Convergence of random variables and large deviations

Pierre-Loïc Méliot

Abstract. The goal of these lectures is to study the qualitative and quantitative aspects of the convergence of a sequence of random variables. For the qualitative approach, we shall be interested in the connections between the various possible notions of convergence in a probabilistic setting, and in the underlying topology of the space of probability measures on a topological space. We shall follow mainly the two first chapters of [**Billingsley, 1999**]. Then, given a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ that converges in probability, we shall study the rate of convergence of the sequence by estimating the quantities

$$-s_n \log \mathbb{P}[X_n \in B],$$

with B some fixed measurable subset of the space $\mathfrak{X}$ where the random variables take their values, and $(s_n)_{n\in\mathbb{N}}$ some sequence going to 0 and giving the exponential speed of convergence. This leads to the theory of large deviations, for which we shall follow [**Dembo and Zeitouni, 1998**, **Feng and Kurtz, 2006**]. Though we shall try to treat each subject in the most general setting (usually a polish space which may be infinite-dimensional), the main connection between the different chapters of these lectures will consist in the three following examples.

(1) Consider the mean $Z_n = \frac{1}{n}\sum_{i=1}^n X_i$ of a sequence of independent (or weakly dependent) random variables. It is involved in the two most known limit theorems in probability: the law of large numbers and the central limit theorem. Can one quantify these two convergences and estimate their speed?

(2) Consider a (time-homogeneous) Markov chain $(X_n)_{n\in\mathbb{N}}$. Under certain hypotheses, *e.g.*, the finiteness of the space of states and the irreducibility and aperiodicity of the kernel, the law of X_n is known to converge to a stationary measure. Is it possible to compute the corresponding mixing time?

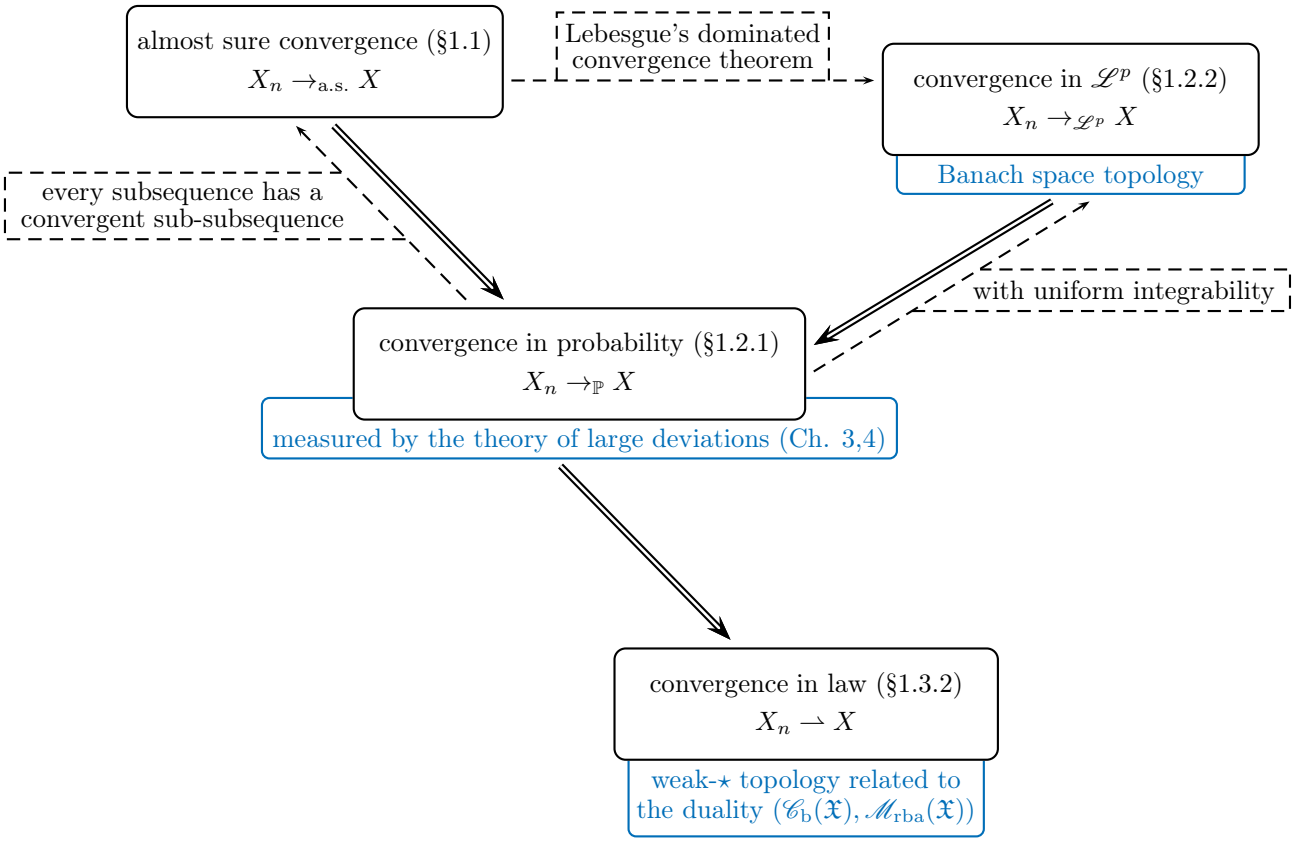
(3) Finally, consider a random walk $(X_n)_{n\in\mathbb{N}}$, *e.g.*, the sum of independent Bernoulli variables. Under appropriate scaling, one expects the random walk to converge to the continuous analogue of random walks, namely, the Brownian motion. Can one give a precise meaning to this convergence, and deduce from this certain fine properties of the Brownian motion?

Contents

CHAPTER 1

Notions of convergence in a probabilistic setting

In this first chapter, we present the most common notions of convergence used in probability: almost sure convergence, convergence in probability, convergence in $\mathscr{L}^p$-norms and convergence in law. We show the connections between these notions, and we detail the topology of convergence in law, which is to be understood as a structure of polish space on the space $\mathscr{M}^1(\mathfrak{X})$ of probability measures on a space $\mathfrak{X}$. Our main sources are [**Billingsley, 1999**, **Kallenberg, 2001**].



Let us fix some notations. The usual sets of integer, rational, real and complex numbers are denoted $\mathbb{N}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$. A measurable space is usually denoted $(\Omega, \mathcal{B})$, where $\mathcal{B}$ is the σ-field of measurable events; and a probability space is usually denoted $(\Omega, \mathcal{B}, \mathbb{P})$. If $\Omega = \mathfrak{X}$ is a topological space, then unless explicitly stated, $\mathcal{B} = \mathcal{B}(\mathfrak{X})$ will be taken to be the set of Borelians, that is to say the smallest σ-field that contains open (and closed)

subsets. We will call random variable a measurable map $X : (\Omega, \mathcal{B}, \mathbb{P}) \to (\mathfrak{X}, \mathcal{B}(\mathfrak{X}))$; it induces a probability measure $\mathbb{P}_X$ on the space of states $(\mathfrak{X}, \mathcal{B}(\mathfrak{X}))$, the law of X, which is defined by

$$\mathbb{P}_X[A \in \mathcal{B}(\mathfrak{X})] = \mathbb{P}[X^{-1}(A)].$$

Sometimes we shall also denote $\mathbb{P}_X = X_*\mathbb{P}$. The set of all probability measures on a measurable space $(\mathfrak{X}, \mathcal{B})$ is denoted $\mathscr{M}^1(\mathfrak{X}, \mathcal{B})$, or simply $\mathscr{M}^1(\mathfrak{X})$. On the other hand, the set of continuous real-valued functions on a topological space $\mathfrak{X}$ will be denoted $\mathscr{C}(\mathfrak{X})$, and $\mathscr{L}^p(\Omega) = \mathscr{L}^p(\Omega, \mathcal{B}, \mathbb{P})$ will stand for the set of measurable real-valued functions on $(\Omega, \mathcal{B})$ with

$$\int_\Omega |f(\omega)|^p \, \mathbb{P}(d\omega) < +\infty,$$

where p is an exponent in $[1, +\infty)$.

Remark. The main difficulty of these lectures lies probably in the topological background, as we shall need to manipulate general topological spaces, that is to say more general than the usual normed vector spaces. The first chapters of [Lang, 1993] form an excellent reference for this theory. If needed, the reader can replace most of the statements involving a general topological space by a statement involving a metric space. In this setting, many proofs are eased by the fact that the topology can be described in terms of convergence of sequences, and measured explicitely by the metric.

1.1. Almost sure convergence

Consider a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables defined on a common probability space $(\Omega, \mathcal{B}, \mathbb{P})$, and with values in a topological space $\mathfrak{X}$. We recall that a sequence $(x_n)_{n\in\mathbb{N}}$ of points in $\mathfrak{X}$ is said to converge to $x \in \mathfrak{X}$ if for every open subset U of $\mathfrak{X}$ containing x,

$$\exists N = N(U), \ \ \forall n \geq N, \ \ x_n \in U.$$

When $\mathfrak{X} = (\mathfrak{X}, d)$ is a metric space, this is equivalent to the usual definition with small ε's. A topology being fixed on $\mathfrak{X}$, we denote the convergence of a sequence by a simple arrow: $x_n \to x$.

Definition 1.1. *The sequence $(X_n)_{n\in\mathbb{N}}$ is said to* ***converge almost surely*** *to a random variable $X : (\Omega, \mathcal{B}, \mathbb{P}) \to \mathfrak{X}$ if there exists a subset $\Omega' \subset \Omega$ of probability 1 and such that*

$$\forall \omega \in \Omega', \ \ X_n(\omega) \to X(\omega).$$

Notation: $X_n \to_{\text{a.s.}} X$.

This is the most natural way to define the convergence of a sequence of random variables, in the sense that it is the direct adaptation of the deterministic setting. The almost sure convergence occurs for instance in the famous strong law of large numbers. Recall

that a family of random variables $(X_i)_{i\in I}$ is said independent if, for every finite subset $\{i_1,\dots,i_r\}\subset I$ and for every measurable subsets $A_1,\dots,A_r\in\mathcal{B}(\mathfrak{X})$,

$$\mathbb{P}[X_{i_1}\in A_1,\dots,X_{i_r}\in A_r]=\prod_{j=1}^{r}\mathbb{P}[X_{i_j}\in A_j].$$

In other words, the law of $(X_i)_{i\in I}$ is the direct product of the laws $\mathbb{P}_{X_i}$.

THEOREM 1.2 (Kolmogorov). *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent real random variables with the same law $\mathbb{P}_{X_1}$, and such that X_1 (whence all the X_n's) is in $\mathscr{L}^1(\Omega)$. One has the almost sure convergence*

$$\frac{1}{n}\sum_{i=1}^{n}X_i\to_{\text{a.s.}}\mathbb{E}[X_1]=\int_{\mathbb{R}}x\,\mathbb{P}_{X_1}(dx)=\int_{\Omega}X_1(\omega)\,\mathbb{P}(d\omega).$$

PROOF. Though probably the most well-known result in probability, the ***law of large numbers*** is not at all easy to prove; we follow here [**Kallenberg, 2001**, Chapter 4], which makes a clever use of the Borel-Cantelli lemma and of Chebyshev's inequality. Considering $X_n-\mathbb{E}[X_n]$ instead of X_n, one can of course assume $\mathbb{E}[X_n]=0$ for all n. Denote X an other independent copy of the X_n's, and for any $n\in\mathbb{N}$,

$$X'_n(\omega)=\mathbf{1}_{|X_n|\le n}(\omega)\,X_n(\omega).$$

It is sufficient to show the almost sure convergence of $Z'_n=\frac{1}{n}\sum_{i=1}^{n}X'_i$ instead of $Z_n=\frac{1}{n}\sum_{i=1}^{n}X_i$. Indeed,

$$\begin{aligned}\sum_{n\in\mathbb{N}_{>0}}\mathbb{P}[X'_n\neq X_n]&=\sum_{n\in\mathbb{N}_{>0}}\mathbb{P}[|X_n|>n]=\sum_{n\in\mathbb{N}_{>0}}\mathbb{P}[|X|>n]\\&\le\int_0^{\infty}\mathbb{P}[|X|>x]\,dx=\mathbb{E}[|X|]<\infty,\end{aligned}$$

so by Borel-Cantelli, $X'_n=X_n$ for every n but a finite number (almost surely), and Z_n and Z'_n have therefore the same asymptotic behavior.

Set $Y_n=\frac{X'_n}{n}$; the Y_n's are centered independent random variables, bounded in absolute value by 1, and we are going to show that $\sum_{n=1}^{\infty}Y_n$ converges almost surely. Notice that

$$\begin{aligned}\sum_{n\ge1}\mathbb{E}[(Y_n)^2]&=\sum_{n\ge1}\frac{1}{n^2}\,\mathbb{E}[X^2\,\mathbf{1}_{|X|\le n}]\le1+\int_1^{\infty}\frac{1}{x^2}\,\mathbb{E}[X^2\,\mathbf{1}_{|X|\le x}]\,dx\\&\le1+\mathbb{E}\left[X^2\int_{|X|}^{\infty}\frac{1}{t^2}\,dt\right]=1+\mathbb{E}[|X|]<+\infty.\end{aligned}$$

Set $S_n=Y_1+Y_2+\cdots+Y_n$ and

$$\tau=\inf\{k\ge n_1,\ \ |S_k-S_{n_1}|>\varepsilon\},$$

where ε is a fixed positive real number. The independence of the Y_n's ensures that for any $n_2 \geq k \geq n_1$, $(S_{n_2} - S_k)$ and $(S_k - S_{n_1})\,\mathbf{1}_{\tau=k}$ are independent. Then,

$$\begin{aligned}\sum_{k=n_1+1}^{n_2} \mathbb{E}[(Y_k)^2] = \mathbb{E}[(S_{n_2} - S_{n_1})^2] &\geq \sum_{k=n_1+1}^{n_2} \mathbb{E}[(S_{n_2} - S_{n_1})^2\,\mathbf{1}_{\tau=k}] \\ &\geq \sum_{k=n_1+1}^{n_2} \mathbb{E}[(S_k)^2\,\mathbf{1}_{\tau=k}] + \mathbb{E}[2\,S_k\,(S_{n_2} - S_k)^2\,\mathbf{1}_{\tau=k}] + \mathbb{E}[(S_{n_2} - S_k)^2] \\ &\geq \sum_{k=n_1+1}^{n_2} \mathbb{E}[(S_k)^2\,\mathbf{1}_{\tau=k}] \geq \varepsilon^2\,\mathbb{P}[\tau \leq n_2].\end{aligned}$$

With n_2 going to infinity, we thus get

$$\frac{1}{\varepsilon^2} \sum_{k=n_1+1}^{\infty} \mathbb{E}[(Y_k)^2] \geq \mathbb{P}\left[\sup_{k \geq n_1} |S_k - S_{n_1}| > \varepsilon\right],$$

and since the left-hand side goes to zero when n_1 goes to infinity, so does the right-hand side. Hence, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{k \geq n} |S_k - S_n| \leq \varepsilon\right] = 1.$$

Now, let us analyze the event "$(S_n)_{n\in\mathbb{N}}$ does not converge". It also means that this is not a Cauchy sequence, hence, we look at

$$\bigcup_{\varepsilon>0} \bigcap_{n\geq 1} \left\{\sup_{k\geq n} |S_k - S_n| > \varepsilon\right\} = \bigcup_{m\geq 1} \bigcap_{n\geq 1} \left\{\sup_{k\geq n} |S_k - S_n| > \frac{1}{m}\right\}.$$

But we have just seen that

$$\mathbb{P}\left[\bigcap_{n\geq 1} \left\{\sup_{k\geq n} |S_k - S_n| > \frac{1}{m}\right\}\right] \leq \inf_{n\geq 1} \mathbb{P}\left[\sup_{k\geq n} |S_k - S_n| > \frac{1}{m}\right] = 0.$$

Taking a countable union over $m \geq 1$ does not change the zero probability, so we have indeed shown that $(S_n)_{n\in\mathbb{N}}$ was almost surely a Cauchy sequence, whence convergent. Finally,

$$\begin{aligned}S_n(\omega) - Z'_n(\omega) &= \sum_{i=1}^{n} \left(1 - \frac{i}{n}\right) \frac{X'_i}{i} = \sum_{i=1}^{n} \left(\int_{\frac{i}{n}}^{1} 1\,dx\right) \frac{X'_i}{i} \\ &= \int_0^1 \left(\sum_{i\leq nx} \frac{X'_i}{i}\right) dx = \int_0^1 S_{\lfloor nx \rfloor}(\omega)\,dx\end{aligned}$$

for a fixed ω in the set of convergence of $(S_n)_{n\in\mathbb{N}}$. By dominated convergence of the maps $x \mapsto S_{\lfloor nx \rfloor}(\omega)$, the right-hand side converges to $\int_0^1 (\lim_{n\to\infty} S_n(\omega))\,dx = \lim_{n\to\infty} S_n(\omega)$, so $\lim_{n\to\infty} Z'_n(\omega)$ exists also and is zero. □

The proof of the strong law of large numbers is a good illustration of many features of the almost sure convergence: in particular, most of the time, one has first to find good estimates on the probability of "bad behavior" of the sequence, and then use the ***Borel-Cantelli lemma*** to prove the good behavior outside an event of probability zero. In this context, the almost sure convergence appears as a refinement of weaker notions of convergence, which are prerequisites in the proof of a strong convergence theorem. This motivates our next paragraph, where we shall deal with the convergences in probability and $\mathscr{L}^p$-norms. One of the main advantages of these notions is that their proof in a concrete situation usually involves totally explicit computations, *e.g.*, computations of estimates of moments; most of the time this is far easier than a direct proof of the strong (almost sure) convergence.

1.2. Convergence in probability and in $\mathscr{L}^p$-spaces

Again, consider a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables defined on a common probability space and with values in a topological space $\mathfrak{X}$. For simplicity we shall assume that $\mathfrak{X}$ is metrizable, that is to say that there is a distance d on $\mathfrak{X}$ such that a basis of neighborhoods of any point $x \in \mathfrak{X}$ consists in the open balls $B_{(x,\varepsilon)} = \{y \in \mathfrak{X},\ d(x,y) < \varepsilon\}$. We could also have treated the notion of convergence in probability in the setting of topological vector spaces, but the choice of metrizable spaces was more coherent with the discussion of §1.3.

1.2.1. Convergence in probability.

Definition 1.3. *The sequence $(X_n)_{n\in\mathbb{N}}$ **converges in probability** to a random variable X if for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}[d(X_n, X) \geq \varepsilon] = 0.$$

We shall denote $X_n \to_{\mathbb{P}} X$ for the convergence in probability. In fact, the notion only depends on the underlying topology of $\mathfrak{X}$, and not on the metric compatible with the topology.

The independence from the metric of the notion of convergence in probability relies on the following result, which relates convergence in probability and almost sure convergence.

Proposition 1.4. *A sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to X if and only if every subsequence $(X_{\phi(n)})_{n\in\mathbb{N}}$ has a sub-subsequence $(X_{\phi\circ\psi(n)})_{n\in\mathbb{N}}$ that converges almost surely to X. As a consequence, $X_n \to_{\text{a.s.}} X$ implies $X_n \to_{\mathbb{P}} X$.*

Proof. Suppose $X_n \to_{\mathbb{P}} X$, and fix a subsequence $(Y_n)_{n\in\mathbb{N}}$. We first prove that for $\eta > 0$, there is a subset $\Omega_\eta \subset \Omega$ of probability greater than $1-\eta$, and an extraction $\phi_\eta : \mathbb{N} \to \mathbb{N}$ such that $Y_{\phi_\eta(n)}(\omega) \to X(\omega)$ for all $\omega \in \Omega$. There exists an integer $n_{\eta,1} \geq 1$ such that

$$\mathbb{P}\left[d(Y_{n_{\eta,1}}, X) \geq \frac{1}{2}\right] \leq \frac{\eta}{2};$$

then, an integer $n_{\eta,2} > n_{\eta,1}$ such that

$$\mathbb{P}\left[d(Y_{n_{\eta,2}}, X) \geq \frac{1}{4}\right] \leq \frac{\eta}{4};$$

etc.; we choose each time $n_{\eta,k} > n_{\eta,k-1}$ such that

$$\mathbb{P}\left[d(Y_{n_{\eta,k}}, X) \geq \frac{1}{2^k}\right] \leq \frac{\eta}{2^k}.$$

The probability of the union $\bigcup_{k\geq 1}\{d(Y_{n_{\eta,k}}, X) \geq \frac{1}{2^k}\}$ is smaller than $\frac{\eta}{2} + \frac{\eta}{4} + \cdots = \eta$, so

$$\mathbb{P}\left[\forall k \geq 1,\ \ d(Y_{n_{\eta,k}}, X) \leq \frac{1}{2^k}\right] \geq 1 - \eta.$$

Setting $\phi_\eta(k) = n_{\eta,k}$, we have shown the previous statement. Then, we proceed by ***diagonal extraction***. Let

- $\psi_1 = \phi_{\frac{1}{2}}$ be an extraction for the sequence $(Y_n)_{n\in\mathbb{N}}$ and corresponding to the parameter $\eta = \frac{1}{2}$;
- $\psi_2 = \phi_{\frac{1}{4}}$ be an extraction for the sequence $(Y_{\psi_1(n)})_{n\in\mathbb{N}}$ and corresponding to the parameter $\eta = \frac{1}{4}$;

and so on. The injection $\psi_k : \mathbb{N} \to \mathbb{N}$ is an extraction for the sequence $(Y_{\psi_1\circ\psi_2\circ\cdots\circ\psi_{k-1}(n)})_{n\in\mathbb{N}}$ and corresponding to the parameter $\eta = \frac{1}{2^k}$. The previous inequality shows that if

$$\Psi(n) = \psi_1 \circ \psi_2 \circ \cdots \circ \psi_n(n),$$

then

$$\mathbb{P}\left[\forall n \geq N,\ \ d(Y_{\Psi(n)}, X) \leq \frac{1}{2^n}\right] \geq 1 - \frac{1}{2^N},$$

because the sequence considered in this estimate of probabilities is a subsequence of $Y_{\psi_1\circ\cdots\circ\psi_N}$. It follows that

$$\mathbb{P}[d(Y_{\Psi(n)}, X) \to 0] \geq 1 - \frac{1}{2^N}$$

for every N, so in fact this probability is 1. Hence, any subsequence of a sequence of r.v. that converges in probability to X has a sub-subsequence that converges almost surely.

Conversely, suppose that $(X_n)_{n\in\mathbb{N}}$ does not converge in probability to X; then there is an $\varepsilon > 0$ such that $\mathbb{P}[d(X_n, X) \geq \varepsilon]$ stays bigger than some $\eta > 0$ for an infinite number of n's, say, along a subsequence $(Y_n)_{n\in\mathbb{N}}$. Now, if the second statement were true, then this subsequence would have a further subsequence $(Y_{\Psi(n)})_{n\in\mathbb{N}}$ that also converges almost surely, though with

$$\forall n \in \mathbb{N},\ \ \mathbb{P}[d(Y_{\Psi(n)}, X) \geq \varepsilon] \geq \eta.$$

The functions $\min\{d(Y_{\psi(n)}, X), 1\}$ are then dominated by 1 and converge almost surely to 0, so

$$\mathbb{E}[\min\{d(Y_{\psi(n)}, X), 1\}] \to 0$$

by the dominated convergence theorem. However, it stays also bigger that $\varepsilon\eta$ for $\varepsilon \leq 1$; contradiction.

Since the almost sure convergence is a purely topological notion, this criterion shows that the convergence in probability does not depend directly on the metric (by that we mean that two distances compatible with the topology yield the same notion of convergence in probability). □

The previous criterion has numerous applications: indeed, almost every result related to the convergence of deterministic sequences trivially "survives" to the almost sure setting, and then, using Proposition 1.4, one gets the analogue result for the convergence in probability. Thus, let us state easy "transformation results" for these two notions of convergence:

(1) Let $f : \mathfrak{X} \to \mathfrak{W}$ be a continuous map. If $X_n \to_{\text{a.s.}} X$, then $f(X_n) \to_{\text{a.s.}} f(X)$. Similarly, if $X_n \to_{\mathbb{P}} X$, then $f(X_n) \to_{\mathbb{P}} f(X)$.

(2) Let $(W_n)_{n\in\mathbb{N}}$ and $(X_n)_{n\in\mathbb{N}}$ be two sequences of random variables defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$, and taking their values in two spaces $\mathfrak{W}$ and $\mathfrak{X}$. Assume $W_n \to_{\text{a.s.}} W$ and $X_n \to_{\text{a.s.}} X$. The pairs (W_n, X_n), which are random variables in $\mathfrak{W} \times \mathfrak{X}$, converge almost surely to (W, X). The same statement holds for the convergence in probability.

(3) As a consequence of the two previous results, if $\mathfrak{X}$ is a topological vector space (respectively, a topological algebra), then any linear combination (respectively, any polynomial) of convergent sequences of r.v. also converges (for the almost sure convergence or the convergence in probability).

1.2.2. Convergence in $\mathscr{L}^p$ and uniform integrability. Contrary to the almost sure convergence, there exist various numerical criterions in order to prove the convergence in probability of a sequence of random variables. For simplicity we assume in the following that $\mathfrak{X} = \mathbb{R}$, though most of the theory could be stated in the setting of random variables with values in a normed vector space.

DEFINITION 1.5. *Suppose that $X_n \in \mathscr{L}^p(\Omega)$ for any n, with $p \in [1, +\infty)$. The sequence $(X_n)_{n\in\mathbb{N}}$ **converges in $\mathscr{L}^p$** towards a random variable $X \in \mathscr{L}^p(\Omega)$ if*

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

Then, $X_n \to_{\mathbb{P}} X$. Hence, the convergence in $\mathscr{L}^p$ implies the convergence in probability.

