = X_1 + \dots X_n$ takes values in $\{-n, -n+1, \dots, n-1, n\}$, by (4.8) and (4.7) H_n and $\mathbb{P}_{n,\beta}$ attain their strict maximum at $\omega = \pm(1, \dots, 1)$. This means that the most likely single configurations correspond to the alignment of all spins in one of the possible two directions. ♢

We now study the asymptotic distribution of the average spin per site $\frac{S_n}{n}$ with respect to the Gibbs measure $\mathbb{P}_{n,\beta}$: in the limit as $n \to \infty$, we will observe alignment effects again. This is the line of reasoning we will follow to determine the weak limit of the measures $\mathbb{P}_{n,\beta}\left(\frac{S_n}{n} \in \cdot\right)$: since for any n the random variables $X_1, \dots, X_n$ are i.i.d. with respect to $\mathbb{P}^n$ with uniform law on $\{-1, 1\}$, it is immediate to gather from Cramér's Theorem that the laws of $\frac{S_n}{n}$ with respect to $\mathbb{P}^n$ satisfy a LDP with good rate function I, the Legendre transform of the logarithmic moment generating functions of X_i. Thanks to the Varadhan-Bryc Theorem, we will show that the laws of $\frac{S_n}{n}$ with respect to $\mathbb{P}_{n,\beta}$ also satisfy a LDP with a certain good rate function I_β. Finally, from the analysis of I_β we will easily deduce the weak limit of these measures. First of all, we need to compute the rate function I (namely, the Legendre transform of the logarithmic moment generating function) for the uniform distribution on $\{-1, 1\}$.

Example 4.16. Let X be a real-valued random variable with uniform law on $\{-1, 1\}$. Then the rate function I has interior domain $\mathring{\mathcal{D}}_I = (\inf(\text{supp}(X)), \sup(\text{supp}(X))) = (-1, 1)$, and the moment generating function is

$$M(t) = \mathbb{E}(e^{tX}) = \frac{1}{2}(e^t + e^{-t}) = \cosh(t).$$

Therefore $\mathring{\mathcal{D}}_M = \mathbb{R}$ and

$$\log M(t) = \log(\cosh(t)), \qquad (\log M)'(t) = \frac{\sinh(t)}{\cosh(t)} = \tanh(t).$$

For $x \in (-1, 1)$, $x = (\log M)'(t)$ if and only if

$$t = \operatorname{arctanh}(x) = \frac{1}{2} \log\left(\frac{1+x}{1-x}\right).$$

Recalling the hyperbolic trigonometric equalities

$$\cosh(t)^2 - \sinh(t)^2 = 1 \qquad \Longrightarrow \qquad \cosh(t)^2 = \frac{1}{1 - \tanh(t)^2},$$

we obtain that for $x \in \mathring{\mathcal{D}}_I = (-1,1)$

$$\begin{aligned}
I(x) &= \operatorname{arctanh}(x)x - \log M(\operatorname{arctanh}(x)) \\
&= \frac{1}{2}x\log\left(\frac{1+x}{1-x}\right) - \log(\cosh(\operatorname{arctanh}(x))) \\
&= \frac{1}{2}x\log\left(\frac{1+x}{1-x}\right) - \log\left(\frac{1}{\sqrt{1-x^2}}\right) \\
&= \frac{1}{2}x\log(1+x) - \frac{1}{2}x\log(1-x) + \frac{1}{2}\log(1+x) + \frac{1}{2}\log(1-x) \\
&= \frac{1}{2}(1+x)\log(1+x) + \frac{1}{2}(1-x)\log(1-x).
\end{aligned}$$