PROOF. This is just a reformulation of the ***Markov-Bienaymé-Chebyshev inequality***:

$$\mathbb{P}[|X_n - X| \geq \varepsilon] = \mathbb{P}[|X_n - X|^p \geq \varepsilon^p] \leq \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p} \to 0.$$

□

Notice that on a probability space, if $p < q$, then by Hölder's inequality,

$$\|X\|_p = \|1 \times X\|_p \leq \|1\|_{\frac{1}{\frac{1}{p}-\frac{1}{q}}} \|X\|_q = \|X\|_q,$$

so $\mathscr{L}^q(\Omega) \subset \mathscr{L}^p(\Omega)$ and the convergence in $\mathscr{L}^q$ implies the convergence in $\mathscr{L}^p$.

EXAMPLE (Weak law of large numbers). Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables, all in $\mathscr{L}^2(\Omega)$. The $\mathscr{L}^2$-norm of the centered mean $Z_n = \frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X]$ is

$$\mathbb{E}[(Z_n)^2] = \sum_{i=1}^n \frac{1}{n^2}\operatorname{Var}[X_i] = \frac{\operatorname{Var}[X]}{n} \to 0,$$

so $Z_n \to_{\mathscr{L}^2} 0$ and consequently, $Z_n \to_{\mathbb{P}} 0$. This statement is weaker than Theorem 1.2, but obviously its proof is way simpler.

Notice that the implications previously shown cannot be reversed: hence, convergence in probability does not imply the almost sure convergence or the $\mathscr{L}^p$-convergence.

EXAMPLE. To any real number $\omega \in [0,1)$, we associate an increasing sequence of integers as follows. Write the proper binary expansion of ω:

$$\omega = \frac{a_1}{2} + \frac{a_2}{4} + \frac{a_3}{8} + \cdots = \sum_{k=1}^{\infty} \frac{a_k}{2^k},$$

with each $a_i \in \{0,1\}$. We then set

$$n_k(\omega) = 2^k + \sum_{i=1}^{k} a_i 2^{k-i} \in \left[\!\left[2^k, 2^{k+1}-1\right]\!\right];$$

the sequence $(n_k(\omega))_{k\geq 1}$ is strictly increasing, with exactly one element in each interval $\left[\!\left[2^k, 2^{k+1}-1\right]\!\right]$. Define then $X_n(\omega) = 1$ if $n = n_k(\omega)$ for some k, and 0 otherwise. Each X_n is the characteristic function of an interval of length

$$\ell = \frac{1}{2^{\lfloor \frac{\log n}{\log 2} \rfloor}},$$

so in particular it is measurable. The X_n's converge in probability to 0; indeed, $\mathbb{E}[|X_n|] = \mathbb{E}[X_n] = \ell \to 0$, so one has convergence in $\mathscr{L}^1([0,1), dx)$. On the other hand, every real number $\omega \in [0,1)$ corresponds to an infinite sequence $(n_k(\omega))_{k\geq 1}$, so to an infinite sequence of $X_n(\omega)$ with $d(X_n(\omega), 0) = 1$; hence, $X_n(\omega)$ never converges to zero. Conclusion: one can have convergence in probability without almost sure convergence.

EXAMPLE. Take the same example as before, but this time with $X_n(\omega) = n^{\frac{1}{p}}$ if $n = n_k(\omega)$ for some k, and 0 otherwise. One still has the convergence in probability of $(X_n)_{n\in\mathbb{N}}$ towards 0: indeed, for any $\varepsilon > 0$,

$$\mathbb{P}[|X_n| \geq \varepsilon] = \ell \to 0.$$

But on the other hand, $\|X_n(\omega)\|_p = (\ell \times n)^{\frac{1}{p}} \geq 1$ for every n, so one does not have convergence in $\mathscr{L}^p$.

To conclude this section, let us see under which conditions one has the converse of Proposition 1.5, *i.e.*, a sequence of random variables that converges in probability also converges in $\mathscr{L}^p(\Omega)$; this leads to the notion of ***uniform integrability***.

DEFINITION 1.6. *A class of functions $\mathscr{F} \subset \mathscr{L}^p(\Omega)$ is said uniformly integrable in $\mathscr{L}^p$ (or simply uniformly integrable when $p = 1$) if for every $\varepsilon > 0$, there exists K such that*

$$\mathbb{E}[|X|^p \, \mathbf{1}_{|X| \geq K}] \leq \varepsilon$$

for all $X \in \mathscr{F}$. In other words, $\lim_{K \to \infty} \sup_{X \in \mathscr{F}} \mathbb{E}[|X|^p \, \mathbf{1}_{|X| \geq K}] = 0$.

EXAMPLE. Every family of functions $\mathscr{F}$ all bounded a.s. by a constant K is uniformly integrable in $\mathscr{L}^p$, for all p.

EXAMPLE. Suppose $(\Omega, \mathcal{B}, \mathbb{P}) = ([0,1], dx)$, and consider the family of functions

$$\mathscr{F}_d = \{x \mapsto x^c\}_{c > -d}$$

with $d \in [0, \frac{1}{p}]$. If $d < \frac{1}{p}$, then $\mathscr{F}_d$ is uniformly integrable in $\mathscr{L}^p(\Omega)$, because

$$\sup_{X \in \mathscr{F}_d} \mathbb{E}[|X|^p \, \mathbf{1}_{|X| \geq K}] = \int_0^1 x^{-dp} \, \mathbf{1}_{x^{-d} \geq K} \, dx = \int_0^{K^{-\frac{1}{d}}} x^{-dp} \, dx = \frac{1}{(1-dp) \, K^{\frac{1-dp}{d}}} \to_{K \to \infty} 0.$$

However if $d = \frac{1}{p}$ then the supremum is always $+\infty$ and the family is not uniformly integrable.

PROPOSITION 1.7. *Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of random variables in $\mathscr{L}^p(\Omega)$. The following assertions are equivalent:*

(i) $X_n \to_{\mathscr{L}^p} X$.

(ii) $X_n \to_{\mathbb{P}} X$ *and* $(X_n)_{n \in \mathbb{N}}$ *is uniformly integrable in* $\mathscr{L}^p$.

PROOF. Suppose $X_n \to_{\mathscr{L}^p} X$; we have already shown that $X_n \to_{\mathbb{P}} X$. Set $\varepsilon > 0$; since $|X|^p \, \mathbf{1}_{|X| \geq K}$ converges to zero almost surely as K goes to infinity, by Lebesgue's dominated convergence theorem, for K big enough, $\int_\Omega |X(\omega)|^p \, \mathbf{1}_{|X(\omega)| \geq K} \, \mathbb{P}(d\omega) \leq \varepsilon$. A similar inequality holds for any finite family of random variables in $\mathscr{L}^p(\Omega)$; in other words, finite subsets of $\mathscr{L}^p(\Omega)$ are uniformly integrable in $\mathscr{L}^p$. Now, since $\|X_n - X\|_p \to 0$, there exists N such that for every $n \geq N$,

$$\int_\Omega |X_n(\omega) - X(\omega)|^p \, \mathbb{P}(d\omega) \leq \varepsilon.$$

Choose K such that $\int_\Omega |X_n(\omega)|^p \, \mathbf{1}_{|X_n(\omega)| \geq K} \, \mathbb{P}(d\omega) \leq \varepsilon$ for every $n < N$ and also for $X_\infty = X$. Then, for every $n \geq N$, one has also

$$\left(\int_\Omega |X_n|^p \, \mathbf{1}_{|X_n| \geq 2K} \right)^{\frac{1}{p}} \leq \left(\int_\Omega |X|^p \, \mathbf{1}_{|X_n| \geq 2K} \right)^{\frac{1}{p}} + \left(\int_\Omega |X_n - X|^p \, \mathbf{1}_{|X_n| \geq 2K} \right)^{\frac{1}{p}}$$

$$\leq \left(\int_\Omega |X|^p \, \mathbf{1}_{|X| \geq K} \right)^{\frac{1}{p}} + \left(\int_\Omega |X|^p \, \mathbf{1}_{|X| < K \text{ and } |X_n| \geq 2K} \right)^{\frac{1}{p}} + \left(\int_\Omega |X_n - X|^p \right)^{\frac{1}{p}}$$

$$\leq \left(\int_\Omega |X|^p \, \mathbf{1}_{|X| \geq K} \right)^{\frac{1}{p}} + 2 \left(\int_\Omega |X_n - X|^p \right)^{\frac{1}{p}} \leq 3 \, \varepsilon^{\frac{1}{p}}$$

being understood that all the integrals are taken against $\mathbb{P}(d\omega)$. So,

$$\int_\Omega |X_n(\omega)|^p \, \mathbf{1}_{|X_n(\omega)|\geq 2K} \, \mathbb{P}(d\omega) \leq 3^p \varepsilon$$

for every $n \in \mathbb{N}$ and K big enough, and the family $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable in $\mathscr{L}^p$.

Conversely, suppose $X_n \to_{\mathbb{P}} X$ and $(X_n)_{n\in\mathbb{N}}$ uniformly integrable in $\mathscr{L}^p$. Fix $\varepsilon > 0$ and K an ε-module of uniform integrability for the family $\{X_n - X, \;\; n \in \mathbb{N}\}$ — one sees easily that removing X does not change the uniform integrability. Then,

$$\begin{aligned}
&\left(\int_\Omega |X_n - X|^p\right)^{\frac{1}{p}} \\
&\leq \left(\int_{|X_n-X|<\varepsilon^{\frac{1}{p}}} |X_n - X|^p\right)^{\frac{1}{p}} + \left(\int_{\varepsilon^{\frac{1}{p}}\leq|X_n-X|<K} |X_n - X|^p\right)^{\frac{1}{p}} + \left(\int_{|X_n-X|\geq K} |X_n - X|^p\right)^{\frac{1}{p}} \\
&\leq \varepsilon^{\frac{1}{p}} + K\,\mathbb{P}[|X_n - X| \geq \varepsilon^{\frac{1}{p}}]^{\frac{1}{p}} + \varepsilon^{\frac{1}{p}}.
\end{aligned}$$

Since one has convergence in probability, for n big enough, the probability in the middle term is smaller than $K^{-p}\varepsilon$, and consequently $\mathbb{E}[|X_n - X|^p] \leq 3^p\varepsilon$, so one has convergence in $\mathscr{L}^p(\Omega)$. □

The previous result is often used as follows. Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of r.v. in $\mathscr{L}^1(\Omega)$ that are uniformly integrable, and that converge in probability to some random variable X. Then one also has $|X_n| \to_{\mathbb{P}} |X|$ since $x \mapsto |x|$ is continuous, and there is a subsequence $|X_{\phi(n)}| \to_{\text{a.s.}} |X|$. By Fatou's lemma,

$$\mathbb{E}[|X|] \leq \liminf_{n\to\infty} \mathbb{E}[|X_{\phi(n)}|] \leq \sup_{n\in\mathbb{N}} \mathbb{E}[|X_n|]$$

the supremum on the right-hand side existing by hypothesis of uniform integrability. Hence, $X \in \mathscr{L}^1(\Omega)$, and the previous discussion shows that $\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n]$.

1.3. Convergence in law and Prohorov's topology

Though more accessible than the almost sure convergence, the convergence in probability is not really adapted to describe certain phenomena. The typical case is the central limit theorem: if Z_n is the mean of independent and identically distributed random variables $X_1, \ldots, X_n$ in $\mathscr{L}^2(\Omega)$, then for any $a < b$,

$$\lim_{n\to\infty} \mathbb{P}\left[a \leq \sqrt{n}\,(Z_n - \mathbb{E}[X]) \leq b\right] = \mathcal{N}_{(0,\mathrm{Var}[X])}(a,b) = \frac{1}{\sqrt{2\pi\,\mathrm{Var}[X]}} \int_a^b \mathrm{e}^{-\frac{x^2}{2\,\mathrm{Var}[X]}}\, dx,$$

but there is no natural Gaussian random variable G such that $\sqrt{n}\,(Z_n - \mathbb{E}[X])$ converges almost surely or in probability to G. This leads to a third important notion of probabilistic convergence, the convergence in law (or in distribution, or weak convergence). In this last section, we give several equivalent definitions of it, and we show that it is induced by the weak topology related to the duality between $\mathscr{C}_b(\mathfrak{X})$ and a space of "generalized" measures $\mathscr{M}_{rba}(\mathfrak{X})$.

1.3.1. Continuous bounded functions and their dual space. A topological space $\mathfrak{X}$ being fixed, we denote $\mathscr{C}_b(\mathfrak{X})$ the space of real continuous functions on $\mathfrak{X}$ that are uniformly bounded:

$$\mathscr{C}_b(\mathfrak{X}) = \{f \in \mathscr{C}(\mathfrak{X}) \mid \exists M \geq 0,\ \ \forall x \in \mathfrak{X},\ \ |f(x)| \leq M\}.$$

This is a complete normed vector space with respect to the norm $\|f\|_\infty = \sup_{x\in\mathfrak{X}} |f(x)|$. It is therefore natural to ask what is the ***topological dual*** of this Banach space, that is to say the set of linear functionals $\phi : \mathscr{C}_b(\mathfrak{X}) \to \mathbb{R}$ that are bounded in the sense that

$$\exists C(\phi),\ \ \forall f \in \mathscr{C}_b(\mathfrak{X}),\ \ |\phi(f)| \leq C(\phi)\, \|f\|_\infty.$$

This is quite a difficult question; see [**Dunford and Schwartz, 1988**, Theorem IV.6.2]. In the simpler case of a compact and Hausdorff space $\mathfrak{X}$, the well-known Riesz' theorem ensures that the dual of $\mathscr{C}_b(\mathfrak{X}) = \mathscr{C}(\mathfrak{X})$ can be identified with $\mathscr{M}(\mathfrak{X})$, the space of signed (bounded) measures; see *e.g.* [**Lang, 1993**, Chapter IX]. This can be generalized in several ways to the case of a locally compact Hausdorff space. Hence, looking for instance at the space $\mathscr{C}_0(\mathfrak{X})$ of continuous functions that vanish at infinity, under some additional topological assumptions on $\mathfrak{X}$, the dual of the Banach space is $\mathscr{M}(\mathfrak{X})$. However, the case of $\mathscr{C}_b(\mathfrak{X})$ is quite different and involves more general objects than measures. Before going on, we shall impose additional conditions on the space $\mathfrak{X}$, and always assume that they hold in the following.

Definition 1.8. *A topological space $\mathfrak{X}$ is called **polish** if:*

1. *The space is separable,* i.e., *there exists a dense countable subset $(a_m)_{m\in\mathbb{N}}$.*
2. *The space is metrizable, that is to say that there exists a distance d on $\mathfrak{X}$ such that a basis of neighborhoods of any point $x \in \mathfrak{X}$ consists in the open balls $B_{(x,\varepsilon)}$.*
3. *The distance d can be chosen so that $(\mathfrak{X}, d)$ is a complete metric space,* i.e., *every Cauchy sequence (with respect to d) has a limit.*

As a metric space, any polish space is in particular normal, which means that points are closed and that any two disjoint closed subsets can be separated by open subsets. One may ask why one does not deal directly with complete separable metric spaces, and only ask for "metrizable by a complete distance". The reason is that the space of states $\mathfrak{X}$ may admit a "natural distance" for which it is not complete, but such that the corresponding topology is metrizable by another distance which is complete but whose manipulation is quite cumbersome.

Example. Consider the open segment $\mathfrak{X} = (0, 1)$, endowed with the restriction of the natural distance $d(x, y) = |x - y|$ of the real line $\mathbb{R}$. This space is obviously separable as $\mathbb{Q}$ is dense in $\mathbb{R}$, but it is not complete with respect to d, since it is not closed in the complete metric space $(\mathbb{R}, d)$. However, the topologically equivalent distance

$$D(x, y) = \left|\tan\left(\frac{\pi(2x-1)}{2}\right) - \tan\left(\frac{\pi(2y-1)}{2}\right)\right|$$

makes $\mathfrak{X}$ into a complete metric space, because it comes from the homeomorphism $x \mapsto \tan\frac{\pi(2x-1)}{2}$ from $(0, 1)$ to $\mathbb{R}$ and from the complete usual distance on $\mathbb{R}$. So, $(0, 1)$ is a

polish space. But in practice, it will be easier to stick to the natural distance d, and we shall neither use D.

EXAMPLE. Another famous example of polish space without convenient complete metric is Skorohod's space $\mathscr{D}$ of *càdlàg* paths, see the discussion at the end of Chapter 2. This example is more convincing than $(0,1)$, because one could have dealt with the non-completeness of $((0,1),d)$ just by completing the space and looking at $[0,1]$. On the contrary, this completion is not manipulable in the case of Skorohod's space.

We now come back to the problem of the dual of $\mathscr{C}_b(\mathfrak{X})$, assuming that $\mathfrak{X}$ is a polish space (this assumptions holds till the end of the chapter). Call finitely additive measure on $\mathfrak{X}$ a function $\mu : \mathcal{B}(\mathfrak{X}) \to \mathbb{R}$ such that for all finite family $A_1, \dots, A_n$ of disjoint measurable subsets,

$$\mu\left(\bigsqcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i).$$

One does not ask here for countable additivity, which is the main difference with usual measures. A finitely additive measure is called bounded if for every subset A,

$$|\mu|(A) = \sup_{A=B_1 \sqcup B_2 \sqcup \cdots \sqcup B_n} \left(\sum_{i=1}^{n} |\mu|(A_i)\right) < \infty.$$

Endowed with the total variation $\|\mu\| = |\mu|(\mathfrak{X})$, the set $\mathscr{M}_{ba}(\mathfrak{X})$ of bounded and finitely additive measures on $\mathfrak{X}$ becomes a Banach space; of course it contains the space $\mathscr{M}(\mathfrak{X})$ of bounded (signed) measures. A closed subspace of $\mathscr{M}_{ba}(\mathfrak{X})$ is the set $\mathscr{M}_{rba}(\mathfrak{X})$ of ***regular bounded finitely additive measures***, which on top of the previous hypotheses satisfy:

$$\forall A \in \mathcal{B}(\mathfrak{X}),\ \ \forall \varepsilon > 0,\ \ \exists C \subset A \subset O \text{ with } C \text{ closed and } O \text{ open},\ \ \forall B \subset O \backslash C,\ \ |\mu|(B) \leq \varepsilon.$$

Beware of the difference between this notion of regularity and the notion of sets of continuity developed hereafter. In a similar way to Lebesgue's integration theory, one can define the integral of a bounded continuous function against a measure in $\mathscr{M}_{rba}(\mathfrak{X})$. Then:

THEOREM 1.9 (Dunford-Schwartz). *The dual of the Banach space $\mathscr{C}_b(\mathfrak{X})$ is exactly $\mathscr{M}_{rba}(\mathfrak{X})$. For a polish space, $\mathscr{M}_{rba}(\mathfrak{X})$ contains as a closed subset $\mathscr{M}^1(\mathfrak{X})$, the set of probability measures.*

The proof is quite technical and again we refer to [**Dunford and Schwartz, 1988**]. The second part of the statement of the theorem uses the following facts:

(1) Any probability measure on a metric (or metrizable) space is regular; then it is also obviously bounded and finitely additive.

(2) In $\mathscr{M}_{rba}(\mathfrak{X})$, the vector space of regular bounded countably additive measures is closed.

Beware that by "closed" we refer to the strong Banach space topology on $\mathscr{M}_{rba}(\mathfrak{X})$; as we shall explain in a moment, $\mathscr{M}^1(\mathfrak{X})$ is not closed for the convergence in law inside $\mathscr{M}_{rba}(\mathfrak{X})$.

1.3.2. Weak convergence of random variables. We are now ready to define the weak convergence of a sequence of random variables $(X_n)_{n\in\mathbb{N}}$. Denote $\mu_n = \mathbb{P}_{X_n}$ the law of the r.v. X_n, which is in $\mathscr{M}^1(\mathfrak{X})$; notice that we do not require the X_n's to be defined on a common probability space $(\Omega, \mathcal{B}, \mathbb{P})$, and that we shall only deal with the laws μ_n. The integral of a continuous bounded function f with respect to any probability measure μ with be denoted $\mu(f) = \int_{\mathfrak{X}} f(x)\,\mu(dx)$.

DEFINITION 1.10. *The sequence of random variables $(X_n)_{n\in\mathbb{N}}$ is said to* ***converge in law*** *towards a random variable X if for any bounded continuous function $f \in \mathscr{C}_b(\mathfrak{X})$,*

$$\lim_{n\to\infty} \mu_n(f) = \mu(f),$$

where μ is the law of X. We then write $X_n \rightharpoonup X$, and since the notion only depends on the laws, we may also write $\mu_n \rightharpoonup \mu$.

If the reader knows about weak topologies, he will see that the convergence of laws $\mu_n \rightharpoonup \mu$ is exactly the ***restriction of the weak-$\star$ topology*** induced by the duality $(\mathscr{C}_b(\mathfrak{X}), \mathscr{M}_{rba}(\mathfrak{X}))$ from $\mathscr{M}_{rba}(\mathfrak{X})$ to $\mathscr{M}^1(\mathfrak{X})$. Notice that the space $\mathscr{M}^1(\mathfrak{X})$ is not closed w.r.t. weak-$\star$ topology; actually this will be one of the main motivation for the notion of tightness developed in Chapter 2. A first important property of the convergence in law is the compatibility with continuous maps:

PROPOSITION 1.11. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of probability measures on a (polish) space $\mathfrak{X}$, with $\mu_n \rightharpoonup \mu$. For every continuous function $f : \mathfrak{X} \to \mathfrak{Y}$, the image laws satisfy $f_\star(\mu_n) \rightharpoonup f_\star(\mu)$.*

PROOF. Let g be a bounded continuous function on $\mathfrak{Y}$. Notice that $f_\star(\mu_n)(g) = \mu_n(g \circ f)$, and $g \circ f$ is a bounded continuous function on $\mathfrak{X}$. Therefore,

$$\lim_{n\to\infty} f_\star(\mu_n)(g) = \mu(g \circ f) = f_\star(\mu)(g)$$

which means that $f_\star(\mu_n) \rightharpoonup f_\star(\mu)$. □

THEOREM 1.12 (Portmanteau). *For probability measures $(\mu_n)_{n\in\mathbb{N}}$ and μ on $\mathfrak{X}$, the following assertions are equivalent:*

(i) $\mu_n \rightharpoonup \mu$.

(ii) for every bounded and uniformly continuous function f, $\lim_{n\to\infty} \mu_n(f) = \mu(f)$.

(iii) for every closed subset F, $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$.

(iv) for every open subset U, $\liminf_{n\to\infty} \mu_n(U) \geq \mu(U)$.

(v) if $\mu(A^\circ) = \mu(A) = \mu(\overline{A})$, then $\lim_{n\to\infty} \mu_n(A) = \mu(A)$.

PROOF. The implication $(i) \Rightarrow (ii)$ is obvious, and so is the equivalence $(iii) \Leftrightarrow (iv)$, since the complementary of a closed subset is an open subset, and $\nu(A^c) = 1 - \nu(A)$ for any probability measure ν.

For $(ii) \Rightarrow (iii)$, recall that the uniform continuity means that

$$\forall \varepsilon > 0,\ \ \exists \eta > 0,\ \ \forall x, y \in \mathfrak{X},\ \ d(x,y) \leq \eta \Rightarrow |f(x) - f(y)| \leq \varepsilon.$$

This is in particular satisfied if f is Lipschitz. Set

$$f_K(x) = (1 - K\, d(x,F))^+ = \max\{1 - K\, d(x,F), 0\},$$

where $K > 0$ and $d(x,F) = \inf\{d(x,y) \mid y \in F\}$. For any K, f_K is Lipschitz with constant K, whence uniformly continuous. Moreover, f_K takes its values in $[0,1]$, and $f_K(x) = 1$ if and only if $x \in F$. It follows that for any measure μ_n, $\mu_n(F) \leq \mu_n(f_K)$. Fix $\varepsilon > 0$. As K goes to infinity, f_K converges pointwise to the indicator of F, and this convergence is dominated by the constant function 1. Hence, $\lim_{K\to\infty} \mu(f_K) = \mu(F)$ and there is a $K = K_\varepsilon$ such that $\mu(f_K) \leq \mu(F) + \varepsilon$. Since f_K is uniformly continuous and bounded, $\lim_{n\to\infty} \mu_n(f_K) = \mu(f_K)$, so for $n \geq N$,

$$\mu_n(F) \leq \mu_n(f_K) \leq \mu(f_K) + \varepsilon \leq \mu(F) + 2\varepsilon,$$

and therefore $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F) + 2\varepsilon$. Since this is true for every $\varepsilon > 0$, one concludes that $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$.

The implication $(iii) + (iv) \Rightarrow (v)$ is easy. Suppose that A is a set of continuity of μ, which means that the measure of A is also the measure of its interior A° or the measure of its closure $\overline{A}$. Then,

$$\liminf_{n\to\infty} \mu_n(A) \geq \liminf_{n\to\infty} \mu_n(A^\circ) = \mu(A^\circ) = \mu(\overline{A}) \geq \limsup_{n\to\infty} \mu_n(\overline{A}) \geq \limsup_{n\to\infty} \mu_n(A),$$

so the limit of $\mu_n(A)$ exists and all the inequalities above are equalities. In particular, $\lim_{n\to\infty} \mu_n(A) = \mu(A)$.

Finally, assume (v) and let us prove that $\mu_n \rightharpoonup \mu$. Fix $f \in \mathscr{C}_b(\mathfrak{X})$; by linearity one can assume that f take its values in $[0,1]$. Then, one can write

$$\mu(f) = \int_0^1 \mu(\{x \mid f(x) > t\})\, dt$$

and similarly for the μ_n's. As f is continuous, the closure of $\{x \mid f(x) > t\}$ is the set $\{x \mid f(x) \geq t\}$, so if $\{x \mid f(x) > t\}$ is not a continuity set of μ, then $\mu(\{x \mid f(x) = t\} > 0$. This can happen only for a countable subset $\{t_1, t_2, \ldots\}$ of $[0,1]$. As a consequence, almost everywhere on $[0,1]$, the functions

$$x \mapsto \mu_n(\{x \mid f(x) > t\})$$

converge towards $\mu(\{x \mid f(x) > t\})$. Finally, by dominated convergence on $[0,1]$, $\mu(f) = \lim_{n\to\infty} \mu_n(f)$. □

COROLLARY 1.13. *Suppose $\mathfrak{X} = \mathbb{R}$, and assume for simplicity that μ has a continuous* ***cumulative distribution function*** *$x \mapsto F_\mu(x) = \mu((-\infty, x])$, which is the case for instance if μ is absolutely continuous w.r.t. the Lebesgue measure. Then, a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ converges in law to μ if and only if the cumulative distribution functions $F_{X_n}(x) = \mathbb{P}[X_n \leq x]$ converge pointwise to $F_\mu(x)$.*

REMARK. Without the hypothesis of continuity of F_μ, the condition is replaced by pointwise convergence of the F_{X_n}'s at every point of continuity of F_μ.

EXAMPLE. The classical central limit theorem ensures, that given a sequence of i.i.d. random variables $(X_n)_{n\in\mathbb{N}}$ in $\mathscr{L}^2(\Omega)$, the recentered and rescaled mean $\sqrt{n}\,(Z_n - \mathbb{E}[X])$ converges in law towards a Gaussian variable of mean 0 and variance $\mathrm{Var}[X]$.

For real valued random variables, another powerful criterion of convergence in law is the one actually used in the proof of the central limit theorem, namely, the so-called Lévy continuity theorem. Recall that the ***characteristic function*** of a probability measure $\mu \in \mathscr{M}^1(\mathbb{R})$ is $\Phi_\mu(\zeta) = \mu(\mathrm{e}^{\mathrm{i}\zeta X}) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\zeta x}\,\mu(dx)$.

THEOREM 1.14 (Lévy). *A sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to a probability measure μ if and only if, for every $\zeta \in \mathbb{R}$, $\lim_{n\to\infty} \Phi_{\mu_n}(\zeta) = \Phi_\mu(\zeta)$.*

We shall prove this theorem quite easily in Chapter 2 by using the theory of tightness. Now, let us see the connections between the convergence in law and the other modes of convergence. The main result is the following:

PROPOSITION 1.15. *Suppose $X_n \to_{\mathbb{P}} X$. Then, $X_n \rightharpoonup X$, and the converse is true if X is a constant random variable.*

PROOF. We use the third criterion of Portmanteau's theorem. Let F be a closed subset of $\mathfrak{X}$; we denote F^ε the ε-boundary of F, that is to say that

$$F^\varepsilon = \{x \in \mathfrak{X} \mid \exists y \in F,\ \ d(x,y) < \varepsilon\}.$$

For $\varepsilon > 0$ fixed, if $X_n \to_{\mathbb{P}} X$, then for n big enough the probability that $d(X_n, X) > \varepsilon$ is arbitrary small, say, smaller than ε. Thus,

$$\begin{aligned}\mu_n(F) = \mathbb{P}[X_n \in F] &\le \mathbb{P}[X_n \in F \text{ and } d(X_n,X) \le \varepsilon] + \mathbb{P}[d(X_n,X) > \varepsilon] \\ &\le \mathbb{P}[X \in F^\varepsilon] + \varepsilon = \mu(F^\varepsilon) + \varepsilon\end{aligned}$$

for $n \ge N(\varepsilon)$. Now, since $F = \bigcap_{\varepsilon\downarrow 0} F^\varepsilon$ as a closed set, $\lim_{\varepsilon\to 0} \mu(F^\varepsilon) = \mu(F)$, so for a given $\varepsilon' > 0$, there exists $\varepsilon \in (0, \varepsilon')$ such that $\mu(F^\varepsilon) \le \mu(F) + \varepsilon'$. If follows that for any $\varepsilon' > 0$,

$$\limsup_{n\to\infty} \mu_n(F) \le \mu(F) + 2\varepsilon'$$

and the supremum limit is even smaller than $\mu(F)$ since the previous inequality is true for any $\varepsilon' > 0$. We have therefore shown that $X_n \rightharpoonup X$. Suppose now that $X_n \rightharpoonup a$; considering an open ball $B_{(a,\varepsilon)}$, we have by Portmanteau's theorem that

$$\liminf_{n\to\infty} \mu_n(B_{(a,\varepsilon)}) \ge \delta_a(B_{(a,\varepsilon)}) = 1,$$

which can be rewritten as : $\lim_{n\to\infty} \mathbb{P}[d(X_n, X) < \varepsilon] = 1$ for any $\varepsilon > 0$. So in the case of a constant, the converse implication is true. □

Thus, we have the chain of implications $(X_n \to_{\text{a.s.}} X) \Rightarrow (X_n \to_{\mathbb{P}} X) \Rightarrow (X_n \rightharpoonup X)$.

REMARK. It can actually be shown that conversely, if $\mu_n \rightharpoonup \mu$ on a polish space $\mathfrak{X}$, then there exists a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and random variables $X_n, X : \Omega \to \mathfrak{X}$ such that X_n has for law μ_n, X has for law μ, and $X_n \to_{\text{a.s.}} X$; see [**Billingsley, 1999**, Theorem 6.7]. So in some sense one does not lose much between the notions of almost sure convergence and weak convergence.

1.3.3. Prohorov's topology. As we have defined a notion of (weak) convergence in the space of probability measures, it is now interesting to study the corresponding topology; ultimately this problem will be solved in §2.1, but we can already state basic results.

PROPOSITION 1.16 (Prohorov). *There exists a distance d on $\mathscr{M}^1(\mathfrak{X})$ such that $\mu_n \rightharpoonup \mu$ if and only if $d(\mu_n, \mu) \to 0$. Consequently, the topology of weak convergence of probability measures is metrizable.*

PROOF. Define the ***Prohorov-Lévy metric*** by

$$d(\mu,\nu) = \inf\{\varepsilon > 0,\ \ \forall A \in \mathcal{B}(\mathfrak{X}),\ \ \mu(A) \leq \nu(A^\varepsilon) + \varepsilon \text{ and } \nu(A) \leq \mu(A^\varepsilon) + \varepsilon\},$$

where as before A^ε denotes the ε-boundary of A. This is well-defined as obviously $\varepsilon = 1$ is in the set considered. Let us first verify that d is a distance on $\mathscr{M}^1(\mathfrak{X})$. If $d(\mu,\nu) = 0$, then for any Borel set A,

$$\mu(A) \leq \inf\{\nu(A^\varepsilon),\ \varepsilon > 0\} = \nu(\overline{A}) \quad \text{and} \quad \nu(A) \leq \mu(\overline{A}).$$

In particular, μ and ν agree on closed sets, and as is well-known from measure theory, this ensures that $\mu = \nu$. The symmetry of d is obvious, and for the triangular inequality, we proceed as follows. Let μ, ν, ρ be three probability measures, and ε and η be such that $d(\mu,\nu) < \varepsilon$ and $d(\nu,\rho) < \eta$. If A is a Borel set, then

$$\mu(A) \leq \nu(A^\varepsilon) + \varepsilon \leq \rho((A^\varepsilon)^\eta) + \eta + \varepsilon,$$

and by the triangular inequality, $(A^\varepsilon)^\eta \subset A^{\varepsilon+\eta}$, so $\varepsilon + \eta$ is a correct bound on $d(\mu,\rho)$ — by symmetry the previous argument also works from ρ to μ. Since this is true for any $\varepsilon > d(\mu,\nu)$ and any $\eta > d(\nu,\rho)$, we conclude that

$$d(\mu,\rho) \leq d(\mu,\nu) + d(\nu,\rho),$$

which ends the proof that d is a distance. It should be noticed that the Prohorov-Lévy metric could have been defined by the simpler equation

$$d(\mu,\nu) = \inf\{\varepsilon > 0,\ \ \forall A \in \mathcal{B}(\mathfrak{X}),\ \ \mu(A) \leq \nu(A^\varepsilon) + \varepsilon\},$$

which does not seem a priori symmetric in μ and ν. Indeed, suppose that $\mu(A) \leq \nu(A^\varepsilon) + \varepsilon$ for every Borelian subset. Then,

$$\mu(A^\varepsilon) = 1 - \mu((A^\varepsilon)^{\mathrm{c}}) \geq 1 - \nu(((A^\varepsilon)^{\mathrm{c}})^\varepsilon) - \varepsilon$$

and $x \in ((A^\varepsilon)^{\mathrm{c}})^\varepsilon$ means that there is a y such that $d(x,y) < \varepsilon$, and $d(y,A) \geq \varepsilon$. By the triangular inequality x cannot be in A, so $((A^\varepsilon)^{\mathrm{c}})^\varepsilon \subset A^{\mathrm{c}}$ and $1 - \nu(((A^\varepsilon)^{\mathrm{c}})^\varepsilon) \geq \nu(A)$. Thus, one also has $\nu(A) \leq \mu(A^\varepsilon) + \varepsilon$ for every $A \in \mathcal{B}(\mathfrak{X})$.

Now, suppose that $d(\mu_n,\mu)$ converges to 0, and fix a Borelian subset A which is a continuity set for μ. If $\varepsilon'>0$ is fixed, there is an $\varepsilon\in(0,\varepsilon')$ such that $\mu(A)\leq\mu(A^\varepsilon)\leq\mu(A)+\varepsilon'$. Then, for n big enough, $d(\mu_n,\mu)\leq\varepsilon$, so

$$\mu_n(A)\leq\mu(A^\varepsilon)+\varepsilon\leq\mu(A)+2\varepsilon'.$$

Looking at A^{c} which is also a continuity set for μ, one gets a converse inequality, so $|\mu_n(A)-\mu(A)|\leq 2\varepsilon'$ for n big enough, and by Theorem 1.12 this ensures the convergence in law. Conversely, assume $\mu_n\rightharpoonup\mu$. Since the space is assumed separable, there is a sequence $(a_m)_{m\in\mathbb{N}}$ such that for every $\varepsilon>0$,

$$\mathfrak{X}=\bigcup_{m\in\mathbb{N}}B_{(a_m,\varepsilon)}.$$

Fix $\varepsilon>0$ and let M be a finite subset of $\mathbb{N}$ such that $\mu(\bigcup_{m\in M}B_{(a_m,\varepsilon)})$ is already bigger than $1-\varepsilon$. The set $\mathcal{F}$ of all open subsets F of type

$$(B_{(a_{i_1},\varepsilon)}\cup\cdots\cup B_{(a_{i_r},\varepsilon)})^\varepsilon\quad\text{with }\{i_1,\ldots,i_m\}\subset M$$

is finite, so for n big enough, $\mu_n(F)\geq\mu(F)-\varepsilon$ for every $F\in\mathcal{F}$ — this is assertion (iv) of Portmanteau's theorem. Now, fix $A\in\mathcal{B}(X)$, and denote

$$F(A)=\left(\bigcup_{m\in M,\ B_{(a_m,\varepsilon)}\cap A\neq\emptyset}B_{(a_m,\varepsilon)}\right)^\varepsilon;$$

this is an element in $\mathcal{F}$, and the μ-measure of the complementary of $F(A)$ in A is smaller than ε. On the other hand, $F(A)\subset A^{3\varepsilon}$. So,

$$\mu(A)\leq\mu(F(A))+\mu(A\setminus F(A))\leq\mu_n(F(A))+2\varepsilon\leq\mu_n(A^{3\varepsilon})+3\varepsilon$$

for n big enough, and by a previous remark this is sufficient to assert that $d(\mu_n,\mu)\leq 3\varepsilon$ or n big enough. □

Proposition 1.17. *Starting from a polish (or even metric separable) space $\mathfrak{X}$, the space of probability measures $\mathscr{M}^1(\mathfrak{X})$ is also separable.*

Proof. Fix a dense sequence $(a_m)_{m\in\mathbb{N}}$ in $\mathfrak{X}$, and denote $\mathscr{E}$ the set of probability measures that are rational linear combinations $\sum_{m\in M}r_m\,\delta_{a_m}$ of the Dirac measures at the points a_m; we aim to prove that $\mathscr{E}$, which is countable, is dense in $\mathscr{M}^1(\mathfrak{X})$. We fix $\mu\in\mathscr{M}^1(\mathfrak{X})$ and $\varepsilon>0$, and as in the proof of the previous proposition, we choose M finite subset of $\mathbb{N}$ such that $\mu(\bigcup_{m\in M}B_{(a_m,\varepsilon)})$ is bigger than $1-\varepsilon$. By removing intersections, one can then choose open subsets $F_m\subset B_{(a_m,\varepsilon)}$ such that

$$\bigcup_{m\in M}B_{(a_m,\varepsilon)}=\bigsqcup_{m\in M}F_m.$$

In each F_m we can assume that there is still the point a_m. Now, fix a Borelian subset A, and denote $F(A)=\bigcup_{m\in M,\ F_m\cap A\neq\emptyset}F_m$; as before,

$$\mu(A)\leq\mu(F(A))+\mu(A\setminus F(A))\leq\left(\sum_{m\in M,\ F_m\cap A\neq\emptyset}\mu(F_m)\right)+\varepsilon.$$

We then take in $\mathscr{E}$ the measure $\pi = \sum_{m\in M} r_m\,\delta_{a_m}$, where $\sum_{m\in M}|r_m - \mu(F_m)| \le 2\varepsilon$ and $\sum_{m\in M} r_m = 1$; this is always possible (and with rational numbers). The previous inequality rewrites then as

$$\mu(A) \le \left(\sum_{m\in M,\ F_m\cap A\neq\emptyset} r_m\right) + 3\varepsilon = \pi(F(A)) + 3\varepsilon \le \pi(A^{3\varepsilon}) + 3\varepsilon,$$

so $d(\mu,\pi)\le 3\varepsilon$ and we have shown the density of $\mathscr{E}$ in $\mathscr{M}^1(\mathfrak{X})$. □

Hence, we have shown that starting from a separable metrizable space $\mathfrak{X}$ (the completeness has not been used so far), the space of probability measures $\mathscr{M}^1(\mathfrak{X})$ was also separable and metrizable for the topology of convergence in law. This result will be completed in the next chapter, where it will be shown that completeness is also conserved. We have now ended our presentation of the modes of probabilistic convergence in which we will be interested, and what we lack basically is a characterization of compactness w.r.t. these convergences, that is to say abstract criterions that guarantee the existence of limits of (sub)sequences. This will prove extremely useful in order to construct complicate random objects by approximation; *cf.* §2.2 where the example of Brownian motions will be treated. Before going on, let us mention that there exists other mode of convergences for random variables or their laws. In particular, one can look at the strong convergence on $\mathscr{M}^1(\mathfrak{X})$ (restricted from $\mathscr{M}_{rba}(\mathfrak{X})$), given by the ***total variation distance***

$$d_{\mathrm{TV}}(\mu,\nu) = \sup_{A\in\mathcal{B}(\mathfrak{X})} |\mu(A)-\nu(A)| \ \in [0,1].$$

Unless $\mathfrak{X}$ is a finite set, the convergence in total variation distance is much stronger than the convergence in law: for instance, if $x_n\to x$ in $\mathbb{R}$, then $\delta_{x_n}\rightharpoonup\delta_x$ (use for instance the fifth criterion in Theorem 1.12), but $d_{\mathrm{TV}}(\delta_{x_n},\delta_x)=1$ if $x_n\neq x$ for every $n\in\mathbb{N}$. However, the total variation distance is often used to measure the weak convergence on a finite (combinatorial) set, *e.g.*, the convergence to stationarity of a Markov chain (see §3.2). It is then much more sensible than Prohorov's metric: indeed, on a large finite set $\mathfrak{X}$, even if $\mu_n\rightharpoonup\mu$, one can still find "exceptional" events A such that $|\mu_n(A)-\mu(A)|$ stays large for quite a long time.

If μ and ν are absolutely continuous with respect to a third probability measure ρ, then their total variation distance can be seen as a $\mathscr{L}^1$-norm:

$$d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\int_{\mathfrak{X}} \left|\frac{d\mu(x)}{d\rho} - \frac{d\nu(x)}{d\rho}\right| \rho(dx).$$

Indeed, it is easily seen that $d_{\mathrm{TV}}(\mu,\nu)$ is attained on the set $A=\{x : \frac{d\mu(x)}{d\rho}\ge\frac{d\nu(x)}{d\rho}\}$, or its complementary (up to a set of $(\mu+\nu)$-measure zero).

As a consequence,

$$\begin{aligned} d_{\mathrm{TV}}(\mu,\nu) &= |\mu(A)-\nu(A)| = \frac{1}{2}\left(|\mu(A)-\nu(A)| + |\mu(A^{\mathrm{c}})-\nu(A^{\mathrm{c}})|\right) \\ &= \frac{1}{2}\int_A \frac{d\mu(x)}{d\rho} - \frac{d\nu(x)}{d\rho}\,\rho(dx) + \frac{1}{2}\int_{A^{\mathrm{c}}} \frac{d\nu(x)}{d\rho} - \frac{d\mu(x)}{d\rho}\,\rho(dx) \\ &= \frac{1}{2}\int_{A\sqcup A^{\mathrm{c}}} \left|\frac{d\mu(x)}{d\rho} - \frac{d\nu(x)}{d\rho}\right| \rho(dx) = \frac{1}{2}\int_{\mathfrak{X}} \left|\frac{d\mu(x)}{d\rho} - \frac{d\nu(x)}{d\rho}\right| \rho(dx). \end{aligned}$$

Most of the time one takes the reference measure ρ to be one of the measure μ of ν, say ν, so one looks at the $\mathscr{L}^1$-norm of $\frac{d\mu}{d\nu} - 1$, defined to be $+\infty$ if μ is not absolutely continuous with respect to ν. A natural generalization of the total variation distance is then of course the $\mathscr{L}^p$-distance

$$\left(\int_{\mathfrak{X}} \left|\frac{d\mu(x)}{d\nu} - 1\right|^p \nu(dx)\right)^{\frac{1}{p}},$$

again defined to be $+\infty$ if μ is not absolutely continuous w.r.t. ν. Notice that these notions are quite different from the $\mathscr{L}^p$-convergences studied in §1.2. However, for every exponent $p \in [1,+\infty)$, if $\mu_n \to_{\mathscr{L}^p(\mathfrak{X},d\mu)} \mu$, then $\mu_n \rightharpoonup \mu$.

CHAPTER 2

Compactness and tightness

If X is a topological space, recall that it is called ***compact*** if every covering $X = \bigcup_{i\in I} U_i$ by open subsets U_i has a finite sub-cover $X = \bigcup_{i\in J\subset I} U_i$. On a metric or metrizable space, this is equivalent to the sequential compactness: every sequence $(x_n)_{n\in\mathbb{N}}$ in X has a convergent subsequence $(x_{\phi(n)})_{n\in\mathbb{N}}$. Suppose now X Hausdorff (this means that distinct points can be separated by open subsets), or even metrizable. Then a compact subset $Y \subset X$ is necessarily closed, which leads to the notion of relative compactness: a subset Y is relatively compact if its closure $\overline{Y}$ is compact in X (for the induced topology). On a metrizable space, this is equivalent to the following statement: every sequence $(y_n)_{n\in\mathbb{N}}$ in Y has a convergent subsequence, with its limit in $\overline{Y}$. For instance, the relatively compact subsets of $\mathbb{R}^N$ are exactly the bounded subsets by Bolzano-Weierstrass theorem. This notion is extremely useful in functional analysis, for the following reason. Assume that X is a metrizable space of "functions", for instance, the space of all continuous functions on $[0,1]$ (endowed with the sup norm). Consider now $(x_n)_{n\in\mathbb{N}}$ a sequence in X that is relatively compact, and such that all the convergent subsequences of $(x_n)_{n\in\mathbb{N}}$ have the same limit x (we shall see in this chapter many situations where this unicity is guaranteed). Then:

LEMMA 2.1. *In the previous situation, assuming only X metrizable, $x_n \to x$.*

PROOF. In the converse situation, there exists $\varepsilon > 0$ and a subsequence $(x_{\phi(n)})_{n\in\mathbb{N}}$ such that $x_{\phi(n)} \notin B_{(x,\varepsilon)}$ for every $n \in \mathbb{N}$. Since $(x_n)_{n\in\mathbb{N}}$ is relatively compact, there is a further subsequence $(x_{\phi\circ\psi(n)})_{n\in\mathbb{N}}$ that is convergent, and by hypothesis its limit is x since it is extracted from $(x_n)_{n\in\mathbb{N}}$: contradiction. □

This lemma is of prime importance in approximation theory: in order to construct a complicate object in a functional space (space of functions, or space of measures, *etc.*), it suffices now to find a sequence $(x_n)_{n\in\mathbb{N}}$ that is built in such a way that the unicity of the limit is ensured, and to prove the relative compactness of the sequence (which most of the time amounts to prove that the sequence is "correctly bounded", in a sense to be precised hereafter in each case). In particular, in a probabilistic setting, this method allows one to construct very subtle random objects, *e.g.*, random trees, surfaces, continuous paths, *etc.*; we shall detail in §2.2 the example of Brownian motions.

The main problem is then to give useful criterions of relative compactness in the usual functional spaces. In this chapter, we shall do it for:

(1) the space of probability measures $\mathscr{M}^1(\mathfrak{X})$ over any polish space; this leads to the notion of tightness, and will allow us to have a better understanding of the weak topology (§2.1).

(2) the space of continuous functions $\mathscr{C}(X)$ on a separable locally compact space, see §2.2.

In Section 2.3, we shall also explain a generalization of the discussion of Section 2.2 to possibly non-continuous random paths; for this section we won't give all details and we refer to [**Billingsley, 1999**, Chapter 3].

2.1. Compactness in $\mathscr{M}^1(\mathfrak{X})$ and tightness

If $\mathfrak{X} = [\![1, N]\!]$ is a finite set, then $\mathscr{M}^1(\mathfrak{X})$ is a simplex in $\mathbb{R}^N$ and it is obviously compact for the strong topology, which is the same as the weak-$\star$ topology corresponding to the convergence in law. However, in the general case of a polish space $\mathfrak{X}$, there is no hope to have this compactness. Indeed, assuming for instance $\mathfrak{X} = \mathbb{R}$, a sequence $(\delta_{x_n})_{n\in\mathbb{N}}$ of Dirac masses at points x_n such that $x_n \to +\infty$ has no limit in $\mathscr{M}^1(\mathbb{R})$ with respect to the convergence in law. Indeed, if it were the case, then the limit μ would satisfy $\mu(U) = 0$ for every open set $U = (-n, n)$, and then, by countable additivity, for every Borelian subset,

$$\mu(A) = \lim_{n\to\infty} \mu(A \cap (-n, n)) = 0,$$

which is not possible for $\mu \in \mathscr{M}^1(\mathbb{R})$ since $\mu(\mathbb{R}) = 1$. In particular, the sequence $(\delta_n)_{n\in\mathbb{N}}$ cannot have any convergent subsequence in $\mathscr{M}^1(\mathbb{R})$.

However, $(\delta_n)_{n\in\mathbb{N}}$ has a convergent subsequence in $\mathscr{M}_{rba}(\mathbb{R})$. Indeed, by Banach-Alaoglu's theorem, a closed ball for the norm in the dual of a Banach space is always weak-$*$ compact, so $\mathscr{M}^1(\mathfrak{X})$ is weak-$*$-relatively compact inside $\mathscr{M}_{rba}(\mathfrak{X})$, the dual of $\mathscr{C}_b(\mathfrak{X})$. But as we have just seen, it is not closed, and one may therefore asks for a criterion of relative compactness inside $\mathscr{M}^1(\mathfrak{X})$; or in other words, a criterion that ensures that the limits of convergent sequences stay in $\mathscr{M}^1(\mathfrak{X})$.

2.1.1. Tightness and Prohorov's theorem.

Definition 2.2. *A family of probability measures $\mathscr{P} \subset \mathscr{M}^1(\mathfrak{X})$ is called* ***tight*** *if for every $\varepsilon > 0$, there is a compact subset K of $\mathfrak{X}$ such that*

$$\mu(K^{\mathrm{c}}) \leq \varepsilon \quad \textit{for every } \mu \in \mathscr{P}.$$

Example. A single probability measure μ on a polish space is always tight. Indeed, consider a dense sequence $(a_m)_{m\in\mathbb{N}}$ in $\mathfrak{X}$; since $\bigcup_{m\in\mathbb{N}} B_{(a_m,\eta)} = \mathfrak{X}$ for every $\eta > 0$, for every $\varepsilon > 0$ and every $m \geq 0$, there is an index $\phi(m)$ such that

$$\mu\left(\bigcup_{m'\leq\phi(m)} B_{(a_{m'},2^{-m})}\right) \geq 1 - \frac{\varepsilon}{2^m}.$$

Consider then the closure of $Y = \bigcap_{m\geq 0}\bigcup_{m'\leq\phi(m)} B_{(a_{m'},2^{-m})}$; its μ-measure is bigger than $1-\sum_{m\geq 0}\frac{\varepsilon}{2^m} = 1-2\varepsilon$. Let $(y_n)_{n\in\mathbb{N}}$ be a sequence in Y. For every $m\geq 0$, there is an subsequence of $(y_n)_{n\in\mathbb{N}}$ such that all the terms fall in the same ball $B_{(a_{m'},2^{-m})}$ with $m'\leq\phi(m)$. Hence, by diagonal extraction, one can construct a subsequence $(y_{\Psi(n)})_{n\in\mathbb{N}}$ such that

$$\forall n_1,n_2\geq N,\ \ |y_{\Psi(n_1)}-y_{\Psi(n_2)}|\leq\frac{1}{2^{N-1}}.$$

Then, the extracted subsequence is Cauchy, and since $\mathfrak{X}$ is polish, convergent. So, $\overline{Y}$ is compact and the tightness is shown.

EXAMPLE. The sequence $(\delta_n)_{n\in\mathbb{N}}$ is not tight in $\mathscr{M}^1(\mathbb{R})$, since for every compact set K there is an n such that $\delta_n(K)=0$.