We conclude that

$$I(x) = \begin{cases} \frac{1}{2}(1+x)\log(1+x) + \frac{1}{2}(1-x)\log(1-x) & x \in [-1,1], \\ +\infty & x \in \mathbb{R} \setminus [-1,1], \end{cases} \tag{4.10}$$

with the convention that $0\log 0 = 0$. ♢

Theorem 4.17. For any $\beta > 0$, the sequence of probability measures $\mathbb{P}_{n,\beta}\left(\frac{S_n}{n} \in \cdot\right)$ on $[-1,1]$ satisfies the LDP with rate n and good rate function I_β given by

$$I_\beta(x) \doteq I(x) - \frac{\beta}{2}x^2 - \inf_{y\in[-1,1]}\left[I(y) - \frac{\beta}{2}y^2\right] \qquad \forall x \in [-1,1], \tag{4.11}$$

where

$$I(x) \doteq \frac{1}{2}(1+x)\log(1+x) + \frac{1}{2}(1-x)\log(1-x) \qquad \forall x \in [-1,1].$$

Proof. Since S_n takes values in $\{-n,\dots,n\}$, $\frac{S_n}{n}$ takes values in $[-1,1]$, hence $\mathbb{P}_{n,\beta}\left(\frac{S_n}{n} \in \cdot\right)$ may be considered as a probability measure on $[-1,1]$. To prove the LDP, by the Varadhan-Bryc Theorem (Theorem 2.43) it suffices to show that I_β is a good rate function and

$$\lim_{n\to\infty} \frac{1}{n}\log\int_{[-1,1]} \exp(nF(x))\mathbb{P}_{n,\beta}\left(\frac{S_n}{n} \in dx\right) = \sup_{x\in[-1,1]}[F(x) - I_\beta(x)] \tag{4.12}$$

for any $F\colon [-1,1] \to \overline{\mathbb{R}}$ continuous and bounded from above.

We first note that, if $(\Omega', \mathcal{E}', \mathbb{P}')$ is any probability space and $\{X_i'\}_{i\in\mathbb{N}}$ is a sequence of real-valued i.i.d. random variable with uniform law on $\{-1,1\}$, and $\frac{S_n'}{n}$ is their empirical average, then for any $n \in \mathbb{N}$ the law of $\frac{S_n'}{n}$ with respect to $\mathbb{P}'$ correspond to the laws of $\frac{S_n}{n}$ with respect to $\mathbb{P}^n$, i.e.

$$\mu_n(dx) \doteq \mathbb{P}'\left(\frac{S_n'}{n} \in dx\right) = \mathbb{P}_n\left(\frac{S_n}{n} \in dx\right),$$

since $X_1, \dots, X_n$ are i.i.d. with uniform law on $\{-1,1\}$, with respect to $\mathbb{P}_n$. By applying Cramér's Theorem 3.1, the sequence μ_n satisfy the LDP with rate n and rate function I given by (4.10).

Recalling (4.7), (4.8) and (4.9), for any fixed $F\colon [-1,1] \to \overline{\mathbb{R}}$ continuous and bounded from above

$$\begin{aligned}
&\int_{[-1,1]} \exp(nF(x))\mathbb{P}_{n,\beta}\Big(\frac{S_n}{n} \in dx\Big) = \int_{\Omega^n} \exp\Big(nF\Big(\frac{S_n}{n}\Big)\Big) d\mathbb{P}_{n,\beta} \\
&= \int_{\Omega^n} \exp\Big(nF\Big(\frac{S_n}{n}\Big)\Big) Z_{n,\beta}^{-1} \exp\Big(\beta \frac{n}{2}\Big(\frac{S_n}{n}\Big)^2\Big) d\mathbb{P}^n \\
&= \Big(\int_{\Omega^n} \exp\Big(\beta\frac{n}{2}\Big(\frac{S_n}{n}\Big)^2\Big) d\mathbb{P}^n\Big)^{-1} \Big(\int_{\Omega^n} \exp\Big(nF\Big(\frac{S_n}{n}\Big) + \beta\frac{n}{2}\Big(\frac{S_n}{n}\Big)^2\Big) d\mathbb{P}^n\Big) \\
&= \Big(\int_{[-1,1]} \exp\Big(n\frac{\beta}{2}y^2\Big)\mu_n(dy)\Big)^{-1} \Big(\int_{[-1,1]} \exp\Big(n\Big(F(x) + \frac{\beta}{2}x^2\Big)\Big)\mu_n(dx)\Big).
\end{aligned}$$