THEOREM 2.3 (Prohorov). *A class of probability measures $\mathscr{P}$ over a polish space is tight if and only if it is relatively compact for the topology of weak convergence.*

PROOF. The easy thing to prove is: relative compactness implies tightness. Indeed, as in the case of a single probability measure, for every $\varepsilon>0$ and every $m\geq 0$, there is an index $\phi(m)$ such that

$$\mu\left(\bigcup_{m'\leq\phi(m)} B_{(a_{m'},2^{-m})}\right)\geq 1-\frac{\varepsilon}{2^m}$$

for every μ in the relatively compact class $\mathscr{P}$. Otherwise, one could construct a sequence $(\mu_n)_{n\in\mathbb{N}}$ in $\mathscr{P}$ such that

$$\mu_n\left(\bigcup_{m'\leq n} B_{(a_{m'},2^{-m})}\right)<1-\frac{\varepsilon}{2^m}$$

and by relative compactness and the forth criterion in Theorem 1.12, a limit μ of a convergent subsequence would satisfy $\mu(\mathfrak{X})\leq 1-\frac{\varepsilon}{2^m}$, which is absurd. Then the same proof as in the case of a single measure ensures the existence of a compact set such that $\mu(K)\geq 1-\varepsilon$ for every fixed $\varepsilon>0$ and any $\mu\in\mathscr{P}$.

The other direction is more difficult, and we shall admit the following facts that have previously been discussed:

(1) Suppose that X is a compact topological space. Then the topological dual of the Banach space $\mathscr{C}(\mathfrak{X})$ is $\mathscr{M}(\mathfrak{X})$, the set of bounded signed (countably additive) measures (this is ***Riesz' representation theorem***).

(2) Let B be any Banach space, and $B^\star$ be its topological dual. One endows $B^\star$ with the weak-$\star$ topology: a basis of neighborhoods of $\phi\in B^\star$ is given by finite intersections of sets

$$U^x_{\phi,\varepsilon}=\{\psi\in B^\star\mid|\psi(x)-\phi(x)|<\varepsilon\}$$

with $\varepsilon > 0$ and $x \in B$. Then the unit ball $\{\phi \in B^\star \mid \forall x \in B,\ |\phi(x)| \leq \|x\|\}$ is topologically compact for the weak-$\star$ topology, and on the other hand it is metrizable if B is separable (this is the ***Banach-Alaoglu theorem***).

Now the first step in the proof of the remaining half of Prohorov's theorem is the following statement: if $\mathfrak{X}$ is a compact space, then $\mathscr{M}^1(\mathfrak{X})$ is also compact. This is a direct consequence of the previous results: indeed, $\mathscr{M}^1(\mathfrak{X})$ is a subset of the unit ball of $\mathscr{M}(\mathfrak{X})$, which is compact by Banach-Alaoglu's theorem since $\mathscr{M}(\mathfrak{X}) = (\mathscr{C}(\mathfrak{X}))^\star$. Thus, it suffices to show that $\mathscr{M}^1(\mathfrak{X})$ is closed in $\mathscr{M}(\mathfrak{X})$ w.r.t. the weak-$\star$ topology (convergence in law). But it writes as the intersection of the sets

$$\{\mu \in \mathscr{M}(\mathfrak{X}) \mid \mu(\mathfrak{X}) = 1\}$$

and

$$\{\mu \in \mathscr{M}(\mathfrak{X}) \mid \mu(f) \geq 0\} \quad \text{for } f \in \mathscr{C}(\mathfrak{X}) \text{ non-negative},$$

which are all closed.

Now, suppose only that $\mathfrak{X}$ is polish, and fix a dense sequence $(a_m)_{m\in\mathbb{N}}$ and a complete metric d. The map

$$\begin{aligned} \psi : \mathfrak{X} &\to \mathfrak{W} = [0,1]^{\mathbb{N}} \\ x &\mapsto (\min(d(x,a_m),1))_{m\in\mathbb{N}} \end{aligned}$$

is an homeomorphism of $\mathfrak{X}$ into a compact space. Endowed with the convergence of all coordinates, $[0,1]^{\mathbb{N}}$ is indeed compact (use diagonal extraction, or even Tychonoff's theorem), and even metrized by

$$\delta((v_n)_{n\in\mathbb{N}},(w_n)_{n\in\mathbb{N}}) = \sum_{n=0}^{\infty} \frac{1}{2^n} |v_n - w_n|.$$

If $d(x,y) \leq \varepsilon$, then for every $m \in \mathbb{N}$,

$$|d(x,a_m) - d(y,a_m)| \leq \varepsilon \qquad ; \qquad |\min(d(x,a_m),1) - \min(d(y,a_m),1)| \leq \varepsilon,$$

and therefore $\delta(\psi(x),\psi(y)) \leq \sum_{n=0}^{\infty} \frac{\varepsilon}{2^n} = 2\varepsilon$ and ψ is continuous and even Lipschitz. Assume now $\psi(x) = \psi(y)$. Then, for every m, $d(x,a_m) = d(y,a_m)$, so taking a subsequence $(a_{\phi_m})_{m\in\mathbb{N}}$ that converges to x in $\mathfrak{X}$, one gets $d(x,x) = 0 = d(y,x)$ by continuity of the distance, and therefore $x = y$. So, ψ realizes an embedding of $\mathfrak{X}$ into $\mathfrak{W}$. To see that it is an homeomorphism, one has to show that

$$x_n \to x \iff \psi(x_n) \to \psi(x).$$

The implication $\Rightarrow$ is clear: if $x_n \to x$, then $d(x_n,a_m) \to d(x,a_m)$ for every m, so $\psi(x_n) \to \psi(x)$. Conversely, if $x_n \not\to x$, then there is a subsequence $(x_{\phi(n)})_{n\in\mathbb{N}}$ and $\varepsilon > 0$ such that $d(x_{\phi(n)},x) \geq \varepsilon$ for every n. Take a_m such that $d(x,a_m) \leq \frac{\varepsilon}{3}$; then $d(x_{\phi(n)},a_m) \geq \frac{2\varepsilon}{3}$ for every n, and therefore

$$\delta(\psi(x_{\phi(n)}),\psi(x)) \geq \frac{1}{2^m}\left(d(x_{\phi(n)},a_m) - d(x,a_m)\right) \geq \frac{\varepsilon}{3\,2^m}.$$

It follows that $\psi(x_n) \not\to \psi(x)$, and we have shown that ψ was an homeomorphism onto its image.

Consider finally a tight class of probability measures $\mathscr{P} \subset \mathscr{M}^1(\mathfrak{X})$. Denote ψ_* the map from $\mathscr{M}^1(\mathfrak{X})$ to $\mathscr{M}^1(\mathfrak{W})$ which associates to a measure μ the probability measure $\psi_*(\mu) = \mu \circ \psi^{-1}$. Given a sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ in $\mathscr{P}$, there is a convergent subsequence of measures $(\psi_*(\mu_{\phi(n)}))_{n\in\mathbb{N}}$ in $\mathscr{M}^1(\mathfrak{W})$, because this space is compact. Denote ν its limit. Since $\mathscr{P}$ is tight, there is for every $m \geq 0$ a compact K_m in $\mathfrak{X}$ such that

$$\forall n \in \mathbb{N}, \;\; \mu_{\phi(n)}(K_m) \geq 1 - \frac{1}{2^m}.$$

As a consequence, $\psi^*(\mu_{\phi(n)})(\psi(K_m)) \geq 1 - \frac{1}{2^m}$ and since ψ is an homeomorphism, $\psi(K_m)$ is compact inside $\mathfrak{W}$, whence closed. By Portmanteau's theorem, it follows that

$$\nu(\psi(K_m)) \geq \limsup_{n\to\infty} \psi_*(\mu_{\phi(n)})(\psi(K_m)) = \limsup_{n\to\infty} \mu_{\phi(n)}(K_m) \geq 1 - \frac{1}{2^m},$$

and therefore, $\nu(\psi(\mathfrak{X})) \geq \nu(\psi(\bigcup_{m\in\mathbb{N}} K_m))) \geq 1$, *i.e.*, $\nu \in \psi_*(\mathscr{M}^1(\mathfrak{X}))$. We have therefore a natural candidate for a limit of $(\mu_{\phi(n)})_{n\in\mathbb{N}}$ in $\mathscr{M}^1(\mathfrak{X})$, namely, $\mu = (\psi^{-1})_*(\nu)$. Fix a closed subset $F \subset \mathfrak{X}$ and $\varepsilon > 0$. Since ψ is an homeomorphism, $\psi(F)$ is closed in $\psi(\mathfrak{X})$, which means that there is a closed subset G of $\mathfrak{W}$ such that $\psi(F) = G \cap \psi(\mathfrak{X})$ — beware that $\psi(F)$ needs not to be closed in $\mathfrak{W}$. Then,

$$\begin{aligned}\limsup_{n\to\infty} \mu_{\phi(n)}(F) &= \limsup_{n\to\infty} \psi_*(\mu_{\phi(n)})(\psi(F)) = \limsup_{n\to\infty} \psi_*(\mu_{\phi(n)})(G) \\ &\leq \nu(G) = \nu(\psi(F)) = \mu(F),\end{aligned}$$

so we have finally proved that $\mu_{\phi(n)} \rightharpoonup \mu$ and that $\mathscr{P}$ is relatively compact. □

REMARK. In our proof of Prohorov's theorem, we have seen that if $\mathfrak{X}$ is compact, then the same holds for $\mathscr{M}^1(\mathfrak{X})$. In fact, the converse is true. Indeed, if $\mathscr{M}^1(\mathfrak{X})$ is compact, then

$$\begin{aligned}\mathfrak{X} &\to \mathscr{M}^1(\mathfrak{X}) \\ x &\mapsto \delta_x\end{aligned}$$

is an homeomorphism into a closed subset of $\mathscr{M}^1(\mathfrak{X})$, so $\mathfrak{X}$ is compact.

2.1.2. Completeness of the space of measures. The rest of the chapter is devoted to consequences and corollaries of Prohorov's theorem 2.3. A first abstract consequence is the following:

PROPOSITION 2.4. *The space of probability measures $\mathscr{M}^1(\mathfrak{X})$ on a polish space is itself polish with respect to the topology of convergence in law.*

PROOF. We already know that $\mathscr{M}^1(\mathfrak{X})$ is separable and metrizable, so the only thing that needs to be proved is the completeness, and in view of Theorem 2.3, it suffices to show that a Cauchy sequence $(\mu_n)_{n\in\mathbb{N}}$ in $\mathscr{M}^1(\mathfrak{X})$ is tight. Then, two possible limits μ_1 and μ_2 of convergent subsequences $(\mu_{\phi_1(n)})_{n\in\mathbb{N}}$ and $(\mu_{\phi_2(n)})_{n\in\mathbb{N}}$ satisfy for every $\varepsilon > 0$

$$\exists N, \;\; \forall n_1, n_2 \geq N, \;\; d_{\mathrm{PL}}(\mu_{\phi_1(n)}, \mu_{\phi_2(n)}) \leq \varepsilon$$

where d_{PL} is the Prohorov-Lévy metric. With $n_1, n_2 \to \infty$, one has therefore $d_{\mathrm{PL}}(\mu_1, \mu_2) \leq \varepsilon$ for every ε, so $\mu_1 = \mu_2$, and by unicity of the limit the Cauchy sequence has then to converge.

As in the proof of the tightness of a single probability measure, the following is a sufficient criterion for tightness: for every $m \geq 0$, there is a finite family of open balls $(B_{(x_{m,m'},2^{-m})})_{m' \leq \phi(m)}$ of radius $\frac{1}{2^m}$ such that

$$\mu\left(\bigcup_{m' \leq \phi(m)} B_{(x_{m,m'},2^{-m})}\right) \geq 1 - \frac{1}{2^m}$$

for every μ in the family. Fix $m \geq 0$ and $N \geq 1$ such that $d(\mu_{n_1}, \mu_{n_2}) \leq \frac{1}{2^m}$ for $n_1, n_2 \leq N$. There is a finite family of balls $B_{(x_{m,m'},2^{-m-1})}$ such that for all $n \leq N$,

$$\mu_n\left(\bigcup_{m' \leq \phi(m)} B_{(x_{m,m'},2^{-m-1})}\right) \geq 1 - \frac{1}{2^{m+1}}.$$

Since $\bigcup_{m' \leq \phi(m)} B_{(x_{m,m'},2^{-m})}$ is included in the $\frac{1}{2^{m+1}}$-boundary of $\bigcup_{m' \leq \phi(m)} B_{(x_{m,m'},2^{-m-1})}$, for $n \geq N$,

$$\mu_n\left(\bigcup_{m' \leq \phi(m)} B_{(x_{m,m'},2^{-m})}\right) \geq \mu_n\left(\bigcup_{m' \leq \phi(m)} B_{(x_{m,m'},2^{-m-1})}\right) - \frac{1}{2^{m+1}} \geq 1 - \frac{1}{2^m},$$

because $d_{\mathrm{PL}}(\mu_n, \mu_N) \leq \frac{1}{2^{m+1}}$; and the same inequality holds for $n \leq N$. The tightness is then proved. □

The completeness of $\mathscr{M}^1(\mathfrak{X})$ is an important result: indeed, in (functional) analysis, one always looks for complete spaces so that there are "sufficiently many limits" — this is why standard analysis is done on $\mathbb{R}$ and not on $\mathbb{Q}$, and integration theory is done with the Lebesgue spaces $\mathscr{L}^1(\Omega)$ obtained by completion of the spaces of step functions. Now, the notion of tightness also eases many proofs of convergence in law. Indeed, assuming that a limit μ has already been identified for a sequence of probability measures $(\mu_n)_{n \in \mathbb{N}}$, one does not need anymore to prove that for every $f \in \mathscr{C}_b(\mathfrak{X})$, $\mu_n(f) \to \mu(f)$; the tightness is sufficient.

Example. Following [**Kallenberg, 2001**, Chapter 5], let us give a simple proof of Lévy's continuity theorem 1.14. If $\mu_n \rightharpoonup \mu$, then obviously $\Phi_{\mu_n} \to \Phi_\mu$ pointwise since each function $x \mapsto \mathrm{e}^{\mathrm{i}\zeta x}$ is continuous and bounded. Conversely, let us suppose the pointwise convergence of the characteristic functions Φ_{μ_n} to Φ_μ. By Fubini's theorem, we get the following estimate: for any $c > 0$,

$$\begin{aligned}\int_{-c}^{c} (1 - \Phi_\mu(\zeta))\, d\zeta &= \int_{\mathbb{R}} \mu(dx) \int_{-c}^{c} (1 - \mathrm{e}^{\mathrm{i}\zeta x})\, d\zeta = 2c \int_{\mathbb{R}} \left(1 - \frac{\sin cx}{cx}\right) \mu(dx) \\ &\geq c\, \mu(\{x \mid |cx| \geq 2\})\end{aligned}$$

since $\sin x \leq \frac{x}{2}$ for $x \geq 2$. As a consequence,

$$\mu(\{x \mid |x| \geq R\}) \leq \frac{R}{2} \int_{-\frac{2}{R}}^{\frac{2}{R}} (1 - \Phi_\mu(\zeta))\, d\zeta,$$

i.e., the tail of μ at infinity is controlled by how close Φ_μ stays to 1 around 0. By Lebesgue's dominated convergence theorem, one has therefore

$$\limsup_{n\to\infty} \mu_n(\{x \mid |x| \geq R\}) \leq \lim_{n\to\infty} \frac{R}{2} \int_{-\frac{2}{R}}^{\frac{2}{R}} (1 - \Phi_{\mu_n}(\zeta))\, d\zeta = \frac{R}{2} \int_{-\frac{2}{R}}^{\frac{2}{R}} (1 - \Phi_\mu(\zeta))\, d\zeta$$

which ensures the tightness since the right-hand side goes to zero as R goes to infinity (by continuity of Φ_μ at 0). Since the map $\mu \mapsto \Phi_\mu$ is injective, the limit of a convergent subsequence of $(\mu_n)_{n\in\mathbb{N}}$ is necessarily μ, so Theorem 1.14 is proved.

2.2. Compactness in $\mathscr{C}(X)$ and Donsker's theorem

In $\mathbb{R}^d$, relative compactness is equivalent to boundedness, so the tightness of a family of probability measures $\mathscr{P}$ is equivalent to the existence of balls such that

$$\mu(B^{\mathbb{R}^d}_{(0,R)}) \geq 1 - \varepsilon$$

for every $\mu \in \mathscr{P}$, and for an arbitrary $\varepsilon > 0$. However, in a functional (polish, or even Banach) space, the relative compactness is less easy to characterize, and so is the tightness of a family of probability measures $\mathscr{P} \subset \mathscr{M}^1(\mathfrak{X})$. In this section, we detail the case of $\mathscr{C}(X)$, the space of continuous functions on a compact, or even separable locally compact metric space (X, d); this example is specially important since it is the natural space of states of many random processes. We give criterions for relative compactness in $\mathfrak{X} = \mathscr{C}(X)$, and then for tightness in $\mathscr{M}^1(\mathfrak{X})$; and as an application we prove Donsker's theorem, which under weak assumptions ensures the convergence of the rescaled sums of independent and identically distributed increments towards an universal continuous random process, the ***Brownian motion***.

2.2.1. Arzelà-Ascoli criterion. To begin with, we assume that X is a compact metric space, with distance d. Then every continuous function $f \in \mathscr{C}(X)$ is in fact uniformly continuous (Heine's theorem), that is to say that for every $\varepsilon > 0$ there exists a modulus δ of uniform continuity such that

$$d(x, y) \leq \delta \Rightarrow |f(x) - f(y)| \leq \varepsilon.$$

We define $\omega_f(\delta)$ to be the maximum of $|f(x) - f(y)|$ when $d(x, y) \leq \delta$; the above statement is equivalent to $\lim_{\delta\to 0} \omega_f(\delta) = 0$. A family of functions $\mathscr{F} \subset \mathscr{C}(X)$ is said ***equicontinuous*** if

$$\lim_{\delta\to 0} \left(\omega_{\mathscr{F}}(\delta) = \sup_{f\in\mathscr{F}} \omega_f(\delta) \right) = 0.$$

In other words, for every $\varepsilon > 0$, there is a common ε-modulus of uniform continuity δ for all $f \in \mathscr{F}$.

We endow $\mathscr{C}(X)$, the space of continuous real-valued functions on X, with the sup norm $\|f\|_\infty = \max_{x\in X} |f(x)|$; then $\mathscr{C}(X)$ is a complete normed vector space (whence polish).

THEOREM 2.5 (Arzelà-Ascoli). *A family of functions $\mathscr{F} \subset \mathscr{C}(X)$ on a compact space is relatively compact if and only if it is equicontinuous and bounded, where by bounded we mean that*

$$\sup_{f\in\mathscr{F}} \|f\|_\infty = \sup_{f\in\mathscr{F}} \sup_{x\in X} |f(x)| = M < \infty.$$

PROOF. Consider a compact family $\mathscr{F} \subset \mathscr{C}(X)$. Since the map $\|\cdot\|_\infty : f \mapsto \|f\|_\infty$ is continuous, the image of $\mathscr{F}$ by $\|\cdot\|_\infty$ is compact in $\mathbb{R}$, whence bounded; so $\mathscr{F}$ is bounded in $\mathscr{C}(X)$ by some constant M. If $\mathscr{F}$ were not equicontinuous, then one could find a sequence of functions $(f_n)_{n\in\mathbb{N}}$ and points x_n, y_n in X such that $d(x_n, y_n) \to 0$ and $|f_n(x_n) - f_n(y_n)| \geq \varepsilon > 0$ for every n. Up to extractions in $[-M, M]$ and in $\mathscr{F}$, one can then suppose $x_n \to x$, $y_n \to y = x$ and $f_n \to f$. But then, since f is continuous and $\|f - f_n\|_\infty \to 0$,

$$\begin{aligned}
\varepsilon &\leq \liminf_{n\to\infty} |f_n(x_n) - f_n(y_n)| \\
&\leq \liminf_{n\to\infty} (|f_n(x_n) - f(x_n)| + |f(x_n) - f(y_n)| + |f_n(y_n) - f(y_n)|) \\
&\leq \liminf_{n\to\infty} (|f(x_n) - f(y_n)| + 2\|f - f_n\|_\infty) = 0;
\end{aligned}$$

whence a contradiction. So, a (relatively) compact family $\mathscr{F} \subset \mathscr{C}(X)$ is bounded and equicontinuous.

Conversely, suppose that $\mathscr{F} \subset \mathscr{C}(X)$ is bounded and equicontinuous. We fix a dense sequence $(a_m)_{m\in\mathbb{N}}$ in X, and a sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathscr{F}$. For every $m \in \mathbb{N}$, the sequence $(f_n(a_m))_{n\in\mathbb{N}}$ is in a compact interval $[-M, M]$, so it has a convergent subsequence. By diagonal extraction we can then make all the $(f_{\Psi(n)}(a_m))_{n\in\mathbb{N}}$ converge simultaneously. More precisely, we choose an extraction $\phi_0 : \mathbb{N} \to \mathbb{N}$ such that $(f_{\phi_0(n)}(a_0))_{n\in\mathbb{N}}$ converges; then, a further extraction $\phi_1 : \mathbb{N} \to \mathbb{N}$ such that $(f_{\phi_0\circ\phi_1(n)}(a_1))_{n\in\mathbb{N}}$ converges; *etc.* The extraction $\phi_k : \mathbb{N} \to \mathbb{N}$ is chosen so that $(f_{\phi_0\circ\phi_1\circ\cdots\circ\phi_k(n)}(a_k))_{n\in\mathbb{N}}$ converges. If

$$\Psi(n) = \phi_0 \circ \phi_1 \circ \cdots \circ \phi_n(n)$$

then after rank m, $(\Psi(n))_{n\in\mathbb{N}}$ is extracted from $\phi_0 \circ \phi_1 \circ \cdots \circ \phi_m$, so $f_{\Psi(n)}(a_m)$ has a limit; and this for every $m \in \mathbb{N}$. Now take $x \in X$, and $(b_m = a_{\phi(m)})_{m\in\mathbb{N}}$ a sequence that converges to x. For every k, l,

$$\begin{aligned}
&|f_{\Psi(k)}(x) - f_{\Psi(l)}(x)| \\
&\leq |f_{\Psi(k)}(x) - f_{\Psi(k)}(b_m)| + |f_{\Psi(k)}(b_m) - f_{\Psi(l)}(b_m)| + |f_{\Psi(l)}(x) - f_{\Psi(l)}(b_m)| \\
&\leq |f_{\Psi(k)}(b_m) - f_{\Psi(l)}(b_m)| + 2\,\omega_{\mathscr{F}}(d(x, b_m))
\end{aligned}$$

which proves that $(f_{\Psi(n)}(x))_{n\in\mathbb{N}}$ is Cauchy for every $x \in X$ (not only the x's in the fixed dense sequence). Therefore, $f(x) = \lim_{n\to\infty} f_{\Psi(n)}(x)$ exists for every $x \in X$. Finally, if $d(x, y) \leq \delta$, then

$$|f(x) - f(y)| = \lim_{n\to\infty} |f_{\Psi_n}(x) - f_{\Psi_n}(y)| \leq \omega_{\mathscr{F}}(\delta),$$

so $f \in \mathscr{C}(X)$ and $\mathscr{F}$ is relatively compact. □

2.2.2. Donsker's theorem. A combination of Prohorov's theorem 2.3 and of Arzelà-Ascoli criterion 2.5 leads to:

COROLLARY 2.6. *A family of probability measures $\mathscr{P} \subset \mathscr{M}^1(\mathscr{C}(X))$ is tight if and only if*

(i) $\lim_{M\to\infty} \sup_{\mu\in\mathscr{P}} \mu(\{\|f\|_\infty \geq M\}) = 0$*;*

(ii) and $\lim_{\delta\to 0} \sup_{\mu\in\mathscr{P}} \mu(\{\omega_f(\delta) \geq \varepsilon\}) = 0$ *for every* $\varepsilon > 0$.

PROOF. We denote $B_R^{\|\cdot\|_\infty}$ the ball of radius R and center 0 w.r.t. the sup norm, and $B_\varepsilon^{\omega(\delta)}$ the set of continuous functions f such that $\omega_f(\delta) < \varepsilon$. Suppose $\mathscr{P}$ tight, and fix $\varepsilon > 0$. There is a compact subset $K \subset \mathscr{C}(X)$ such that $\mu(K) \geq 1-\varepsilon$ for any $\mu \in \mathscr{P}$, and by Arzelà-Ascoli, an $M \geq 0$ such that

$$K \subset B_M^{\|\cdot\|_\infty} \cap \left(\bigcap_{\varepsilon>0} \bigcup_{\delta>0} B_\varepsilon^{\omega(\delta)} \right).$$

From this one deduces:

(i) If $m \geq M$, then for all $\mu \in \mathscr{P}$,

$$\mu(\{\|f\|_\infty \geq m\}) \leq \mu(\{\|f\|_\infty \geq M\}) \leq \mu(K^{\mathrm{c}}) \leq \varepsilon,$$

which proves that $\lim_{M\to\infty} \sup_{\mu\in\mathscr{P}} \mu(\{\|f\|_\infty \geq M\}) = 0$.

(ii) Fix $\varepsilon' > 0$ and a δ_0 such that $K \subset B_{\varepsilon'}^{\omega(\delta_0)}$. If $\delta \leq \delta_0$, then for all $\mu \in \mathscr{P}$,

$$\mu(\{\omega_f(\delta) \geq \varepsilon\}) \leq \mu(\{\omega_f(\delta_0) \geq \varepsilon'\}) \leq \mu(K^{\mathrm{c}}) \leq \varepsilon,$$

so $\lim_{\delta\to 0} \sup_{\mu\in\mathscr{P}} \mu(\{\omega_f(\delta) \geq \varepsilon'\}) = 0$ for every $\varepsilon' > 0$.