Since the functions $y \to \frac{\beta}{2}y^2$ and $x \to F(x) + \frac{\beta}{2}x^2$ are continuous and bounded from above on $[-1,1]$, by the Varadhan-Bryc Theorem the LDP satisfied by μ_n suffices to conclude that

$$\lim_{n\to\infty} \frac{1}{n} \log \int_{[-1,1]} \exp\Big(n\frac{\beta}{2}y^2\Big)\mu_n(dy) = \sup_{y\in[-1,1]} \Big[\frac{\beta}{2}y^2 - I(y)\Big],$$
$$\lim_{n\to\infty} \frac{1}{n} \log \int_{[-1,1]} \exp\Big(n\Big(F(x) + \frac{\beta}{2}x^2\Big)\Big)\mu_n(dx) = \sup_{x\in[-1,1]} \Big[F(x) + \frac{\beta}{2}x^2 - I(x)\Big].$$

Therefore,

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n} \log \int_{[-1,1]} \exp(nF(x))\mathbb{P}_{n,\beta}\Big(\frac{S_n}{n} \in dx\Big) \\
&= -\lim_{n\to\infty} \frac{1}{n} \log \int_{[-1,1]} \exp\Big(n\frac{\beta}{2}y^2\Big)\mu_n(dy) + \lim_{n\to\infty} \frac{1}{n} \log \int_{[-1,1]} \exp\Big(n\Big(F(x) + \frac{\beta}{2}x^2\Big)\Big)\mu_n(dx) \\
&= -\sup_{y\in[-1,1]} \Big[\frac{\beta}{2}y^2 - I(y)\Big] + \sup_{x\in[-1,1]} \Big[F(x) + \frac{\beta}{2}x^2 - I(x)\Big] \\
&= \sup_{x\in[-1,1]} \Big[F(x) - \Big(I(x) - \frac{\beta}{2}x^2 - \inf_{y\in[-1,1]} \Big[I(y) - \frac{\beta}{2}y^2\Big]\Big)\Big] \\
&= \sup_{x\in[-1,1]} [F(x) - I_\beta(x)],
\end{aligned}$$

where I_β is defined by (4.11): this proves (4.12). Moreover, I_β is a good rate function on $[-1,1]$: I_β is nonnegative since

$$I(x) - \frac{\beta}{2}x^2 \geq \inf_{y\in[-1,1]} \Big[I(y) - \frac{\beta}{2}y^2\Big] \qquad \forall x \in [-1,1],$$

and I_β is finite and continuous on the compact interval $[-1,1]$ since I is so, which implies that I_β has compact level sets. □

Studying the properties of the good rate function I_β will allow us to determine the weak limit of the measures $\mathbb{P}_{n,\beta}\big(\frac{S_n}{n} \in \cdot\big)$, as the following theorem shows.

Theorem 4.18. The laws of the average spin per site $\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^n \omega_i$ with respect to the probability $\mathbb{P}_{n,\beta}$ have the following weak limit:

$$\mathbb{P}_{n,\beta}\Big(\frac{S_n}{n} \in \cdot\Big) \rightharpoonup \begin{cases} \delta_0 & \text{if } 0 < \beta \le 1, \\ \frac{1}{2}\delta_{-m_\beta} + \frac{1}{2}\delta_{m_\beta} & \text{if } \beta > 1, \end{cases}$$

for some m_β such that $\sqrt{(\beta-1)/\beta} < m_\beta < 1$.

Proof. Since

$$I_\beta(x) = \frac{1}{2}(1-x)\log(1-x) + \frac{1}{2}(1+x)\log(1+x) - \frac{\beta}{2}x^2 - \inf_{y\in[-1,1]}\Big[I(y) - \frac{\beta}{2}y^2\Big],$$

deriving twice we obtain for all $x \in (-1,1)$

$$\begin{aligned} I_\beta'(x) &= -\frac{1}{2}\log(1-x) - \frac{1}{2} + \frac{1}{2}\log(1+x) + \frac{1}{2} - \beta x = \frac{1}{2}[\log(1+x) - \log(1-x)] - \beta x \\ &= \frac{1}{2}\log\Big(\frac{1+x}{1-x}\Big) - \beta x = \operatorname{arctanh}(x) - \beta x, \\ I_\beta''(x) &= \frac{1}{2}\Big[\frac{1}{1+x} + \frac{1}{1-x}\Big] - \beta = \frac{1}{2}\frac{1-x+1+x}{(1+x)(1-x)} - \beta = \frac{1}{1-x^2} - \beta. \end{aligned}$$

We note the following general properties of I_β:

- The good rate function I_β is continuous on the compact $[-1,1]$, hence it attains its minimum 0 at some point of the interval.
- I_β is an even function, i.e. $I_\beta(-x) = I_\beta(x)$ for all $x \in [-1,1]$.
- The stationary points, i.e. the points with zero derivative, are the solutions in $(-1,1)$ of the so-called *mean field equation* $\tanh(\beta x) = x$. In particular, 0 is always a stationary point (there may exist other ones).
- $I_\beta''(x) > 0$ if and only if $x^2 > (\beta-1)/\beta$.

We now distinguish the cases $0 < \beta \le 1$ and $\beta > 1$.

Let first $0 < \beta \le 1$. In this case, $I_\beta''(x) > 0$ for all $x \in (-1,1)$ ($I_\beta(x) = 0$ only if $x = 0$, in the case $\beta = 1$), hence I_β is strictly convex on $[-1,1]$ and the stationary point 0 is the unique minimum point. To prove that $\mathbb{P}_{n,\beta}\big(\frac{S_n}{n}\in\cdot\big)$ converges weakly to δ_0, for example we may apply Theorem 2.23: it suffices to show that

$$\limsup_{n\to\infty}\mathbb{P}_{n,\beta}\Big(\frac{S_n}{n}\in C\Big)\le\delta_0(C)\qquad\forall C\subseteq[-1,1]\text{ closed.}$$

If $0\in C$, $\delta_0(C)=1$ and the claim is trivial, so assume that $0\notin C$. Since by Theorem 4.17 $\mathbb{P}_{n,\beta}\big(\frac{S_n}{n}\in\cdot\big)$ satisfies the LDP with rate n and good rate function I_β,

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}_{n,\beta}\Big(\frac{S_n}{n}\in C\Big)\le-\inf_{x\in C}I_\beta(x)<0$$

(the infimum is strictly positive since $I_\beta(x)>0$ for all $x\in C$ and C is closed). In other words,

$$\limsup_{n\to\infty}\mathbb{P}_{n,\beta}\Big(\frac{S_n}{n}\in C\Big)\le\limsup_{n\to\infty}\exp\Big(-n\inf_{x\in C}I_\beta(x)\Big)=0=\delta_0(C),$$

which concludes the proof of the claim.

Let now $\beta>1$. In this case, $I_\beta''(x)>0$ if and only if $-1<x<-\sqrt{(\beta-1)/\beta}$ or $\sqrt{(\beta-1)/\beta}<x<1$, and $I_\beta''(x)=0$ if and only if $x=\pm\sqrt{(\beta-1)/\beta}$. We deduce that:

- I_β is convex on the intervals $\big[-1,-\sqrt{(\beta-1)/\beta}\big]$ and $\big[\sqrt{(\beta-1)/\beta},1\big]$, and concave on $\big[-\sqrt{(\beta-1)/\beta},\sqrt{(\beta-1)/\beta}\big]^\dagger$; in particular, I_β has two inflection points at $\pm\sqrt{(\beta-1)/\beta}$.
- I_β has a local maximum point at 0, and two minimum points, in the intervals $\big(-1,-\sqrt{(\beta-1)/\beta}\big)$ and $\big(\sqrt{(\beta-1)/\beta},1\big)$ respectively; by the symmetry of I_β, these minimum points are opposite, i.e. they are m_β and $-m_\beta$ for some $m_\beta\in\big(\sqrt{(\beta-1)/\beta},1\big)$. As a consequence, $I_\beta(\pm m_\beta)=0$, and $\pm m_\beta$ are also stationary points, i.e. they satisfy $\tanh(\pm\beta m_\beta)=\pm m_\beta$.