The direct sense is therefore proved. Conversely, if the two assumptions are satisfied, fix $\varepsilon > 0$ and choose real numbers M and $\delta_{n\geq 1}$ so that

$$\forall m \geq M,\ \ \forall \mu \in \mathscr{P},\ \ \mu(\{\|f\|_\infty \geq m\}) \leq \varepsilon;$$
$$\forall \delta \leq \delta_n,\ \ \forall \mu \in \mathscr{P},\ \ \mu\left(\left\{\omega_f(\delta) \geq 2^{-n}\right\}\right) \leq \frac{\varepsilon}{2^n}.$$

The μ-measure of the union $\{\|f\|_\infty \geq M\} \cup \bigcup_{n\geq 1} \{\omega_f(\delta_n) \geq 2^{-n}\}$ is smaller than 2ε for every $\mu \in \mathscr{P}$. The complementary of this set is

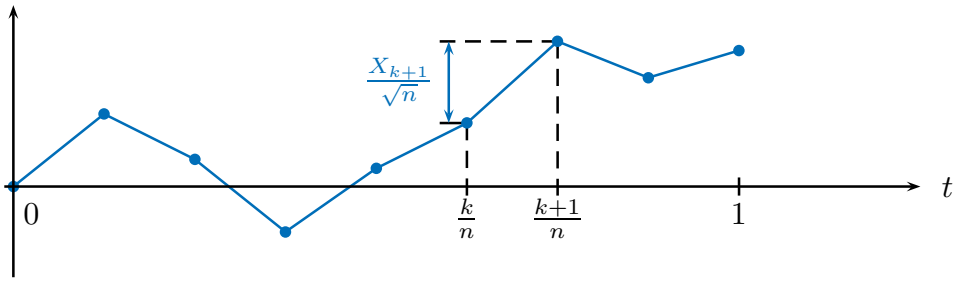
$$B_M^{\|\cdot\|_\infty} \cap \left(\bigcap_{n\geq 1} B_{2^{-n}}^{\omega(\delta_n)} \right) \subset B_M^{\|\cdot\|_\infty} \cap \left(\bigcap_{n\geq 1} \bigcup_{\delta>0} B_{2^{-n}}^{\omega(\delta)} \right) = B_M^{\|\cdot\|_\infty} \cap \left(\bigcap_{\varepsilon>0} \bigcup_{\delta>0} B_\varepsilon^{\omega(\delta)} \right)$$

and it is relatively compact. □

EXAMPLE. Fix a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables that are independent, identically distributed and in $\mathscr{L}^2(\Omega,\mathcal{B},\mathbb{P})$, with mean $\mathbb{E}[X]=0$ and variance $\mathbb{E}[X^2]=1$. We associate to such a sequence a family of random continuous functions on the segment $[0,1]$:

$$W^{(n)}(\omega,t)=\frac{1}{\sqrt{n}}\left(S_{\lfloor nt\rfloor}(\omega)+(nt-\lfloor nt\rfloor)\,X_{\lfloor nt\rfloor+1}(\omega)\right),$$

where $S_n=X_1+X_2+\cdots+X_n$. The map $W^{(n)}$ is the unique affine map starting from 0 and such that the increment from time $t=\frac{k}{n}$ to time $t=\frac{k+1}{n}$ is linear and equal to $\frac{X_{k+1}}{\sqrt{n}}$; see the figure hereafter.



For every n, $W^{(n)}$ can be considered as a $\mathscr{C}([0,1])$-valued random variable, since it is a continuous (hence measurable) function $\mathbb{R}^n\to\mathscr{C}([0,1])$ of the random vector $(X_1,\dots,X_n)$. Denote $\mu^{(n)}$ the law of $W^{(n)}$; we claim that $\{\mu^{(n)}\}_{n\in\mathbb{N}}$ is tight inside $\mathscr{M}^1(\mathscr{C}([0,1]))$. In order to prove this, notice that in Corollary 2.6, on any path connected compact space X, the first criterion (i) can be replaced by

$$\lim_{M\to\infty}\sup_{\mu\in\mathscr{P}}\mu(|f(x_0)|\geq M\})=0$$

where $x_0\in X$ is some fixed point. Indeed, if $\varepsilon>0$ is fixed and assuming (ii) true, one has then

$$\sup_{\mu\in\mathscr{P}}\mu(\{|f(x_0)|\geq M\})\leq\frac{\varepsilon}{2}$$

for M big enough, but also

$$\sup_{\mu\in\mathscr{P}}\mu(\{\omega_f(\delta)\geq M\})\leq\frac{\varepsilon}{2}$$

for δ small enough. It follows that the measure of the complementary of these sets is bigger than $1-\varepsilon$, and on this event,

$$|f(x)|\leq|f(x_0)|+\left\lceil\frac{d(x_0,x)}{\delta}\right\rceil M\leq\left(1+\left\lceil\frac{\mathrm{diam}(X)}{\delta}\right\rceil\right)M,$$

so $\|f\|_\infty$ is bounded by a constant with arbitrary big μ-probability uniformly in $\mu\in\mathscr{P}$. In our case, $W^{(n)}(0)=0$ almost surely, so it is sufficient to verify that (ii) in Corollary 2.6 is satisfied. This is done by using ***Kolmogorov's tightness criterion***, in the following form:

PROPOSITION 2.7 (Kolmogorov). *In the previous setting, if*

$$\lim_{\lambda\to\infty}\limsup_{n\to\infty}\lambda^2\,\mathbb{P}\left[\max_{k\in[\![1,n]\!]}|S_k|\geq\lambda\sqrt{n}\right]=0,$$

then one has tightness for $(W^{(n)})_{n\in\mathbb{N}}$.

PROOF. First, let us check that this criterion holds. This is a consequence of the subtle Etemadi maximal inequality, which is true for any sequence of independent centered random variables in $\mathscr{L}^2(\Omega)$. Set $\tau=\inf\{k\geq 1,\ \ |S_k|\geq 3M\}$ — the same kind of argument has been used in the proof of Theorem 1.2. Then,

$$\begin{aligned}
\mathbb{P}\left[\max_{k\in[\![1,n]\!]}|S_k|\geq 3M\right]&\leq\mathbb{P}[|S_n|\geq M]+\sum_{k=1}^{n-1}\mathbb{P}[(|S_n|<M)\wedge(\tau=k)]\\
&\leq\mathbb{P}[|S_n|\geq M]+\sum_{k=1}^{n-1}\mathbb{P}[(|S_n-S_k|>2M)\wedge(\tau=k)]\\
&\leq\mathbb{P}[|S_n|\geq M]+\sum_{k=1}^{n-1}\mathbb{P}[(|S_n-S_k|>2M)]\;\mathbb{P}[\tau=k]\\
&\leq\mathbb{P}[|S_n|\geq M]+\max_{k\in[\![1,n-1]\!]}\left(\mathbb{P}[|S_n-S_k|\geq 2M]\right)\\
&\leq\mathbb{P}[|S_n|\geq M]+\max_{k\in[\![1,n-1]\!]}\left(\mathbb{P}[|S_n|\geq M]+\mathbb{P}[|S_k|\geq M]\right)\\
&\leq 3\max_{k\in[\![1,n]\!]}\left(\mathbb{P}[|S_k|\geq M]\right).
\end{aligned}$$

Suppose now that $S_n=X_1+\cdots+X_n$ with i.i.d. random variables. By the central limit theorem, $\frac{S_n}{\sqrt{n}}$ converges to a Gaussian random variable G, so if λ is fixed, then for $k\geq k_\lambda$ big enough,

$$\mathbb{P}[|S_k|\geq\lambda\sqrt{n}]\leq\mathbb{P}[|S_k|\geq\lambda\sqrt{k}]\leq\frac{4}{3}\mathcal{N}([-\lambda,\lambda]^c)\leq\frac{4}{3}\frac{\mathbb{E}[G^4]}{\lambda^4}=\frac{4}{\lambda^4}.$$

On the other hand, for $k\leq k_\lambda$, by Chebyshev's inequality in $\mathscr{L}^2(\Omega)$,

$$\mathbb{P}[|S_k|\geq\lambda\sqrt{n}]\leq\frac{\mathbb{E}[(S_k)^2]}{\lambda^2 n}=\frac{k}{\lambda^2 n}\leq\frac{k_\lambda}{\lambda^2 n}.$$

Consequently,

$$\limsup_{n\to\infty}(3\lambda)^2\,\mathbb{P}\left[\max_{k\in[\![1,n]\!]}|S_k|\geq 3\lambda\sqrt{n}\right]\leq 27\limsup_{n\to\infty}\max\left(\frac{k_\lambda}{n},\frac{4}{\lambda^2}\right)=\frac{108}{\lambda^2}\to_{\lambda\to\infty}0$$

and Kolmogorov's criterion is satisfied.

Now, fix $\varepsilon>0$, and let us prove that under Kolmogorov's criterion one has

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mu_n(\{\omega_W(\delta)\geq\varepsilon\})=0;$$

since we are working with a sequence we can replace the supremum over a family by the limit of the supremum in Corollary 2.6. Fix a subdivision $0=t_0\leq t_1\leq\cdots\leq t_r=1$ of

$[0,1]$, and δ smaller than $|t_{i+1}-t_i|$ for every i. If $\omega_W(\delta)\geq 3\varepsilon$, then there exist times $s<t$ with $|t-s|\leq\delta$ and $|W(s)-W(t)|\geq 3\varepsilon$. Either s and t are in the same interval $[t_{i-1},t_i]$, in which case

$$2\max_{i\in[\![1,r]\!]}\sup_{u\in[t_{i-1},t_i]}|W(u)-W(t_i)|\geq|W(s)-W(t_i)|+|W(t)-W(t_i)|\geq|W(s)-W(t)|\geq 3\varepsilon.$$

Otherwise, $s\in[t_{i-1},t_i]$ and $t\in[t_i,t_{i+1}]$ and

$$\begin{aligned}3\max_{i\in[\![1,r]\!]}\sup_{u\in[t_{i-1},t_i]}|W(u)-W(t_i)|&\geq|W(s)-W(t_i)|+|W(t_i)-W(t_{i+1})|+|W(t)-W(t_{i+1})|\\&\geq|W(s)-W(t)|\geq 3\varepsilon.\end{aligned}$$

So in every situation, $\max_{i\in[\![1,r]\!]}\sup_{u\in[t_{i-1},t_i]}|W(u)-W(t_i)|\geq\varepsilon$, and therefore

$$\mathbb{P}[\omega_W(\delta)\geq 3\varepsilon]\leq\sum_{i=1}^{r}\mathbb{P}\left[\sup_{u\in[t_{i-1},t_i]}|W(u)-W(t_i)|\geq\varepsilon\right].$$

Assume now $\mathbb{P}=\mu_n$ and $(t_i=\frac{T_i}{n})_{i\in[\![0,r]\!]}$ subdivision of $(\frac{i}{n})_{i\in[\![0,n]\!]}$. Then for each i, since the function $W^{(n)}$ is almost surely affine by parts with changes of direction at points $\frac{i}{n}$,

$$\sup_{u\in[t_{i-1},t_i]}|W(u)-W(t_i)|=\max_{k\in[\![T_{i-1},T_i]\!]}\left|W\left(\frac{k}{n}\right)-W\left(\frac{T_i}{n}\right)\right|,$$

and therefore,

$$\begin{aligned}\mu_n(\{\omega_W(\delta)\geq 3\varepsilon\})&\leq\sum_{i=1}^{r}\mathbb{P}\left[\max_{k\in[\![T_{i-1},T_i]\!]}|S_k-S_{T_i}|\geq\varepsilon\sqrt{n}\right]\\&\leq\sum_{i=1}^{r}\mathbb{P}\left[\max_{k\in[\![1,T_i-T_{i-1}]\!]}|S_k|\geq\varepsilon\sqrt{n}\right],\end{aligned}$$

assuming that $\delta\leq\frac{T_i-T_{i-1}}{n}$ for all i. Write then $\lambda=\frac{\varepsilon}{\sqrt{4\delta}}$, and $T_i=im$ where $m=\lceil n\delta\rceil$. This condition ensures that

$$\frac{T_i-T_{i-1}}{n}=\frac{\lceil n\delta\rceil}{n}\geq\delta,$$

except maybe around $t=1$: we remove the last T_i and choose $T_r=n$ in such a way that

$$\lceil n\delta\rceil\leq T_r-T_{r-1}<2\lceil n\delta\rceil.$$

The number r is then smaller than $\frac{n}{m}\leq\frac{1}{\delta}$, and on the other hand, for n big enough, $\frac{n}{m}\geq\frac{1}{2\delta}$. Then,

$$\begin{aligned}\mu_n(\{\omega_W(\delta)\geq 3\varepsilon\})&\leq r\,\mathbb{P}\left[\max_{k\in[\![1,2m]\!]}|S_k|\geq\varepsilon\sqrt{n}\right]\leq\frac{1}{\delta}\,\mathbb{P}\left[\max_{k\in[\![1,2m]\!]}|S_k|\geq\frac{\varepsilon}{\sqrt{4\delta}}\sqrt{2m}\right]\\&\leq\frac{4\lambda^2}{\varepsilon^2}\,\mathbb{P}\left[\max_{k\in[\![1,2m]\!]}|S_k|\geq\lambda\sqrt{2m}\right].\end{aligned}$$

The parameters ε and δ being fixed, for n going to infinity, m goes also to infinity, so by Kolmogorov's criterion the right-hand side goes to zero. □

In the previous example, it turns out that there is a unique possible limit for of a convergent subsequence of r.v. $(W^{(\phi(n))})_{n\in\mathbb{N}}$. The proof of this fact relies on the following general result:

PROPOSITION 2.8. *Let $\mathscr{P}$ be a tight subset of $\mathscr{M}^1(\mathscr{C}(X))$, where X is a compact metric space. If for every finite family of points of X,*

$$(\mu_n)_*(f(x_1), f(x_2), \dots, f(x_d)) \rightharpoonup_{\mathbb{R}^d} \mu_*(f(x_1), f(x_2), \dots, f(x_d))$$

with the μ_n's in $\mathscr{P}$, then $\mu_n \rightharpoonup_{\mathscr{C}(X)} \mu$. In other words, tightness and convergence of the finite-dimensional laws ensure the convergence — on the other hand, the finite-dimensional laws entirely determine a probability measure on $\mathscr{C}(X)$.

PROOF. Let us first see why the finite-dimensional laws determine the law of a random function $f \in \mathscr{C}(X)$ taken according to a probability measure μ. Fix $\varepsilon > 0$, $f \in \mathscr{C}(X)$ and a dense sequence $(a_m)_{m\in\mathbb{N}}$ in X. The closed ball $\overline{B}_{(f,\varepsilon)}$ can be written as

$$\begin{aligned}\overline{B}_{(f,\varepsilon)} &= \bigcup_{m=1}^{\infty} \{g \in \mathscr{C}(X),\ \ |g(a_m) - f(a_m)| \leq \varepsilon\} \\ &= \bigcup_{m=1}^{\infty}\uparrow \{g \in \mathscr{C}(X),\ \ \forall m' \leq m,\ \ |f(a_{m'}) - g(a_{m'})| \leq \varepsilon\}.\end{aligned}$$

The sets appearing on the right-hand side are finite-dimensional laws, so the measure of any closed ball is entirely determined by the finite-dimensional laws. From there one can compute the measure of any open ball, and then by separability of any open set. This determines μ by a standard argument of measure theory.

Now, suppose that $(\mu_n)_{n\in\mathbb{N}}$ is a tight sequence on $\mathscr{C}(X)$ and converges in finite-dimensional laws towards μ. Since the maps

$$\begin{aligned}\mathscr{C}(X) &\to \mathbb{R}^d \\ f &\mapsto (f(x_1), \dots, f(x_d))\end{aligned}$$

are continuous, by Proposition 1.11, if ν is the limit a convergent subsequence of $(\mu_n)_{n\in\mathbb{N}}$, then the finite dimensional laws of ν agree with those of ν, so $\mu = \nu$ by the previous discussion. The lemma at the beginning of the chapter allows one to conclude. □

All this leads to the powerful:

THEOREM 2.9 (Donsker). *Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of identically distributed and independent random variables in $\mathscr{L}^2(\Omega, \mathcal{B}, \mathbb{P})$, with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$. The sequence of random paths $(W^{(n)})_{n\geq 1}$ that is associated to it converges in law towards the unique continuous random process W in $\mathscr{C}([0,1])$ such that, for any times $0 = t_0 \leq t_1 \leq \cdots \leq t_r \leq 1$,*

$$W_{t_1}, W_{t_2} - W_{t_1}, \dots, W_{t_r} - W_{t_{r-1}}$$

are independent centered Gaussians of variance $t_1, t_2 - t_1, \dots, t_r - t_{r-1}$. In particular, such a random continuous path exists, and it is called the Brownian motion.

PROOF. Since the tightness is already shown, it suffices to verify that the finite dimensional laws are those indicated above. Fix times $t_1, \dots, t_r$ and intervals $[a_1, b_1], \dots, [a_r, b_r]$; one wants to estimate the probability

$$\mathbb{P}\left[(W^{(n)}_{t_1}, W^{(n)}_{t_2} - W^{(n)}_{t_1}, \dots, W^{(n)}_{t_r} - W^{(n)}_{t_{r-1}}) \in \prod_{i=1}^{r} [a_i, b_i]\right].$$

Each increment $W^{(n)}_{t_i} - W^{(n)}_{t_{i-1}}$ differs from $W^{(n)}_{\frac{\lfloor t_i n \rfloor}{n}} - W^{(n)}_{\frac{\lfloor t_{i-1} n \rfloor}{n}}$ by at most (in absolute value)

$$\Delta^{(n)}_i = \frac{1}{\sqrt{n}} \left(|X_{t_i+1}| + |X_{t_{i-1}+1}| \right).$$

For n big enough these quantities $\Delta^{(n)}_i$ are independent and identically distributed random variables with square mean smaller than $\frac{4}{n}$. It follows that if $\varepsilon > 0$ is fixed, then the probability that all $\Delta^{(n)}_i$ are smaller than ε can be made arbitrary close to 1, so that the asymptotic behavior of the probability to estimate is the same as the one of

$$\mathbb{P}\left[\left(\frac{S_{\lfloor t_1 n \rfloor}}{\sqrt{n}}, \frac{S_{\lfloor t_2 n \rfloor} - S_{\lfloor t_1 n \rfloor}}{\sqrt{n}}, \dots, \frac{S_{\lfloor t_r n \rfloor} - S_{\lfloor t_{r-1} n \rfloor}}{\sqrt{n}}\right) \in \prod_{i=1}^{r} [a_i, b_i]\right].$$

The central limit theorem ensures that this random vector converges in law to a Gaussian vector with independent entries and diagonal variances $t_1, t_2 - t_1, \dots, t_r - t_{r-1}$ (use if needed Lévy's continuity theorem to reprove in a vectorial setting the convergence in law by computation of the asymptotics of the characteristic functions). So, the limit of the probability is as expected

$$\frac{1}{\prod_{i=1}^{r} \sqrt{2\pi(t_i - t_{i-1})}} \int_{\prod_{i=1}^{r} [a_i, b_i]} \mathrm{e}^{-\sum_{i=1}^{r} \frac{(x_i)^2}{2(t_i - t_{i-1})}} \, dx^r$$

and this ends the proof of our theorem □

2.3. Skorohod's space $\mathscr{D}([0,1])$

The proof of Donsker's theorem is a typical use of tightness theory in a functional setting, and it illustrates every important feature of this kind of argument. Hence, one can show in a very similar way the existence of other complicate random objects, such as: solutions of stochastic differential equations, random Brownian trees, random metric spaces that are scaling limits of random combinatorial objects (*e.g.* random maps), *etc.* To conclude this chapter, let us discuss two generalizations of Donsker's theorem.

2.3.1. Extension of Donsker's theorem to locally compact separable spaces. The first one does not cost much and is the extension of the abstract setting developed at the beginning of §2.2 to spaces X that are only locally compact and separable. Locally compact means that every point $x \in X$ has a compact neighborhood, and if one assumes also the separability, then X can be written as an increasing union of compact sets:

$$X = \bigcup_{n=1}^{\infty} \uparrow K_n, \quad \text{with each } K_n \text{ compact set.}$$

A typical example is $\mathbb{R} = \bigcup_{n=1}^{\infty}[-n,n]$. The space $\mathscr{C}(X)$ of continuous functions on a locally compact separable metric space (in short l.c.s.s.) is endowed with the topology of convergence on every compact set, that is to say that a basis of neighborhoods of a function $f \in \mathscr{C}(X)$ consists in the sets

$$\{g \in \mathscr{C}(X),\ \ \forall x \in K,\ \ |f(x) - g(x)| \leq \varepsilon\},$$

where K is a compact set and $\varepsilon > 0$. For a l.c.s.s., this topology is metrizable by

$$d(f,g) = \sum_{n=1}^{\infty} \frac{\|f-g\|_{\infty,K_n}}{2^n} \quad \text{with } \|u\|_{\infty,K} = \sup_{x\in K} |u(x)| \text{ for any compact set } K.$$

We thus get a polish space $\mathscr{C}(X)$ on which Prohorov's theorem can be applied. The analogue of Arzelà-Ascoli theorem 2.5 in the setting of l.c.s.s. is a localized version of equicontinuity: a family $\mathscr{F} \subset \mathscr{C}(X)$ is relatively compact if and only if on any compact set K (or any compact set in an exhaustive sequence $(K_n)_{n\in\mathbb{N}}$ with $X = \bigcap_{n=1}^{\infty} K_n$),

$$\lim_{\delta\to 0} \sup_{f\in\mathscr{F}} \omega_{f,K}(\delta) = 0 \quad \text{with } \omega_{f,K}(\delta) = \sup\{|f(x) - f(y)|,\ x,y \in K \text{ and } d(x,y) \leq \delta\}.$$

As a consequence, a family of probability measures $\mathscr{P} \subset \mathscr{M}^1(\mathscr{C}(X))$ is tight if and only if, for every compact set K and any $\varepsilon > 0$,

$$\lim_{M\to\infty} \sup_{\mu\in\mathscr{P}} \mu(\{\|f\|_{\infty,K} \geq M\}) = 0 \quad ; \quad \lim_{\delta\to 0} \sup_{\mu\in\mathscr{P}} \mu(\{\omega_{f,K}(\delta) \geq \varepsilon\}) = 0.$$

In short: if all the previously discussed criterions hold on any compact subset $K \subset X$, then everything works. So for instance, one gets for free the existence of a continuous random path W on $\mathbb{R}_+ = [0,+\infty)$ (instead of $[0,1]$) that is the universal limit of rescaled sums of independent identically distributed random variables, and such that increments are independent Gaussian variables as in Donsker's theorem.

2.3.2. Extension of Donsker's theorem to càdlàg paths. A far more complex generalization of the previous discussion is the analysis of random paths that may have jumps, and of tight sequences of such random paths. We shall not give all the details, but try to explain the main features of this theory, following [**Billingsley, 1999**, Chapter 3]. Denote $\mathscr{D}([0,1]) \supset \mathscr{C}([0,1])$ the space of ***càdlàg functions*** $f : [0,1] \to \mathbb{R}$ such that:

(i) for every $t \in [0,1]$, $\lim_{s\to t_-} f(s)$ exists;

(ii) for every $t \in [0,1]$, $\lim_{s\to t_+} f(s)$ exists and is equal to $f(t)$.

Hence, f is continuous on the right and with limits on the left. Important functions that are in $\mathscr{D}([0,1])$ but not necessarily in $\mathscr{C}([0,1])$ are the functions with bounded variation (up to modification at a countable number of points of discontinuity so that jumps are *càdlàg*), and in particular, any monotone function. As a consequence, the space $\mathscr{D}([0,1])$, or its infinite counterpart $\mathscr{D}(\mathbb{R}_+)$, is particularly useful in the study of random processes that may have jumps, *e.g.*, Poisson processes, or more generally a process in one of the two most studied classes, namely, Feller processes and semimartingales — up to modification, a real-valued process in one these two classes can always be realized as an element of $\mathscr{D}(\mathbb{R}_+)$, see [**Revuz and Yor, 2004**].

The topology on $\mathscr{D}([0,1])$ corresponds basically to the "uniform" convergence of the graphs of the functions viewed as subsets of $\mathbb{R}^2$. To make this notion rigorous, let us introduce the metric

$$d(f,g) = \inf_{\psi} \max(\|f - g\circ\psi\|_\infty, \|\mathrm{id} - \psi\|_\infty),$$

where the infimum is taken over all the increasing homeomorphisms $\psi : [0,1] \to [0,1]$. This is indeed a distance, and the corresponding topology is called Skorohod's topology; its restriction to $\mathscr{C}([0,1])$ can be shown to be the same as the topology given by the uniform norm $\|\cdot\|_\infty$ — in particular, $\mathscr{C}([0,1])$, since complete, is closed inside $\mathscr{D}([0,1])$. It can be shown that $\mathscr{D}([0,1])$ is a polish space, but new difficulties occur in the proof:

(1) First of all, $\mathscr{D}([0,1])$, though a vector space endowed with a topology, is not a topological vector space, that is, addition of functions is not continuous w.r.t. Skorohod's topology! Consider indeed the two following sequences of functions:

$$f_n = \mathbf{1}_{\left[\frac{1}{2}+\frac{1}{2^{n+1}},1\right]} \quad ; \quad g_n = \mathbf{1}_{\left[\frac{1}{2}-\frac{1}{2^{n+1}},1\right]}.$$

Skorohod's topology is made for such sequences to converge to $h = \mathbf{1}_{[\frac{1}{2},1]}$. Indeed, define ψ_n to be the affine by parts function such that $\psi_n(0) = 0$, $\psi_n(\frac{1}{2}) = \frac{1}{2} + \frac{1}{2^{n+1}}$ and $\psi_n(1) = 1$, and ψ_n is affine between these points. Then $h \circ \psi_n = f_n$, so

$$d(f_n, h) \leq \|\psi_n - \mathrm{id}\|_\infty = \frac{1}{2^{n+1}} \to 0,$$

and the same argument holds for the g_n's. If addition were continuous, then one would have $f_n + g_n \to 2h$ in $\mathscr{D}([0,1])$. However, for every homeomorphism ψ, $h \circ \psi$ has only two different values, 0 and 2, whereas $f_n + g_n$ also takes the value 1; it follows that

$$d(f_n + g_n, 2h) \geq 1 \quad \text{for any } n \geq 1.$$

(2) This oddity put apart, the distance d defined above is not complete. However, one can find a topologically equivalent distance δ that is complete. The idea is to penalize more the homeomorphisms ψ with large fluctuations, hence, one takes

$$\delta(f,g) = \inf_{\psi} \max\left(\|f - g\circ\psi\|_\infty, \sup_{s<t} \log\left|\frac{\psi(t)-\psi(s)}{t-s}\right|\right).$$

This is only useful to show that $\mathscr{D}([0,1])$ is a polish space, since of course δ is then way less manipulable than d.