To prove that $\mathbb{P}_{n,\beta}\big(\frac{S_n}{n}\in\cdot\big)$ converges weakly to $\mu\doteq\frac{1}{2}\delta_{-m_\beta}+\frac{1}{2}\delta_{m_\beta}$, by Theorem 2.23 it suffices to prove that

$$\liminf_{n\to\infty}\mathbb{P}_{n,\beta}\Big(\frac{S_n}{n}\in O\Big)\ge\frac{1}{2}\delta_{-m_\beta}(O)+\frac{1}{2}\delta_{m_\beta}(O)\qquad\forall O\subseteq[-1,1]\text{ open.}$$

We distinguish the following cases:

- If $m_\beta\notin O$ and $-m_\beta\notin O$, then $\mu(O)=0$ and the inequality is trivial.

†This is the first significative example of non-convex good rate function.

- If $m_\beta \in O$ and $-m_\beta \in O$, then O^c is a closed such that $\pm m_\beta \notin O^c$, and it is easy to show that $\mathbb{P}_{n,\beta}\left(\frac{S_n}{n} \in O^c\right) \to 0$, just as in the case $0 < \beta \leq 1$. Therefore, $\mathbb{P}_{n,\beta}\left(\frac{S_n}{n} \in O\right) \to 1 = \frac{1}{2}\delta_{-m_\beta}(O) + \frac{1}{2}\delta_{m_\beta}(O)$ and the inequality holds. We also note that, if we take $O \doteq (-m_\beta - \varepsilon, -m_\beta + \varepsilon) \cup (m_\beta - \varepsilon, m_\beta + \varepsilon)$ for any $0 < \varepsilon < m_\beta \wedge (1 - m_\beta)$, we obtain

$$\lim_{n\to\infty}\left[\mathbb{P}_{n,\beta}\left(-m_\beta - \varepsilon < \frac{S_n}{n} < -m_\beta + \varepsilon\right) + \mathbb{P}_{n,\beta}\left(m_\beta - \varepsilon < \frac{S_n}{n} < m_\beta + \varepsilon\right)\right] = 1.$$

  Since $\frac{S_n}{n}$ and $-\frac{S_n}{n}$ have the same law with respect to $\mathbb{P}_{n,\beta}$ (this symmetry is evident from (4.7) and (4.8)), it follows that the two probabilities above are equal, hence

$$\lim_{n\to\infty}\mathbb{P}_{n,\beta}\left(-m_\beta - \varepsilon < \frac{S_n}{n} < -m_\beta + \varepsilon\right) = \lim_{n\to\infty}\mathbb{P}_{n,\beta}\left(m_\beta - \varepsilon < \frac{S_n}{n} < m_\beta + \varepsilon\right) = \frac{1}{2}. \quad (4.13)$$

- Finally, if $m_\beta \in O$ and $-m_\beta \notin O$, there exists $0 < \varepsilon < m_\beta \wedge (1 - m_\beta)$ such that $(m_\beta - \varepsilon, m_\beta + \varepsilon) \subseteq O$. By (4.13),

$$\begin{aligned}\liminf_{n\to\infty}\mathbb{P}_{n,\beta}\left(\frac{S_n}{n} \in O\right) &\geq \liminf_{n\to\infty}\mathbb{P}_{n,\beta}\left(m_\beta - \varepsilon < \frac{S_n}{n} < m_\beta + \varepsilon\right) \\ &= \frac{1}{2} = \frac{1}{2}\delta_{-m_\beta}(O) + \frac{1}{2}\delta_{m_\beta}(O).\end{aligned}$$ □