The analogue of Arzelà-Ascoli therem for Skorohod's space involves a notion of δ-modulus adapted to functions which may have jumps. Thus, define

$$\omega'_f(\delta) = \inf_{0<t_1<t_2<\cdots<t_r=1} \max_{i\in[\![1,r]\!]} \sup_{x\in[t_{i-1},t_i)} |f(x) - f(t_{i-1})|$$

where the infimum is taken over subdivisions of $[0,1]$ with $|t_i - t_{i-1}| \geq \delta$ for all i. One can show that

$$\omega'_f\left(\frac{\delta}{2}\right) \leq \omega_f(\delta) \leq 2\,\omega'_f(\delta) + \sup_{x\in[0,1]} |f(x) - f(x_-)|$$

for any $f \in \mathscr{D}([0,1])$. Then, a part $\mathscr{F} \subset \mathscr{D}([0,1])$ is relatively compact if and only if it is bounded and $\lim_{\delta\to 0} \sup_{f\in\mathscr{F}} \omega'_f(\delta) = 0$ — remark the similarity with Theorem 2.5. It follows that tightness in $\mathscr{M}^1(\mathscr{D}([0,1]))$ is given by the exact analogue of Corollary 2.6, with $\omega'_f(\delta)$ instead of $\omega_f(\delta)$. Proposition 2.8 also generalizes to the case of functions in $\mathscr{D}([0,1])$. This enables one to deal with more general random paths than continuous paths, but the price is the complexity of Skorohod's topology.

EXAMPLE. Given a sequence $(X_n)_{n\in\mathbb{N}}$ of i.i.d. random variables in $\mathscr{L}^2(\Omega)$, we associate to it a sequence of random paths in $\mathscr{D}([0,1])$:

$$X^{(n)}(t) = \frac{S_{\lfloor nt\rfloor}}{\sqrt{n}}, \quad \text{with } S_k = X_1 + X_2 + \cdots + X_k.$$

This definition is quite simpler than the definition of the affine by parts paths $W^{(n)}$, and one ends up in Skorohod's space $\mathscr{D}([0,1])$. Then, one can show that in $\mathscr{D}([0,1])$, one has convergence in law of $(X^{(n)})_{n\in\mathbb{N}}$ towards the Brownian motion, though it is not obvious then that this random process is in fact continuous. So, there is the analog of Theorem 2.9 in Skohorod's space.

CHAPTER 3

Cramér's and Sanov's theorems

In the previous chapters, we have reviewed the different notions of convergence in probability, and we have given criterions of compactness for sequences of random variables. Another topic in this setting is the study of the speed of convergence, and it leads to the theory of large deviations. More precisely, a way to measure the convergence (say, in probability) of a sequence of random variables $X_n \to_{\mathbb{P}} 0$ is to compute the probabilities for the $|X_n|$'s to stay bigger than some level ε:

$$\mathbb{P}[|X_n| \geq \varepsilon] = ?$$

These probabilities $P(\varepsilon)$ all go to zero, and the speed of convergence of the r.v. is related to their rate of decay. In the case of sums of i.i.d. random variables, the computation of the rate of decay is given by Cramér's theorem, *cf.* §3.1; whereas for empirical laws of finite Markov chains, the convergence is ensured by the so-called ergodic theorems (§3.2) and is measured by Sanov's entropy (see §3.3). Cramér's and Sanov's theorems shall allow us to introduce the main features of the general theory of large deviations, which will in turn be exposed in Chapter 4. Our main references for this chapter and the next one are [**Deuschel and Stroock, 1989**, **Dembo and Zeitouni, 1998**, **Feng and Kurtz, 2006**].

3.1. Legendre-Fenchel transforms and Cramér's large deviations

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables in $\mathbb{R}$. Even when the X_n's do not have the same distribution, one still expects $Z_n = \frac{1}{n}\sum_{i=1}^n X_i$ to be concentrated around its mean $\mathbb{E}[Z_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i]$, at least if the X_i's are "not too big". To make this statement precise, suppose for instance that each X_i is almost surely bounded, with values in an interval $[a_i, b_i]$.

PROPOSITION 3.1 (Hoeffding). *In the previous setting,*

$$\mathbb{P}[|Z_n - \mathbb{E}[Z_n]| \geq \varepsilon] \leq 2\,\exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

PROOF. This inequality follows from a clever use of ***Chernov's inequality*** and the convexity of the exponential map. For any random variable X with mean 0,

$$\mathbb{E}[\mathrm{e}^{\theta X}] \leq \mathbb{E}\left[\frac{X-a}{b-a}\,\mathrm{e}^{\theta b} + \frac{b-X}{b-a}\,\mathrm{e}^{\theta a}\right] = \frac{b\mathrm{e}^{\theta a} - a\mathrm{e}^{\theta b}}{b-a} = \mathrm{e}^{\theta a + \log(\frac{b}{b-a} - \frac{a}{b-a}\,\mathrm{e}^{\theta(b-a)})}$$

for any interval $[a,b]$ containing the values of X. Notice that the function

$$g(\theta) = \theta a + \log\left(\frac{b}{b-a} - \frac{a}{b-a}\,\mathrm{e}^{\theta(b-a)}\right)$$

satisfies $g(0)=0$ and $g'(0)=0$, so, by Taylor expansion,

$$g(\theta) \le \frac{\theta^2}{2}\left(\max_{0\le\phi\le\theta} g''(\phi)\right) = \frac{\theta^2}{2}\,\frac{(b-a)^2}{4} \qquad ; \qquad \mathbb{E}[\mathrm{e}^{\theta X}] \le \mathrm{e}^{\frac{\theta^2(b-a)^2}{8}}.$$

Then, by Chernov's inequality,

$$\begin{aligned}
\mathbb{P}[Z_n - \mathbb{E}[Z_n] \ge \varepsilon] &\le \mathrm{e}^{-\theta\varepsilon}\ \mathbb{E}\Big[\mathrm{e}^{\theta(Z_n-\mathbb{E}[Z_n])}\Big] \\
&\le \mathrm{e}^{-\theta\varepsilon}\ \prod_{i=1}^{n} \mathbb{E}\Big[\mathrm{e}^{\frac{\theta}{n}(X_i-\mathbb{E}[X_i])}\Big] \\
&\le \mathrm{e}^{-\theta\varepsilon+\frac{\theta^2}{8n^2}\sum_{i=1}^n (b_i-a_i)^2} \\
&\le \mathrm{e}^{-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i-a_i)^2}}
\end{aligned}$$

if one takes for the last step $\theta = 4\varepsilon n^2/(\sum_{i=1}^n (b_i-a_i)^2)$. Replacing Z_n by $-Z_n$, one gets the inequality with a coefficient 2 for $\mathbb{P}[|Z_n - \mathbb{E}[Z_n]| \ge \varepsilon]$. □

As a corollary, if the X_n's all take their values in a common interval $[a,b]$, then

$$\mathbb{P}[|Z_n - \mathbb{E}[Z_n]| \ge \varepsilon] \le 2\,\exp\left(-\frac{2\varepsilon^2 n}{(b-a)^2}\right),$$

that is to say that the probability of deviation of the mean from its expected value decreases exponentially fast in n. The theory of large deviations is exactly meant to prove such kind of results, and also to compute precisely the exponents of fluctuations. In the previous example, we have shown that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[|Z_n - \mathbb{E}[Z_n]| \ge \varepsilon] \le -\frac{2\varepsilon^2}{(b-a)^2},$$

but this inequality is quite suboptimal in some cases. Cramér's theorem 3.3 will give a refined and precise estimate in the case of independent and identically distributed random variables.

3.1.1. Legendre-Fenchel transforms. Thus, consider a sequence of i.i.d. random variables $X_1, X_2, \dots$ with values in $\mathbb{R}$, not necessarily in any space $\mathscr{L}^p(\Omega)$. The ***Laplace transform***

$$\Phi_X(s) = \mathbb{E}[\mathrm{e}^{sX}]$$

is well defined as a function $\mathbb{R} \to [0,+\infty]$, since $\mathrm{e}^{sX(\omega)} \ge 0$ for any s and any ω in the probability space $(\Omega, \mathcal{B}, \mathbb{P})$. The independence hypothesis ensures that

$$\Phi_{Z_n}(s) = \left(\Phi_X\left(\frac{s}{n}\right)\right)^n \quad \forall s \in \mathbb{R},\ n \in \mathbb{N},$$

and then, by Chernov's inequality, one has

$$\begin{aligned}\mathbb{P}[Z_n \geq u] &\leq \mathrm{e}^{-su}\,\mathbb{E}[\mathrm{e}^{sZ_n}] = \mathrm{e}^{-su}\,\Lambda_{Z_n}(s) = \exp\left(-n\left(\frac{su}{n} - \log\Phi_X\left(\frac{s}{n}\right)\right)\right)\\ &\leq \exp\left(-n(tu - \log\Phi_X(t))\right)\end{aligned}$$

with $t = \frac{s}{n}$. Since this is true for any value of $t \geq 0$, one deduces that

$$\frac{1}{n}\log\mathbb{P}[Z_n \geq u] \leq -\sup_{t\in\mathbb{R}_+}(tu - \Lambda_X(t))$$

where $\Lambda_X(t) = \log\Phi_X(t)$ is the logarithm of the moment generating function, also known as the cumulant generating function.

This leads us to consider the following transform on (convex) functions.

DEFINITION 3.2. *The* ***Legendre-Fenchel transform*** *of a function $f : \mathbb{R} \to (-\infty, +\infty]$ not equal everywhere to $+\infty$ is*

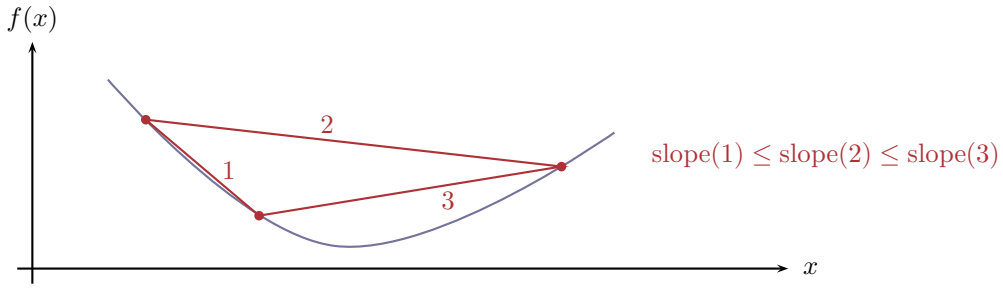
$$f^*(u) = \sup_{t\in\mathbb{R}}(tu - f(t)) \in (-\infty, +\infty].$$

The transformation LF : $f \mapsto f^*$ *takes its values in the set of convex and lower semi-continuous functions, and $f = f^{**}$ if and only if f is convex and lower semi-continuous.*

PROOF. Recall that a function on the real line is said ***convex*** if, for $a \leq x \leq b$,

$$f(x) \leq \frac{b-x}{b-a}\,f(a) + \frac{x-a}{b-a}\,f(b);$$

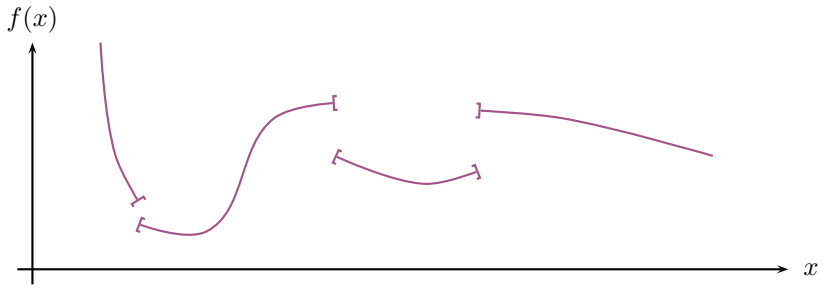
equivalently, the slopes of the function are non-decreasing.



On the other hand, f is said ***lower semi-continuous*** if for every $x \in \mathbb{R}$ and any $\varepsilon > 0$, there exists δ such that

$$d(x, y) \leq \delta \;\Rightarrow\; f(y) \geq f(x) - \varepsilon.$$

In other words, $\liminf_{y\to x} f(y) \geq f(x)$, which means that f cannot decrease with a discontinuity; see the figure below.



If f is a convex function and $f(a)$ and $f(b)$ are finite, then f is continuous on the whole interval $[a,b]$. Indeed, if $x<y$ are two points in $[a,b]$, then

$$\frac{f(x)-f(a)}{x-a} \leq \frac{f(y)-f(a)}{y-a} \leq \frac{f(y)-f(x)}{y-x} \leq \frac{f(b)-f(x)}{b-x},$$

so if $\Delta_f(a,x)$ and $\Delta_f(x,b)$ are the left- and right-hand sides of the previous inequality, then

$$f(x)+\Delta_f(a,x)\,(y-x) \leq f(y) \leq f(x)+\Delta_f(x,b)\,(y-x)$$

around x, which ensures the continuity. It follows that a convex function from $\mathbb{R}$ to $(-\infty,+\infty]$ is continuous on an interval I, and equal to $+\infty$ outside I. If one demands f to be lower semi-continuous, then I is a closed interval, and moreover, one has then

$$f = \sup\{g,\ g \text{ affine and smaller than } f\}.$$

Fix $f:\mathbb{R}\to(-\infty,+\infty]$ not equal everywhere to $+\infty$, and let us check that f^* is convex. If $x\in[a,b]$, then for any t,

$$xt-f(t) = \frac{x-a}{b-a}\,(bt-f(t)) + \frac{b-x}{b-a}\,(at-f(t)) \leq \frac{x-a}{b-a}\,f^*(b) + \frac{b-x}{b-a}\,f^*(a),$$

so by taking the supremum, $f^*(x)$ indeed satisfies the inequality of convexity. The Legendre-Fenchel transform is also automatically lower semi-continuous:

$$\begin{aligned}\liminf_{y\to x} f^*(y) &= \lim_{\varepsilon\to 0}\left(\inf_{d(y,x)\leq\varepsilon}\sup_{t\in\mathbb{R}}(ty-f(t))\right) \geq \lim_{\varepsilon\to 0}\left(\sup_{t\in\mathbb{R}}(tx-f(t))-t\varepsilon\right)\\ &\geq \sup_{t\in\mathbb{R}}(tx-f(t)) = f^*(x).\end{aligned}$$

By the previous discussion, it is even continuous on $\mathcal{D}(f^*)=\{x\in\mathbb{R},\ f^*(x)<+\infty\}$, which is a closed interval. Then, let us prove the following: the biconjugate f^{**} is the largest lower semi-continuous and convex function that is smaller than f. First, for any x,

$$\begin{aligned}f^{**}(x) &= \sup_{t\in\mathbb{R}}(tx-f^*(t)) = \sup_{t\in\mathbb{R}}\left(tx-\sup_{s\in\mathbb{R}}(ts-f(s))\right)\\ &\leq \sup_{t\in\mathbb{R}}(tx-(tx-f(x))) = \sup_{t\in\mathbb{R}} f(x) = f(x),\end{aligned}$$

so f^{**} is indeed a lower semi-continuous and convex function that is smaller than f. Conversely, let $g \leq f$ be an affine function. Then, for any s and x,

$$g(x) = g(s) + t(x-s) \leq f(s) + t(x-s)$$

for some t, so $g(x) \leq \inf_{s\in\mathbb{R}}(f(s) + t(x-s)) = tx - f^*(t)$ for some t, and

$$g(x) \leq \sup_{t\in\mathbb{R}}(tx - f^*(t)) = f^{**}(x).$$

Since a lower semi-continuous convex function is the supremum of the affine functions that are smaller, it follows that indeed

$$f^{**} = \sup\{g,\ g \text{ lower semi-continuous convex function smaller than } f\}.$$

In particular, if $f = f^{**}$, then f is lower semi-continuous and convex, and conversely, if f is lower semi-continuous and convex, then $f \leq f^{**} \leq f$, so $f = f^{**}$. □

EXAMPLE. Let X be a real-valued random variable, and $\Lambda(t) = \log \mathbb{E}[\mathrm{e}^{tX}]$ its cumulant generating series. To begin with, we assume that $|X|$ is almost surely bounded in absolute value by some constant M; this ensures the existence and analyticity of $\Lambda(t)$ on $\mathbb{R}$. Then, it is easy to verify that $\Lambda(t)$ is convex:

$$\Lambda'(t) = \frac{\mathbb{E}[X\,\mathrm{e}^{tX}]}{\mathbb{E}[\mathrm{e}^{tX}]} = \mathbb{E}'[X]$$
$$\Lambda''(t) = \frac{\mathbb{E}[X^2\,\mathrm{e}^{tX}]\,\mathbb{E}[\mathrm{e}^{tX}] - \mathbb{E}[X\,\mathrm{e}^{tX}]^2}{\mathbb{E}[\mathrm{e}^{tX}]^2} = \mathrm{Var}'[X] \geq 0.$$

where the $\mathbb{E}'$ means that one computes the expectation under the new probability

$$d\mathbb{P}' = \frac{\mathrm{e}^{tX}}{\mathbb{E}[\mathrm{e}^{tX}]}\,d\mathbb{P}.$$

In the general case, by Lebesgue's monotone convergence theorem,

$$\Lambda(t) = \sup_{M\in\mathbb{R}_+} \log \mathbb{E}[\mathrm{e}^{tX}\,\mathbf{1}_{|X|\leq M}],$$

so one conserves the inequalities of convexity. On the other hand, as the supremum of continuous functions, Λ is a lower semi-continuous function; indeed, it is easy to verify from the definition that lower semi-continuous functions are closed by upper bound. In particular, $\Lambda^{**} = \Lambda$.

EXAMPLE. Let us detail the previous example for classical distributions. If $X = \mathcal{N}(m, \sigma^2)$ is a Gaussian variable of mean m and variance σ^2, then $\mathbb{E}[\mathrm{e}^{tX}] = \exp(tm + \frac{t^2\sigma^2}{2})$, so

$$\Lambda_X(t) = tm + \frac{t^2\sigma^2}{2}$$

which is indeed convex, and

$$\Lambda_X^*(u) = \frac{1}{2}\left(\frac{u-m}{\sigma}\right)^2.$$

If X is a Bernoulli variable equal to ± 1 with probability $\frac{1}{2}$ for each possibility, then $\mathbb{E}[\mathrm{e}^{tX}] = \cosh t$, $\Lambda_X(t) = \log\cosh t$ and

$$\Lambda_X^*(u) = u\operatorname{arctanh} u - \log\cosh\operatorname{arctanh} u = \begin{cases} \frac{1+u}{2}\log(1+u) + \frac{1-u}{2}\log(1-u) & \text{if } |u|<1; \\ +\infty & \text{otherwise.} \end{cases}$$

Finally, if $X = \mathcal{P}(m)$ is a Poisson variable of mean m, then $\Lambda_X(t) = m(\mathrm{e}^t - 1)$ and

$$\Lambda_X^*(u) = \begin{cases} (m-u) + u\log\frac{u}{m} & \text{if } u > 0, \\ +\infty & \text{otherwise.} \end{cases}$$

3.1.2. Cramér's large deviations. We are now ready to prove that the previously demonstrated upper bound for the probability of deviation of a mean of i.i.d. variables is a sharp bound. We follow very closely [**Dembo and Zeitouni, 1998**, Chapter 2].

THEOREM 3.3 (Cramér). *Denote as usual Z_n the mean of n independent copies of a random variable X. For any closed set $F \subset \mathbb{R}$,*

$$\limsup_{n\to\infty} \frac{1}{n}\log\mathbb{P}[Z_n \in F] \leq -\inf_{u\in F}\Lambda_X^*(u),$$

and for any open set $U \subset \mathbb{R}$,

$$\liminf_{n\to\infty} \frac{1}{n}\log\mathbb{P}[Z_n \in U] \geq -\inf_{u\in U}\Lambda_X^*(u).$$

LEMMA 3.4. *The following are equivalent:*

(i) $\mathcal{D}(\Lambda_X) = \{0\}$, i.e., $\Lambda_X(0) = 0$ *and* $\Lambda_X(t \neq 0) = +\infty$.

(ii) $\Lambda_X^* = 0$ *everywhere on* $\mathbb{R}$.

Otherwise, $\mathbb{E}[X]$ exists, possibly as an extended real number ($\pm\infty$). If $\mathbb{E}[X]$ is finite, then $\Lambda_X^*(\mathbb{E}[X]) = 0$. In every case, $\inf_{u\in\mathbb{R}}\Lambda_X^*(u) = 0$.

PROOF. If $\mathcal{D}(\Lambda_X) = \{0\}$, then for every $u \in \mathbb{R}$,

$$\Lambda_X^*(u) = \sup_{t\in\mathbb{R}}(ut - \Lambda_X(t)) = u\,0 - \Lambda_X(0) = 0,$$

so $(i) \Rightarrow (ii)$. Conversely, since the Legendre-Fenchel transform restricted to the set of convex lower semi-continuous functions is an involution, $(ii) \Rightarrow (i)$. Suppose now $\mathcal{D}(\Lambda_X) \neq \{0\}$, which means that there exists $t \neq 0$ with $\Lambda_X(t) < +\infty$. By symmetry, one can assume for instance $t > 0$. Then, since $X \leq \frac{\mathrm{e}^{tX}}{t}$,

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,d\mathbb{P}(\omega) \leq \int_\Omega \frac{\mathrm{e}^{tX(\omega)}}{t}\,d\mathbb{P}(\omega) = \frac{\exp\Lambda_X(t)}{t} < +\infty$$

so $\mathbb{E}[X]$ exists in $[-\infty, +\infty)$. Similarly, if $\Lambda_X(t) < \infty$ for some $t < 0$, then $\mathbb{E}[X]$ exists in $(-\infty, +\infty]$.

Since $\Lambda_X(0) = 0$, in every case, $\Lambda_X^*(u) = \sup_{t\in\mathbb{R}}(ut - \Lambda_X(t)) \geq u\,0 - \Lambda_X(0) = 0$, that is, Λ_X^* is a non-negative function. Suppose $\mathbb{E}[X] = m$ finite. Then, by Jensen's inequality,

$$\Lambda_X(t) = \log \mathbb{E}[\mathrm{e}^{tX}] \geq \mathbb{E}[\log \mathrm{e}^{tX}] = t\,\mathbb{E}[X] = tm,$$

so $\Lambda_X^*(m) = \sup_{t\in\mathbb{R}}(mt - \Lambda_X(t)) \leq 0$, and one has in fact equality. Thus, we already know that $\inf_{u\in\mathbb{R}} \Lambda_X^*(u) = 0$ in the following situations:

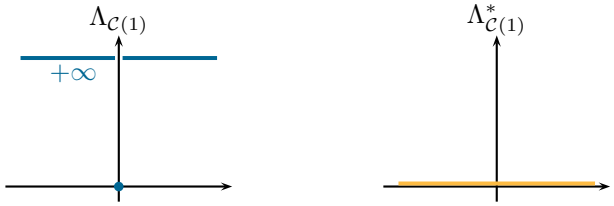
- if $\mathcal{D}(\Lambda_X) = \{0\}$, because in this case $\Lambda_X^* = 0$ everywhere;
- or, if $\mathbb{E}[X] = m$ is finite.

Suppose finally $\mathcal{D}(\Lambda_X) \cap \mathbb{R}_+^* \neq \emptyset$, but $\mathbb{E}[X] = -\infty$ (the other case is symmetric). Then, for every u,

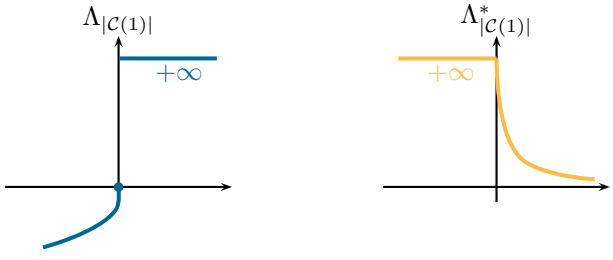
$$\log \mathbb{P}[X - u \geq 0] \leq \inf_{t\in\mathbb{R}_+} \log \mathbb{E}[\mathrm{e}^{tX-tu}] = -\sup_{t\in\mathbb{R}_+}(ut - \Lambda_X(t)) = -\Lambda_X^*(u)$$

because $\Lambda_X(t) = +\infty$ if $t < 0$. It follows that $\lim_{u\to-\infty} \Lambda_X^*(u) = 0$, since the left-hand side in the previous inequality trivially goes to 0. □

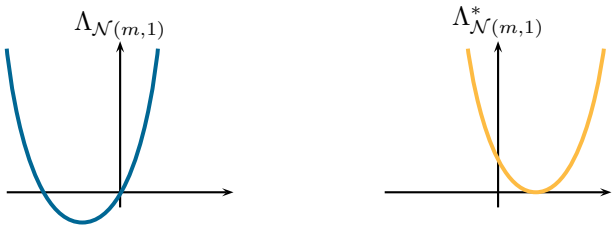
Example. Suppose that X follows a Cauchy law of parameter 1: $d\mathbb{P}_X = \frac{1}{\pi(1+x^2)}\,dx$. Then, $\Lambda_X(t) = \mathbf{1}_{t\neq 0}\,\infty$, which is the first case in the previous lemma.



Consider then the case when X is the absolute value of a Cauchy variable of parameter 1: $d\mathbb{P}_X = \frac{2\,\mathbf{1}_{x\geq 0}}{\pi(1+x^2)}\,dx$. Then, $\Lambda_X(t)$ is finite for $t \leq 0$, but $\mathbb{E}[X] = +\infty$. In this case the aspect of the cumulant generating function and its Legendre-Fenchel transform is:



Finally, if X is a Gaussian variable of mean m and variance 1, then $\Lambda_X(t) = \frac{(t+m)^2-m^2}{2}$ and $\Lambda_X^*(u) = \frac{(u-m)^2}{2}$, and in particular one has indeed $\Lambda_X^*(m) = 0$.



LEMMA 3.5. *Suppose* $\mathbb{E}[X] > -\infty$. *Then, for all* $u \leq \mathbb{E}[X]$,

$$\Lambda_X^*(u) = \sup_{t \in \mathbb{R}_-} (ut - \Lambda_X(t))$$

and Λ_X^* *is non-increasing on* $(-\infty, \mathbb{E}[X])$. *Similarly, if* $\mathbb{E}[X] < +\infty$, *then for all* $u \geq \mathbb{E}[X]$,

$$\Lambda_X^*(u) = \sup_{t \in \mathbb{R}_+} (ut - \Lambda_X(t))$$

and Λ_X^* *is non-decreasing on* $(\mathbb{E}[X], +\infty)$.

PROOF. Let us treat for instance the second case. If $\mathbb{E}[X] = -\infty$, then $\Lambda_X(t) = +\infty$ for any $t < 0$, so $ut - \Lambda(t) = -\infty$ if $t < 0$. Therefore,

$$\sup_{t \in \mathbb{R}} (ut - \Lambda_X(t)) = \sup_{t \in \mathbb{R}_+} (ut - \Lambda_X(t))$$

for any $u \in (-\infty, +\infty)$. On the other hand, if $\mathbb{E}[X] = m$ is finite, then for $u \geq m$, and $t < 0$,

$$ut - \Lambda_X(t) \leq mt - \Lambda_X(t) \leq \Lambda_X^*(m) = 0,$$

so again

$$0 \leq \sup_{t \in \mathbb{R}} (ut - \Lambda_X(t)) = \sup_{t \in \mathbb{R}_+} (ut - \Lambda_X(t)).$$

Since $u \mapsto ut - \Lambda(t)$ is non-decreasing in u for any $t \in \mathbb{R}_+$, this also implies that Λ_X^* is non-decreasing on $(m, +\infty)$. □

PROOF OF THEOREM 3.3. The upper bound is mainly a consequence of Chernov's inequality and of the previous lemmas. Fix a non-empty closed set $F \subset \mathbb{R}$. The inequality

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}[Z_n \in F] \leq - \inf_{u \in F} \Lambda_X^*(u)$$

is obvious if the infimum is equal to 0, so one can suppose $\inf_{u \in F} \Lambda_X^*(u) > 0$. In particular, this implies that Λ_X^* is not identically equal to zero, so by Lemma 3.4, $\mathbb{E}[X]$ exists as an extended real number. Recall that for every $u \in \mathbb{R}$

$$\mathbb{P}[Z_n \geq u] \leq \exp\left(-n \sup_{t \in \mathbb{R}_+} (ut - \Lambda_X(t))\right),$$

and similarly,

$$\mathbb{P}[Z_n \leq u] \leq \exp\left(-n \sup_{t \in \mathbb{R}_-} (ut - \Lambda_X(t))\right).$$

The previous lemma ensures then for all u in $(-\infty, \mathbb{E}[X])$, $\mathbb{P}[Z_n \leq u] \leq \exp(-n\Lambda_X^*(u))$, and for all u in $(\mathbb{E}[X], +\infty)$, $\mathbb{P}[Z_n \geq u] \leq \exp(-n\Lambda_X^*(u))$. Here it is understood that $(-\infty, -\infty) = \emptyset$ and $(+\infty, +\infty) = \emptyset$.

(1) Suppose first that $\mathbb{E}[X] = m$ is finite. Since $\Lambda_X^*(m) = 0$, m is in the complementary of F, which is open; so there exists an interval (a, b), which one can assume maximal, such that $m \in (a, b) \subset F^{\mathrm{c}}$. Then,

$$\mathbb{P}[Z_n \in F] \leq \mathbb{P}[Z_n \leq a] + \mathbb{P}[Z_n \geq b] \leq \mathbf{1}_{a>-\infty}\, \mathrm{e}^{-n\,\Lambda_X^*(a)} + \mathbf{1}_{b<+\infty}\, \mathrm{e}^{-n\,\Lambda_X^*(b)}.$$

Since F is closed, if $a > -\infty$, then $a \in F$, so $\Lambda_X^*(a) \geq \inf_{u\in F} \Lambda_X^*(u)$. Similarly, if $b < +\infty$, then $b \in F$, so $\Lambda_X^*(b) \geq \inf_{u\in F} \Lambda_X^*(u)$. So finally,

$$\mathbb{P}[Z_n \in F] \leq 2\, \mathrm{e}^{-n \inf_{u\in F} \Lambda_X^*(u)},$$

which proves the upper bound in the case of finite expectation random variables.

(2) Suppose now that $\mathbb{E}[X]$ is infinite, say, equal to $-\infty$. We have seen before that in this case, $\lim_{u\to-\infty} \Lambda_X^*(u) = 0$. Since $\inf_{u\in F} \Lambda_X^*(u) > 0$, the set F is therefore bounded from below. Denote $x = \inf F > -\infty$; since $x \in F$,

$$\mathbb{P}[Z_n \in F] \leq \mathbb{P}[Z_n \geq x] \leq \mathrm{e}^{-n\Lambda^*(x)} \leq \mathrm{e}^{-n \inf_{u\in F} \Lambda_X^*(u)}.$$

So, the upper bound is also shown in the case of r.v. with infinite expectation.

For the lower bound, suppose first that the law of X charges both $(0, +\infty)$ and $(-\infty, 0)$, but is supported by a bounded subset of $\mathbb{R}$. This implies that $\Lambda_X(t) < +\infty$ for every $t \in \mathbb{R}$, and also that $\lim_{|t|\to\infty} \Lambda_X(t) = +\infty$. Consequently, Λ_X and Λ_X^* are continuous convex functions with values in $\mathbb{R}$. By Lebesgue's dominated convergence theorem, Λ_X is then even differentiable, with

$$\Lambda_X'(t) = \frac{\mathbb{E}[X\, \mathrm{e}^{tX}]}{\mathbb{E}[\mathrm{e}^{tX}]}.$$

Let t_0 be such that $\Lambda_X(t_0) = \inf_{t\in\mathbb{R}} \Lambda_X(t)$, and therefore $\Lambda_X'(t_0) = 0$. Denote μ the law of X under $\mathbb{P}$, and μ_0 the law of X under the new probability:

$$d\mathbb{P}_0(\omega) = \frac{\mathrm{e}^{t_0 X(\omega)}}{\mathbb{E}[\mathrm{e}^{t_0 X}]}\, d\mathbb{P}(\omega).$$

This ***exponential change of probability measure*** will be a central argument of the theory of large deviations. If A_n is the mean of i.i.d. variables of law μ_0, then for every

$\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}[A_n \in (-\varepsilon, \varepsilon)] &= \int_{\left|\sum_{i=1}^n x_i\right| < n\varepsilon} \mu_0(dx_1) \cdots \mu_0(dx_n) \\
&= \mathrm{e}^{-n\Lambda_X(t_0)} \int_{\left|\sum_{i=1}^n x_i\right| < n\varepsilon} \exp\left(t_0 \sum_{i=1}^n x_i\right) \mu(dx_1) \cdots \mu(dx_n) \\
&\leq \mathrm{e}^{n(|t_0|\varepsilon - \Lambda_X(t_0))} \int_{\left|\sum_{i=1}^n x_i\right| < n\varepsilon} \mu(dx_1) \cdots \mu(dx_n) \\
&\leq \mathrm{e}^{n(|t_0|\varepsilon - \Lambda_X(t_0))}\ \mathbb{P}[Z_n \in (-\varepsilon, \varepsilon)].
\end{aligned}$$

By choice of t_0, the first moment of μ_0 is zero, so by the law of large numbers,

$$\lim_{n\to\infty} \mathbb{P}[A_n \in (-\varepsilon, \varepsilon)] = 1.$$

Therefore, for $\eta > \varepsilon > 0$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[Z_n \in (-\eta, \eta)] \geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[Z_n \in (-\varepsilon, \varepsilon)] \geq \Lambda_X(t_0) - \varepsilon|t_0|.$$

Taking the limit $\varepsilon \to 0$, one concludes that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[Z_n \in (-\eta, \eta)] \geq \Lambda_X(t_0) = -\Lambda_X^*(0)$$

since $\Lambda_X^*(0) = \sup_{t\in\mathbb{R}} -\Lambda_X(t)$ is attained for $t = t_0$.

If $\mu(0, +\infty) = 0$ or $\mu(-\infty, 0) = 0$, then the previous inequality is also satisfied. Indeed, Λ_X is then a monotone function with $\inf_{t\in\mathbb{R}} \Lambda_X(t) = \log \mu(\{0\}) = -\Lambda_X^*(0)$. Then,

$$\begin{aligned}
\mathbb{P}[Z_n \in (-\eta, \eta)] &\geq \mathbb{P}[Z_n = 0] = \mathbb{P}[X_1 = X_2 = \cdots = X_n = 0] \\
&\geq (\mu(\{0\}))^n = \exp(-n\Lambda_X^*(0)).
\end{aligned}$$

Finally, the same inequality also holds when $\mu(0, +\infty) > 0$, $\mu(-\infty, 0) > 0$ but μ has an unbounded support. Indeed, for every $M > 0$ such that $\mu(0, M) > 0$ and $\mu(-M, 0)$, if μ^M is the law of X conditioned to be smaller in absolute value than $|M|$, and if B_n is the mean of i.i.d. variables of law μ^M, then

$$\begin{aligned}
\mathbb{P}[Z_n \in (-\eta, \eta)] &\geq \mathbb{P}[Z_n \in (-\eta, \eta) \text{ and } |X_i| \leq M \text{ for all } i \in [\![1, n]\!]] \\
&\geq \mu(-M, M)^n\ \mathbb{P}[A_n \in (-\eta, \eta)],
\end{aligned}$$

and therefore, by the previous discussion,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[Z_n \in (-\eta, \eta)] \geq \log \mu(-M, M) + \inf_{t\in\mathbb{R}} \Lambda_{X,M}(t)$$

where $\Lambda_{X,M}(t)$ is the cumulant generating function of μ^M. The right-hand side can also be rewritten as

$$\inf_{t\in\mathbb{R}} \left(\log \int_{-M}^{M} \mathrm{e}^{tx}\, \mu(dx)\right) \geq \min(\log \mu(0, M), \log \mu(-M, 0)).$$

Therefore, if $K = \limsup_{M\to\infty} \min_{t\in\mathbb{R}} \left(\log \int_{-M}^{M} \mathrm{e}^{tx}\,\mu(dx)\right)$, then $K > -\infty$, and on the other hand,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[Z_n \in (-\eta,\eta)] \geq -K.$$

For every M, the level set $\{t \in \mathbb{R},\ \log \int_{-M}^{M} \mathrm{e}^{tx}\,\mu(dx) \leq K\}$ is a non-empty compact set. Since the integrals are increasing with M, these compact sets are nested, so there exists by Bolzano-Weierstrass some t_0 in the intersection of all these sets. By Lebesgue monotone convergence theorem,

$$\inf_{t\in\mathbb{R}} \exp(\Lambda_X(t)) \leq \int_{-\infty}^{\infty} \mathrm{e}^{t_0 x}\,\mu(dx) \leq K,$$

so again $\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[Z_n \in (-\eta,\eta)] \geq -\Lambda_X^*(0)$.

Finally, fix an open set U and $a \in U$. There exists an interval $(a-\eta, a+\eta)$ included in U, and by applying the previous inequality to the variable $Z_n - a$,

$$\liminf_{n\to\infty} \mathbb{P}[Z_n \in U] \geq \liminf_{n\to\infty} \mathbb{P}[Z_n - a \in (-\eta,\eta)] \geq -\Lambda_{X-a}^*(0) = -\Lambda_X^*(a).$$

Since this is true for any $a \in U$, the upper bound is proved. □

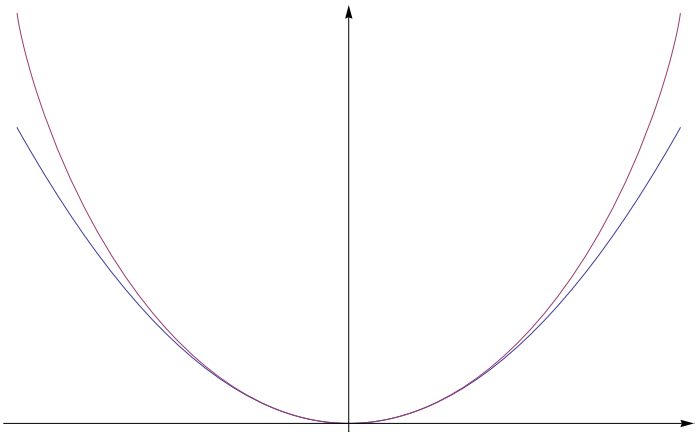
Example. Suppose that X is a Bernoulli variable equal to ± 1 with probability $\frac{1}{2}$. Hoeffding's inequality ensured that

$$\mathbb{P}[|Z_n| \geq \varepsilon] \leq 2\,\mathrm{e}^{\frac{-n\varepsilon^2}{2}}$$

for any $\varepsilon > 0$. Cramér's theorem shows that more precisely,

$$\mathbb{P}[|Z_n| \geq \varepsilon] \leq 2\,\mathrm{e}^{-n\Lambda_X^*(\varepsilon)} = 2\,\mathrm{e}^{-n\left(\frac{1+\varepsilon}{2}\,\log(1+\varepsilon)+\frac{1-\varepsilon}{2}\,\log(1-\varepsilon)\right)}$$

for $\varepsilon \in [0,1)$. The new exponent is always better, as can be seen on the following picture (the new exponent is the upper curve):



3.1.3. Moderate deviations. The rest of this chapter, and Chapter 4 are devoted to extensions and generalizations of Cramér's theorem 3.3. From the proof, it appears that the independence of the X_n's does not play such a big role, except in order to compute the Λ_{Z_n}'s. As a consequence, one can expect that a large deviation principle will also occur when the Λ_{Z_n}'s satisfy certain asymptotic property. This statement will be made rigorous by Ellis-Gärtner theorem, see §4.2. Hence, we shall be able to deal with the large deviations of a sequence of r.v. that does not necessarily come from independent random variables. On the other hand, though many arguments of monotony have been used in the previous discussion, the main tools, namely, Chernov's inequality for the upper bound and an exponential change of probability for the lower bound, can be generalized for instance to $\mathbb{R}^d$, and even to "larger" spaces whose topology will be precised in §4.1. So one expects an analogue of Cramér's theorem to hold in a multi-dimensional setting, or even possibly in an infinite-dimensional setting. In the rest of this chapter, we present in detail a case which gives many hints about what happens in the general setting, and which also constitutes an important example in this theory: the large deviations of empirical measures of finite Markov chains.

Before going on, let us review the different asymptotic regimes of a sum $X_1+\cdots+X_n = S_n$ of independent, identically distributed and centered random variables. The central limit theorem ensures that the fluctuations of S_n are typically of order $\sqrt{n}$, hence,

$$\mathbb{P}[S_n \geq x\sqrt{n}] \simeq \frac{1}{\sqrt{2\pi}\,\sigma}\int_{s=x}^{\infty} \mathrm{e}^{-\frac{s^2}{2\sigma^2}}\,ds \leq \frac{\sigma}{\sqrt{2\pi}\,x}\,\mathrm{e}^{-\frac{x^2}{2\sigma^2}}.$$

On the other hand, Cramér's theorem gives the estimate

$$\mathbb{P}[S_n \geq x\,n] \leq 2\,\mathrm{e}^{-n\,\Lambda^*(x)}$$

for fluctuations of order n, with $\Lambda^*(x)$ Legendre-Fenchel transform of the cumulant generating series of X. One can then ask what happens in between, that is for fluctuations of order a_n with $\sqrt{n} \ll a_n \ll n$. This is covered by the following result:

PROPOSITION 3.6. *Fix a sequence $(a_n)_{n\in\mathbb{N}}$ with $\sqrt{n} \ll a_n \ll n$, and suppose to simplify that Λ_X is defined over $\mathbb{R}$. Then, for $x > 0$,*

$$\lim_{n\to\infty} \frac{n}{(a_n)^2}\,\log \mathbb{P}[S_n \geq x\,a_n] = -\frac{x^2}{2\sigma^2}.$$

This to be compared to

$$\lim_{n\to\infty} \frac{1}{n}\,\log \mathbb{P}[S_n \geq x\,n] = -\Lambda^*(x).$$

PARTIAL PROOF. The upper bound is again a consequence of Chernov's inequality:

$$\mathbb{P}[S_n \geq x\,a_n] \leq \mathbb{E}\left[\mathrm{e}^{\frac{\lambda S_n\,a_n}{n}}\right]\,\mathrm{e}^{-\lambda x\,\frac{(a_n)^2}{n}} = \mathrm{e}^{n\,\Lambda_X(\frac{\lambda\,a_n}{n})-\lambda x\frac{(a_n)^2}{n}}.$$

The asymptotic expansion of Λ_X around 0 is $\Lambda_X(\varepsilon) = \frac{\sigma^2\varepsilon^2}{2} + o(\varepsilon^2)$, so,

$$\limsup_{n\to\infty} \frac{n}{(a_n)^2}\,\log \mathbb{P}[S_n \geq x\,a_n] \leq \left(\frac{\lambda^2\sigma^2}{2} - \lambda x\right).$$

Choosing $\lambda = \frac{x^2}{\sigma^2}$ ends the proof of the upper bound. For the lower bound, we refer to Section 4.3, since the hypotheses of our Proposition will allow us to apply Ellis-Gärtner theorem. □

A way to restate the previous Proposition is to say that n is the smallest order of fluctuations for which the central limit theorem does not hold anymore; before, the rate of decay of the probabilities of fluctuations is given by the universal exponent $-\frac{x^2}{2\sigma^2}$. Actually, much more can be said in the intermediate regime $\sqrt{n} \ll a_n \ll n^{2/3}$: indeed, theorems due to Bahadur and Rao ensure that if X is not distributed on a lattice, then the approximation

$$\mathbb{P}[S_n \geq x\, a_n] \simeq \mathbb{P}[\mathcal{N}_{(0,1)} \geq x\, a_n]$$

holds, that is to say that one has an equivalent of $\mathbb{P}[S_n \geq x\, a_n]$ instead of its logarithm. We refer to [**Dembo and Zeitouni, 1998**, §3.7] for precisions on these results.

3.2. Ergodic theorems for Markov chains

We start by recalling the main features of finite Markov chains.

3.2.1. Irreducible finite Markov chains and Perron-Frobenius theory. Fix a finite space of states, say, $\mathfrak{X} = [\![1, N]\!]$. A (time-homogeneous) ***Markov chain*** on $\mathfrak{X}$ is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ such that for all n, the Markov property is satisfied:

$$\mathbb{P}[X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n] = p(x_n, x_{n+1})$$

where $(p(i,j))_{i,j\in[\![1,N]\!]}$ is a stochastic matrix, that is to say a matrix with non-negative entries such that

$$\sum_{j=1}^{N} p(i,j) = 1$$

for any row $i \in [\![1, N]\!]$. One says then that p is the transition matrix of the Markov chain. A Markov chain is entirely determined in law by its transition matrix and by the law of X_0, which is a vector $(\pi_0(1), \ldots, \pi_0(N))$ in the (finite-dimensional) simplex $\mathscr{M}_1([\![1, N]\!])$. Indeed, by repetitive use of Markov's property, one can then compute all the elementary probabilities

$$\mathbb{P}[X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n] = \pi_0(x_0)\, p(x_0, x_1)\, p(x_1, x_2) \cdots p(x_{n-1}, x_n).$$

In particular, for any time n, the law of X_n is related to the n-th power of p:

$$\pi_n(x) = \mathbb{P}[X_n = x] = (\pi_0\, p^n)(x).$$

In this setting, the ***Perron-Frobenius theorem*** (see Theorem 3.9 hereafter) allows one to prove most of the asymptotic results on finite Markov chains. We first need to explain the notions of irreducibility and aperiodicity of a (non-negative) square matrix.

DEFINITION 3.7. *A square matrix $A \in \mathrm{M}(N, \mathbb{R})$ with non-negative entries is said irreducible if one of the following equivalent properties holds:*

(i) There is no non-trivial and A-invariant coordinate subspace, i.e., *for every part $\{i_1, \dots, i_r\} \neq \emptyset, [\![1, N]\!]$ of $[\![1, N]\!]$, $\mathrm{Vect}(e_{i_1}, \dots, e_{i_r})$ is not A-invariant.*

(ii) There is no permutation matrix P_σ such that

$$P_\sigma A P_{\sigma^{-1}} = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$$

with B and D non-zero matrices.

(iii) For every pair of indices (i, j), there is a natural number $m \geq 1$ such that $(A^m)_{ij} > 0$.

Otherwise, A is called a reducible matrix, and there exists a permutation matrix P_σ such that

$$P_\sigma A P_{\sigma^{-1}} = \begin{pmatrix} B_1 & * & \cdots & * \\ 0 & B_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & B_k \end{pmatrix}$$

where each B_i is either equal to 0 or irreducible.

PROOF. In what follows we always make the matrices act on the right of the vectors, which are therefore considered as row matrices of size N. If $\neg(i)$ holds, choose a permutation σ of $[\![1, N]\!]$ that sends $[\![1, r]\!]$ to $\{i_1, \dots, i_r\}$. Then, $P_\sigma A P_{\sigma^{-1}}$ stabilizes the vector space $\mathrm{Vect}(e_1, \dots, e_r)$, so it writes as a block upper triangular matrix and $\neg(ii)$ is true. Conversely, if $\neg(ii)$ is true, then, denoting r the size of the block B, the coordinate subspace $\mathrm{Vect}(e_{\sigma(1)}, \dots, e_{\sigma(r)})$ is stabilized by A. Therefore, $(i) \Leftrightarrow (ii)$. Suppose (i) false, and fix an invariant coordinate subspace $\mathrm{Vect}(e_{i_1}, \dots, e_{i_r})$. Denote $i = i_1$ and fix $j \notin \{i_1, \dots, i_r\}$. For any m, $e_i A^m \in \mathrm{Vect}(e_{i_1}, \dots, e_{i_r})$ has for j-th coefficient 0, so $(A^m)_{ij} = 0$ and (iii) is false. Conversely, if (iii) is false, fix two indices i, j such that $(A_{ij})^m = 0$ for all m, and denote C_i the set of indices k such that $(A_{ik})^m > 0$ for some m. We claim that $\mathrm{Vect}(\{e_k,\ k \in C_i\})$ is a non-trivial invariant coordinate subspace of $\mathbb{R}^N$. It is non-trivial because it contains e_i but it does not contain e_j. On the other hand, consider a linear combination $\sum_{k \in C_i} \lambda_k \, e_k$, and apply to it A; one obtains

$$\sum_{l=1}^{N} \left(\sum_{k \in C_i} \lambda_k \, A_{kl} \right) e_l.$$

If e_l occurs with a non-zero coefficient above, then some $A_{k_0 l}$ has to be non-zero itself, and on the other hand, $(A^m)_{ik_0} > 0$ for some m. Then,

$$(A^{m+1})_{il} = \sum_k (A^m)_{ik} \, A_{kl} \geq (A^m)_{ik_0} \, A_{k_0 l} > 0$$

so $l \in C_i$ and the stability is shown, as well as the equivalence $(i) \Leftrightarrow (iii)$. The last statement in the definition is a refinement of the previous discussion, where each block

B_s corresponds to indices in a connected component of the digraph with vertex set $[\![1, N]\!]$ and edge set the (i, j)'s such that $A_{ij} > 0$. □

Call period of index i of a non-negative matrix A the greatest common divisor of the integers $m \geq 0$ such that $(A^m)_{ii} > 0$.

LEMMA 3.8. *For an irreducible matrix A, the period $h(A)$ does not depend on i and is called the period of the matrix. There exists a partition of $[\![1, N]\!]$ in $h(A)$ parts $\mathfrak{X}_1 \sqcup \cdots \sqcup \mathfrak{X}_h$ such that, if σ is a permutation that sends consecutive intervals $I_1, \ldots, I_h$ of $[\![1, N]\!]$ to $\mathfrak{X}_1, \ldots, \mathfrak{X}_h$, then*

$$P_\sigma A P_{\sigma^{-1}} = \begin{pmatrix} 0 & B_1 & 0 & \cdots & 0 \\ \vdots & \ddots & B_2 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & 0 & B_{h-1} \\ B_h & 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

THEOREM 3.9 (Perron-Frobenius). *Let A be an irreducible non-negative matrix of size N and period h. There exists a positive real number $r = r(A)$ with the following properties:*

1. *For every h-th root of unity ζ, $r\zeta$ is a simple eigenvalue of A.*
2. *Every complex eigenvalue e of A satisfies $|e| \leq r$, with equality if and only if $e = r\zeta$ for some h-th root of unity.*
3. *The eigenspace associated to r contains real vectors with positive coordinates, and conversely, if v is an eigenvector of A with positive coordinates, then v is an eigenvector for r.*
4. *The Perron-Frobenius eigenvalue r satisfies*

$$\min_{i \in [\![1,N]\!]} \left(\sum_{j=1}^{N} a_{ij} \right) \leq r \leq \max_{i \in [\![1,N]\!]} \left(\sum_{j=1}^{N} a_{ij} \right).$$

In particular, if A is a stochastic matrix, then $r = 1$.

We refer to [**Meyer, 2000**, Chapter 8] for a proof of these classical algebraic results.