The latter theorem allows us to claim that a transition phase also occurs in our model as the dimension n of the system tends to infinity. When the temperature is smaller than a certain critical value (i.e., when $\beta > 1$), the average configuration of spins concentrates on the two values $\pm m_\beta$, which correspond to the minimum points of the good rate function I_β; this shows the spontaneous magnetization of the system, due to the parallel alignment of spins. When the temperature becomes larger than the critical value (i.e., when $\beta \leq 1$), the spins behave as independent random variables satisfying the law of large numbers, and the average spin per site concentrates on the mean value 0, which corresponds to the unique minimum point of the good rate function I_β in this second case; this shows the paramagnetic behavior of the material for high temperatures, when the alignment effects disappear and the spins arrange randomly. In the light of these observations, the denomination *equilibrium macrostates* makes sense for the set of the minimum points of I_β: the equilibrium macrostates are $\pm m_\beta$ if $\beta > 1$ and 0 if $0 < \beta \leq 1$.

Bibliography

[Ale] J. Alexopoulos, *A brief introduction to N–functions and Orlicz function spaces*. Lecture notes available on the web page `http://www.personal.kent.edu/~jalexopo/Miscellaneous%20Topics/A%20brief%20introduction/Lectures.pdf` (2004).

[Bah] R. R. Bahadur, *Some Limit Theorems in Statistics*. Vol. 4 of CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia (1971).

[Bil] P. Billingsley, *Convergence of Probability Measures*. Second edition, Wiley, New York (1999).

[Blo] G. Blom, *Harald Cramer 1893-1985*, The Annals of Statistics, Vol. 15, No. 4, Institute of Mathematical Statistics (December 1987).

[Bou] N. Bourbaki, *Éléments de mathématique. Topologie générale. Chapitres 5 à 10.* Springer-Verlag Berlin Heidelberg (2007).

[Bre] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Springer-Verlag New York (2010).

[Bry] W. Bryc, *Large deviations by the asymptotic value method*. Diffusion Processes and Related Problems in Analysis Vol. 1, ed. M. Pinsky, Birkhäuser, Boston (1990).

[Cho] G. Choquet, *Lectures on Analysis. Volume I. Integration and Topological Vector Spaces.* Benjamin, London (1969).

[CoTh] T. M. Cover, J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons (1991).

[Cra] H. Cramér, *Sur un nouveau théorème-limite de la théorie des probabilités*. Actualités Scientifiques et Industrielles 736 (1938).

[DeZe] A. Dembo, O. Zeitouni, *Large Deviations Techniques and Applications*. 2nd edition, Springer-Verlag New York (1998).

[Ell] R. S. Ellis, *The theory of large deviations: from Boltzmann's 1877 calculation to equilibrium macrostates in 2D turbulence*. Physica D 133, pages 106–136, Elsevier (1999).

[EKM] P. Embrechts, C. Klüppelberg, T. Mikosch, *Modelling Extremal Events for Insurance and Finance.* Springer-Verlag Berlin Heidelberg (1997).

[EtKu] S. N. Ethier and T. G.Kurtz, *Markov Processes. Characterization and Convergence*. John Wiley & Sons, New York (1986).

[Fav] Valentina Favero, *Il modello di Curie-Weiss e le grandi deviazioni*. Bachelor's thesis, Università degli Studi di Padova (2009-2010).

[Hol] Frank den Hollander, *Large Deviations.* Fields Institute Monographs 14. American Mathematical Society, Providence (2000).

[Kec] A. S. Kechris, *Classical Descriptive Set Theory*. Springer-Verlag New York (1995).

[Kle] A. Klenke, *Probability Theory. A comprehensive course.* Springer-Verlag Berlin Heidelberg (2006).

[LeRo] E. L. Lehmann, J. P. Romano, *Testing Statistical Hypotheses.* Springer (2005).

[Lun] F. Lundberg, *Approximerad framställning av sannolikhetsfunktionen. Återförsäkring av kollektivrisker.* Akad. Afhandling. Almqvist och Wiksell, Uppsala (1903).

[Man] M. Manetti, *Topologia*. Springer-Verlag Italia, Milano (2008).

[OBVe] G. L. O'Brien and W. Vervaat, *Capacities, large deviations and loglog laws.* Stable Processes and Related Topics, Progress in Probability 25, pages 43–83, Birkhäuser, Boston (1991).

[Pha] H. Pham, *Some Applications and Methods of Large Deviations in Finance and Insurance.* Paris-Princeton Lectures on Mathematical Finance 2004, Springer-Verlag Berlin Heidelberg (2007).

[Ram] S. Ramasubramanian, *Large deviations: an introduction to 2007 Abel prize*. Proceedings Mathematical Sciences, Vol. 118, Issue 2, Springer (May 2008).

[San] I. N. Sanov, *On the probability of large deviations of random variables*. Matematicheskii Sbornik 42, pages 11-44 (1957). [English translation: Selected Translations in Mathematical Statistics and Probability I, pages 213–244 (1961)].

[Soa] P. M. Soardi, *Analisi Matematica. Nuova edizione*. Città Studi Edizioni (2007).

[Swa] J. M. Swart, *Large Deviation Theory*. Lecture notes available on the web page `staff.utia.cas.cz/swart/LDP10.pdf` (2012).

[Var] S. R. S. Varadhan, *Asymptotic probabilities and differential equations*. Communications on Pure and Applied Mathematics, Vol. 19 (1966).

Notation

$\varnothing$	Empty set.
$A \cap B$	Intersection of two sets A and B.
$A \cup B$	Union of two sets A and B.
$A \dot{\cup} B$	Disjoint union of two sets A and B.
A^c	Complement of a set A.
$A \setminus B$	Intersection of A and B^c.
$\mathring{A}$	Interior of a subset A of a topological space.
$\overline{A}$	Closure of a subset A of a topological space.
∂A	Boundary of a subset A of a topological space.
$\|A\|$	Lebesgue measure of a measurable set $A \subseteq \mathbb{R}^d$.
$B_r(x)$	Open ball of radius r centered at a point x, in a metric space.
$\mathbb{N}$	Strictly positive integers.
$\mathbb{N}_0$	Nonnegative integers.
$\mathbb{Z}$	Integers.
$\mathbb{Q}$	Rational numbers.
$\mathbb{R}$	Real numbers.
$\mathbb{R}^+$	Nonnegative real numbers.
$\overline{\mathbb{R}}$	Extended real line, $\mathbb{R} \cup \{-\infty, +\infty\}$.
$n!$	Factorial of a nonnegative integer n.
$x \wedge y$	Minimum of two numbers x and y.
$x \vee y$	Maximum of two numbers x and y.
x^+	Positive part of a real number x, i.e. $x \wedge 0$.
x^-	Negative part of a real number x, i.e. $(-x) \wedge 0$.
$\inf(A)$	Infimum of a set $A \subseteq \overline{\mathbb{R}}$.
$\sup(A)$	Supremum of a set $A \subseteq \overline{\mathbb{R}}$.
$x_n \to x$	Sequence x_n converging to x.
$x_n \nearrow x$	Increasing real sequence x_n converging to x.
$x_n \searrow x$	Decreasing real sequence x_n converging to x.
$\langle x, y \rangle$	Euclidean inner product of two vectors $x, y \in \mathbb{R}^d$.
$\|x\|$	Euclidean norm of a vector $x \in \mathbb{R}^d$.
$\mathbb{1}_A$	Indicator function of a set A.
$f(A)$	Image of a function f under a set A.

f^{-1}	Inverse image or inverse function of a function f.
$f \circ g$	Composition of two functions f and g.
$f', f'', \dots$	Derivatives of a function f of one variable.
∇f	Gradient of a function f of several variables.
$D^2 f$	Hessian matrix of a function f of several variables.
$\mathbb{P}$	Probability.
$\mu \otimes \nu$	Product measure of two probability measures μ and ν.
$\mathbb{E}$	Expectation.
Var	Variance.
Cov	Covariance.
a.s.	Almost surely.
a.e.	Almost every.
i.i.d.	Independent and identically distributed.
δ_x	Dirac probability measure concentrated on x.
$U(A)$	Continuous uniform distribution on a Borel set $A \subseteq \mathbb{R}^d$.

Index