3.2.2. Ergodic theorems. For Markov chains, the Perron-Frobenius theory implies the following:

THEOREM 3.10 (Ergodic theorem for irreducible Markov chains). *Consider an irreducible Markov chain with space of states $[\![1, N]\!]$, and denote π the unique Perron-Frobenius vector π such that $\sum_{i=1}^{n} \pi(i) = 1$.*

1. *If the chain is aperiodic, meaning that the period h of the transition matrix p is 1, then for any initial measure μ_0, the Markov chain $(X_n)_{n \in \mathbb{N}}$ converges in law towards π:*

$$\pi_n \rightharpoonup \pi.$$

(2) *Even if the chain is not aperiodic, the* ***empirical distribution*** $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, *which is a random measure, converges almost surely towards* π*:*

$$\nu_n \to_{\text{a.s.}} \pi.$$

Here, the topology on measures is the topology of convergence in law, but since the space of states is finite this is also the strong topology inside $\mathbb{R}^N$*.*

Finally, π *is a stationary probability measure for the Markov chain, meaning that if* $\pi_0 = \pi$*, then* $\pi_n = \pi$ *for all* $n \geq 0$*.*

The second part of Theorem 3.10 is an example of ***ergodic theorem***, that is to say a statement that relates asymptotically a mean over time (the empirical measure) to a mean over space (the stationary measure).

PROOF. On a finite set, convergence in law is equivalent to the convergence of all elementary probabilities $\mathbb{P}[X_n = x] \to \pi(x)$; in other words, $\pi_n \to \pi$ in $\mathbb{R}^N$. Suppose that the transition matrix p is irreducible and aperiodic; then by Perron-Frobenius theorem, there exists (up to a scalar) a unique positive eigenvector π for the eigenvalue 1, and all the other eigenvalues are of strictly smaller module. In the following we assume π to be normalized to be a probability measure on $[\![1, N]\!]$. We then expand π_0 in a basis $(\pi, b_2, \dots, b_N)$ such that the operator associated to p has for matrix in this basis the Jordan form

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ & \lambda_2 & * & \cdots & * \\ & & \lambda_3 & \ddots & \vdots \\ & & & \ddots & * \\ & & & & \lambda_N \end{pmatrix}$$

with $|\lambda_i| < 1$ for $i \in [\![2, N]\!]$. If $R < 1$ is strictly bigger than all the modules of these other eigenvalues, then $\pi_0 = \alpha\,\pi + \sum_{i=2}^N \beta_i\, b_i$ implies that

$$\pi_0\, p^n = \alpha\,\pi + \underbrace{\left(\sum_{i=2}^N \beta_i\, b_i\right) p^n}_{\substack{\text{smaller in norm} \\ \text{than } C\,R^n}}.$$

Indeed, the underbraced term writes as $\sum_{i=2}^N \beta_i(m)\,(\lambda_i)^m\, b_i$ with the $\beta_i(m)$'s polynomials in m. Consequently, $\pi_n = \pi_0\, p^n \to \alpha\,\pi$, and since the map $(x_1, \dots, x_N) \to \sum_{i=1}^N x_i$ is continuous, as the left-hand side is always a probability measure, $\alpha = 1$. So, we have shown that $\pi_n \rightharpoonup \pi$ in the aperiodic case.

In the periodic case ($h \geq 2$), the previous result is usually false, as can be seen on the example

$$N = 2 \quad ; \quad p = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad ; \quad \pi_0 = (1, 0) \quad ; \quad \pi = \left(\frac{1}{2}, \frac{1}{2}\right).$$

However, one can still guess on this example that in mean over time, the Markov chain occupies each state i with probability $\pi(i)$. This is precisely what is meant by the second

part of the theorem, and the proof of this ergodic theorem goes as follows. The almost sure convergence claimed in our ergodic theorem is equivalent to the statement

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(X_i=k)} = \nu_n(k) \to_{\text{a.s.}} \pi(k)$$

for any $k \in [\![1, N]\!]$. Let $T_{k\to k}$ be the first time $n \geq 1$ (random) for which $X_n = k$, assuming $X_0 = k$ almost surely. This time is finite almost surely; otherwise, by irreducibility of the Markov chain, there would be a non-negative probability for not returning to the state j for any j (not only k), and then, a non-negative probability for not returning to $[\![1, N]\!]$, and of course that is absurd.

So, $T_{k\to k}$ is finite a.s., and more generally, if $T^{(m)}_{k\to k}$ is the m-th hitting time of k by the Markov chain starting from k, then $T^{(m)}_{k\to k}$ is finite a.s., and by the strong Markov property,

$$T^{(1)}_{k\to k}, T^{(2)}_{k\to k} - T^{(1)}_{k\to k}, \ldots, T^{(m)}_{k\to k} - T^{(m-1)}_{k\to k}, \ldots$$

are i.i.d. random variables. By the law of large numbers,

$$\mathbb{E}[T_{k\to k}] = \lim_{m\to\infty} \frac{T^{(m)}_{k\to k}}{m} \quad \text{almost surely,}$$

this being *a priori* an extended real number in $[0, +\infty]$. Now, denote m the unique (random) integer such that $T_{\pi_0\to k} + T^{(m)}_{k\to k} \leq n < T_{\pi_0\to k} + T^{(m+1)}_{k\to k}$. By definition,

$$\nu_n(k) = \frac{m+1}{n} \leq \frac{m+1}{T_{\pi_0\to k} + T^{(m)}_{k\to k}} \to_{\text{a.s.}} \frac{1}{\mathbb{E}[T_{k\to k}]}$$

and similarly,

$$\nu_n(k) = \frac{m+1}{n} \geq \frac{m+1}{T_{\pi_0\to k} + T^{(m+1)}_{k\to k}} \to_{\text{a.s.}} \frac{1}{\mathbb{E}[T_{k\to k}]}$$

since $T_{\pi_0\to k}$ is a.s. finite for the same reasons as before, and the number m of visits to k goes a.s. to infinity. We have therefore shown the a.s. convergence of the empirical measure towards the vector

$$\left(\frac{1}{\mathbb{E}[T_{1\to 1}]}, \ldots, \frac{1}{\mathbb{E}[T_{N\to N}]}\right).$$

However, the expectation of the empirical measure is at coordinate k

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}[X_i = k] = \frac{1}{n}\sum_{i=1}^{n}(\pi_0\, p^i)(k).$$

Even in the periodic case, one can still expand π_0 over a Jordan adapted basis of the operator p. This expansion writes

$$\pi_0 = \alpha\,\pi + \sum_{j=1}^{h-1} b_{\zeta^j} + \text{remainder associated to smaller eigenvalues,}$$

where ζ is a primitive h-th root of unity, the b_{ζ^j}'s are eigenvectors for the ζ^j, and as before π is the normalized eigenvector for the eigenvalue 1. Consequently,

$$\frac{1}{n}\sum_{i=1}^{n}(\pi_0\, p^i)(k) = \alpha\,\pi(k) + \sum_{j=1}^{h-1}\left(\frac{1}{n}\sum_{i=1}^{n}\zeta^{ij}\right) b_{\zeta^j}(k) + \text{remainder smaller than } C\,R^n,$$

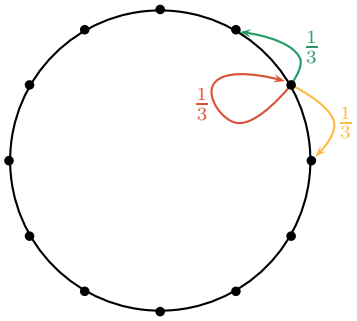
which converges to $\alpha\,\pi(k)$ as n goes to infinity. By taking the sum over $k \in [\![1, N]\!]$, one also gets $\alpha = 1$, so, the limited expectation of the empirical measure is the vector π. Lebesgue dominated convergence theorem ensures now that $\frac{1}{\mathbb{E}[T_{k\to k}]} = \pi(k)$, and the convergence a.s. of the empirical measure to π is shown. In particular, since the vector π is positive by Perron-Frobenius, $\mathbb{E}[T_{k\to k}] < +\infty$ for all k. □

EXAMPLE. If $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables on $[\![1, N]\!]$, then the empirical measure of these realizations of X converges a.s. to the law of X.

EXAMPLE. Consider the random walk on N points placed on a circle, with at each step a probability 1/3 to jump from one site to the left neighbor, 1/3 to jump to the right neighbor, and 1/3 to stay at the same position. The matrix of transition of this Markov chain is given hereafter, and it can be written as

$$p = \frac{C_N + I_N + (C_N)^{-1}}{3},$$

where C_N is the circulant matrix with 1's over the diagonal and in the bottom left corner.



$$\begin{pmatrix} 1/3 & 1/3 & & & 1/3 \\ 1/3 & 1/3 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1/3 \\ 1/3 & & & 1/3 & 1/3 \end{pmatrix}$$

As a consequence, the diagonalization of p is equivalent to the diagonalization of C_N. However, C_N is solution of the equation $X^N - 1 = 0$, so its eigenvalues are the N roots of unity $1, \zeta = \mathrm{e}^{\frac{2\mathrm{i}\pi}{N}}, \zeta^2, \ldots, \zeta^{N-1}$. Consequently, the eigenvalues of p are

$$1, \frac{1+2\cos\zeta}{3}, \frac{1+2\cos 2\zeta}{3}, \ldots, \frac{1+2\cos(N-1)\zeta}{3}$$

and they are all smaller than 1 but the first eigenvalue, associated to the Perron-Frobenius vector $\pi = (\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N})$. It follows from the ergodic theorems that the law of the random walk X_N and the empirical law converges towards this Perron-Frobenius vector, which is the uniform law on $[\![1, N]\!]$.

The proof of the ergodic theorem says nothing about the speed of convergence, but on the other hand, since it uses the law of large numbers for independent random variables (the times $T_{k\to k}^{(m)} - T_{k\to k}^{(m-1)}$), one can expect that an exponential control can indeed be shown. This is the purpose of the next section, which is devoted to Sanov's theorem.

3.3. Entropy and Sanov's theorem

In this section, we fix a finite space of states $\mathfrak{X} = [\![1, N]\!]$ and an irreducible Markov chain on $\mathfrak{X}$ with transition kernel p and stationary measure π.

3.3.1. Relative entropy and a statement of Sanov's theorem. For Markov chains, the role of the Legendre-Fenchel transform (the rate of decay in principles of large deviations for sums of i.i.d.) will be played by the (relative) ***entropy*** of measures, and by generalizations of it.

DEFINITION 3.11. *The entropy of a probability measure $\mu \in \mathscr{M}^1(\mathfrak{X})$ is*

$$H(\mu) = -\sum_{i=1}^{N} \mu(i)\, \log \mu(i),$$

and the relative entropy (also known as Kullback-Leibler divergence) of a probability measure ν with respect to μ is

$$H(\mu||\nu) = \sum_{i=1}^{N} \mu(i)\, \log \frac{\mu(i)}{\nu(i)}.$$

These quantities are non-negative, and $H(\mu||\nu) = 0$ if and only if $\mu = \nu$.

PROOF. The non-negativity of the entropy is obvious, since every term is non-negative; it is not hard to see that $H(\mu) = 0$ if and only if μ is concentrated on one point $i \in [\![1, N]\!]$. For the relative entropy, we use Jensen's inequality on the convex function $\phi(x) = x \log x$:

$$H(\mu||\nu) = \sum_{i=1}^{N} \nu(i)\, \phi(x_i) \geq \phi\left(\sum_{i=1}^{N} \nu(i)\, x_i\right) = \phi(1) = 0$$

where $x_i = \frac{\mu(i)}{\nu(i)}$. Since ϕ is strictly convex, there is equality if and only if $x_i = 1$ for all $i \in [\![1, N]\!]$, *i.e.*, $\mu = \nu$. □

For ν probability measure on $\mathfrak{X}$, we set

$$I(\nu) = \sup_{\mu \in \mathscr{M}^1(\mathfrak{X})} \left(\sum_{i=1}^{N} \nu(i)\, \log \frac{\mu(i)}{(\mu\, p)(i)}\right).$$

Notice that if the Markov chain is a sequence of independent identically distributed random variables, then $\mu p = \pi$ is the law of these r.v. for any μ, and therefore, by calculus of variations with respect to μ,

$$I(\nu) = \sup_{\mu \in \mathscr{M}^1(\mathfrak{X})} \left(\sum_{i=1}^{N} \nu(i) \, \log \frac{\mu(i)}{\pi(i)} \right) = \sum_{i=1}^{N} \nu(i) \, \log \frac{\nu(i)}{\pi(i)} = H(\nu || \pi)$$

is the relative entropy with respect to π. In the general case, one can always give an equation for the measure μ that realizes the supremum, by using the Lagrange principle; see the discussion of Corollary 4.7.

THEOREM 3.12 (Sanov). *As in Theorem 3.10, we denote $\nu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ the random empirical measures of the Markov chain. For any closed set $F \subset \mathscr{M}^1(\mathfrak{X})$,*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}[\nu_n \in F] \leq - \inf_{\nu \in F} I(\nu),$$

and for any open set $U \subset \mathscr{M}^1(\mathfrak{X})$,

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}[\nu_n \in U] \geq - \inf_{\nu \in U} I(\nu).$$

The similarity of this statement with Cramér's theorem 3.3 will lead us to define general principles of large deviations in §4.1, by requiring similar asymptotics for the probabilities of being in a given closed or open set. Actually, this framework (and more precisely, Ellis-Gärtner theorem, see Sections 4.2 and 4.3.4) is needed in order to prove Sanov's theorem in full generality.

3.3.2. Method of types and Sanov's theorem for independent variables. Fortunately, when the X_n's are i.i.d. random variables, one can prove the theorem by elementary combinatorial arguments, and the so-called ***method of types***. This discussion makes appear the notions of entropy in a very natural way.

PROOF OF THEOREM 3.12 FOR I.I.D. RANDOM VARIABLES. To any sequence $x = (x_1, \dots, x_n) \in \mathfrak{X}^n$, we associate the empirical measure $\nu_x = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$. Notice then that for a sequence of i.i.d. random variables with law π,

$$\begin{aligned} \mathbb{P}[X_1 = x_1, \dots, X_n = x_n] &= \prod_{i=1}^{n} \pi(x_i) = \prod_{y \in \mathfrak{X}} \pi(y)^{n \nu_x(y)} \\ &= \exp \left(\sum_{y \in \mathfrak{X}} n \, \nu_x(y) \log \pi(y) \right) = \mathrm{e}^{-n(H(\nu_x) + H(\nu_x || \pi))}. \end{aligned}$$

Call type a class T of sequences $x = (x_1, \dots, x_n)$ in $\mathfrak{X}^n$ that differ by a permutation of the coordinates; for instance, if $N = 2$ and $n = 5$, then

$$T = \{(1,1,1,1,2), (1,1,1,2,1), (1,1,2,1,1), (1,2,1,1,1), (2,1,1,1,1)\}$$

is the type of sequences with four 1's and one 2. The cardinality of the type $T(x)$ of a sequence x is given by the multinomial coefficient

$$\frac{n!}{\prod_{y\in\mathfrak{X}} n\nu_x(y)!}\,.$$

Recall Stirling's estimate for factorials: $k\log\frac{k}{\mathrm{e}} \le \log k! \le k\log\frac{k}{\mathrm{e}} + \frac{1}{2}\log k + 1$. Consequently,

$$\begin{aligned}\log(\operatorname{card} T(x)) &\le n\log\left(\frac{n}{\mathrm{e}}\right) + \frac{1}{2}\log n + 1 - \sum_{y\in\mathfrak{X}} n\nu_x(y)\,\log\left(\frac{n\nu_x(y)}{\mathrm{e}}\right)\\ &\le -n\sum_{y\in\mathfrak{X}}\nu_x(y)\log\nu_x(y) + \frac{1}{2}\log n + 1 = nH(\nu_x) + \frac{1}{2}\log n + 1;\end{aligned}$$

and similarly,

$$\log(\operatorname{card} T(x)) \ge nH(\nu_x) - N\left(\frac{1}{2}\log n + 1\right)$$

So, there exists polynomials in n that are independent of the type and such that

$$\frac{1}{P_1(n)}\,\mathrm{e}^{nH(\nu_x)} \le \operatorname{card} T(x) \le P_2(n)\,\mathrm{e}^{nH(\nu_x)}.$$

Fix now a Borel subset $B \subset \mathscr{M}^1(\mathfrak{X})$, and let us compute bounds for $\mathbb{P}[\nu_n \in B]$. Notice that two sequences have the same type if and only if they yield the same empirical measure ν_x. Denote $\mathscr{T}_n$ the subset of $\mathscr{M}^1(\mathfrak{X})$ that consists in empirical measures coming from n-sequences. Then, for the upper bound, one can write:

$$\begin{aligned}\mathbb{P}[\nu_n \in B] &= \sum_{\nu\in B\cap\mathscr{T}_n} \mathbb{P}[\nu_n = \nu] = \sum_{\nu\in B\cap\mathscr{T}_n}\sum_{\nu_x=\nu} \mathbb{P}[X_1 = x_1,\dots,X_n = x_n]\\ &\le \sum_{\nu\in B\cap\mathscr{T}_n} P_2(n)\,\mathrm{e}^{nH(\nu)}\,\mathrm{e}^{-n(H(\nu)+H(\nu||\pi))}\\ &\le (\operatorname{card}\mathscr{T}_n)\,P_2(n)\,\mathrm{e}^{-n\,\inf_{\nu\in B} H(\nu||\pi)}.\end{aligned}$$

The number of types of order n is the number of way to split n into N non-negative integers, so it is given by the binomial coefficient $\binom{n+N-1}{N-1}$, which is a polynomial in n. Hence, there exists a polynomial $P_3(n)$ such that

$$\frac{1}{n}\log\mathbb{P}[\nu_n \in B] \le \frac{1}{n}\log P_3(n) - \inf_{\nu\in B} H(\nu||\pi),$$

which proves the upper bound since the logarithm of the polynomial is a $o(n)$.

For the lower bound, fix an open set $U \subset \mathscr{M}^1(\mathfrak{X})$ and a measure $\nu_\varepsilon \in U$ such that $H(\nu_\varepsilon||\pi) \le \inf_{\nu\in U} H(\nu||\pi) + \varepsilon$. For n big enough, one can always approximate the weights $\nu_\varepsilon(1), \nu_\varepsilon(2), \dots, \nu_\varepsilon(N)$ by fractions over n, in such a way that the approximation $\nu_{\varepsilon,n}$ is in $\mathscr{T}_n$ and satisfies $H(\nu_{\varepsilon,n}||\pi) \le H(\nu_\varepsilon||\pi) + \varepsilon$. This approximation is in U for n big enough because U is open. Then,

$$\mathbb{P}[\nu_n \in U] \ge \mathbb{P}[\nu_{\varepsilon,n} \in U] \ge \frac{1}{P_1(n)}\,\mathrm{e}^{-n\,H(\nu_{\varepsilon,n}||\pi)},$$

and taking the logarithms and the liminf gives

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[\nu_n \in U] \geq - \inf_{\nu\in U} H(\nu||\pi) - 2\varepsilon.$$

Since this is true for every $\varepsilon > 0$, the theorem is proved in the i.i.d. case. □

3.3.3. Large deviations for empirical pair measures. A modification of the method of types can be used in the general case of Markov chains, but it does not readily gives Sanov's theorem. The starting point is the possibility to compute the probability of any sequence $x = (x_1, \dots, x_n)$. Suppose to simplify that the initial distribution π_0 is the stationary measure π. We set $x_0 = x_n$ for symmetry reasons. Then,

$$\mathbb{P}[X_1 = x_1, \dots, X_n = x_n] = \pi(x_1) \prod_{i=2}^{n} p(x_{i-1}, x_i)$$

wich leads us to introduce the empirical pair measure

$$\nu_x^{(2)} = \frac{1}{n} \sum_{i=1}^{n} \delta_{(x_{i-1},x_i)} \in \mathscr{M}^1(\mathfrak{X} \times \mathfrak{X}),$$

and to show a principle of large deviations for this empirical pair measure instead of the empirical measure. For $\nu \in \mathscr{M}^1(\mathfrak{X} \times \mathfrak{X})$, set

$$J(\nu) = \begin{cases} H(\nu||\nu_1 \otimes p) & \text{if } \nu_1 = \nu_2, \\ +\infty & \text{otherwise,} \end{cases}$$

where ν_1 and ν_2 are the two marginales of ν, and $(\mu \otimes p)(x, y) = \mu(x)\, p(x, y)$ for $\mu \in \mathscr{M}^1(\mathfrak{X})$. A measure in $\mathscr{M}^1(\mathfrak{X} \times \mathfrak{X})$ with equal marginales will be called shift-invariant. Notice that the marginales of an empirical pair measure $\nu_x^{(2)}$ are always equal, and equal to ν_x.

THEOREM 3.13. *For an irreducible Markov chain on $\mathfrak{X}$ with transition kernel p and stationary measure π, assuming moreover $\pi_0 = \pi$, one has*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[\nu_n^{(2)} \in F] \leq - \inf_{\nu\in F} J(\nu)$$

for any closed set $F \subset \mathscr{M}^1(\mathfrak{X} \times \mathfrak{X})$, and

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[\nu_n^{(2)} \in U] \geq - \inf_{\nu\in U} J(\nu)$$

for any open set $U \subset \mathscr{M}^1(\mathfrak{X} \times \mathfrak{X})$.

PROOF. For $g : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$, consider the modified Markov chain with transition kernel

$$p^{(g)}(x, y) = \mathrm{e}^{g(x,y) - U^{(g)}(x)}\, p(x, y),$$

where $U^{(g)}(x)$ is the normalization constant

$$U^{(g)}(x) = \log \left(\sum_{y\in\mathfrak{X}} \mathrm{e}^{g(x,y)}\, p(x, y) \right).$$

It is still irreducible, so the modified stationary measure $\pi^{(g)}$ as well as the stationary measure π are positive on $\mathfrak{X}$. The reader should notice the similarity between this change of process and the change of probability measures involved in the proof of the lower bound of Cramér's theorem 3.3. That said, note that

$$\begin{aligned}
1 &= \sum_{x_1,\dots,x_n} \pi^{(g)}(x_1) \prod_{i=2}^{n} p^{(g)}(x_{i-1},x_i) = \sum_{x_1,\dots,x_n} \pi^{(g)}(x_1) \prod_{i=2}^{n} p(x_{i-1},x_i)\, \mathrm{e}^{g(x_{i-1},x_i)-U^{(g)}(x_{i-1})} \\
&\geq \left(\inf_{(y,z)\in\mathfrak{X}^2} \frac{\pi^{(g)}(y)\, p(y,z)}{\pi(y)\, p^{(g)}(y,z)} \right) \sum_{x_1,\dots,x_n} \left(\pi(x_1) \prod_{i=2}^{n} p(x_{i-1},x_i) \right) \mathrm{e}^{n\int (g(y,z)-U^{(g)}(y))\, \nu_x^{(2)}(dy,dz)} \\
&\geq \left(\inf_{(y,z)\in\mathfrak{X}^2} \frac{\pi^{(g)}(y)\, p(y,z)}{\pi(y)\, p^{(g)}(y,z)} \right) \mathbb{E}\Big[\mathrm{e}^{n\int (g(y,z)-U^{(g)}(y))\, \nu_n^{(2)}(dy,dz)} \Big]
\end{aligned}$$

Fix a closed subset $F \subset \mathscr{M}^1(\mathfrak{X}\times\mathfrak{X})$, and $\nu^{(2)} \in F$. One can find a function g such that

$$\int (g(y,z) - U^{(g)}(y))\, \nu^{(2)}(dy,dz) \geq \sup_h \left(\int (h(y,z) - U^{(h)}(y))\, \nu^{(2)}(dy,dz) \right) - \varepsilon,$$

and by continuity, a neighborhood $B_{(\nu^{(2)},\eta)}$ of $\nu^{(2)}$ such that for $\mu^{(2)}$ in this neighborhood,

$$\int (g(y,z) - U^{(g)}(y))\, \mu^{(2)}(dy,dz) \geq \int (g(y,z) - U^{(g)}(y))\, \nu^{(2)}(dy,dz) - \varepsilon.$$

Then, by Chernov's inequality,

$$\begin{aligned}
\mathbb{P}[\nu_n^{(2)} \in B_{(\nu^{(2)},\eta)}] &\leq \mathbb{E}\Big[\mathrm{e}^{n\int (g(y,z)-U^{(g)}(y))\, \nu_n^{(2)}(dy,dz)} \Big]\ \mathrm{e}^{-n\left(\sup_h \left(\int (h(y,z)-U^{(h)}(y))\, \nu^{(2)}(dy,dz)\right)-2\varepsilon\right)} \\
&\leq C\, \mathrm{e}^{-n\left(\sup_h \left(\int (h(y,z)-U^{(h)}(y))\, \nu^{(2)}(dy,dz)\right)-2\varepsilon\right)}
\end{aligned}$$

with C independent from n. It follows that

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[\nu_n^{(2)} \in B_{(\nu^{(2)},\eta)}] &\leq -\sup_h \left(\int (h(y,z) - U^{(h)}(y))\, \nu^{(2)}(dy,dz) \right) + 2\varepsilon \\
&\leq -\inf_{\mu^{(2)}\in F} \sup_h \left(\int (h(y,z) - U^{(h)}(y))\, \mu^{(2)}(dy,dz) \right) + 2\varepsilon
\end{aligned}$$

and covering the compact set F by a finite number of balls, and then making ε go to zero, we end up with

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}[\nu_n^{(2)} \in F] \leq -\inf_{\mu^{(2)}\in F} \sup_h \left(\int (h(y,z) - U^{(h)}(y))\, \mu^{(2)}(dy,dz) \right).$$

Let us then identify the supremum over functions h. Fixing a measure μ with $\mu_1 = \mu_2$, one looks at

$$\begin{aligned}
\sum_{(y,z)\in\mathfrak{X}^2} (g(y,z) - U^{(g)}(y))\, \mu(y,z) &= \sum_{(y,z)\in\mathfrak{X}^2} \mu(y,z)\, \log \left(\frac{\mu_1(y)\, p^{(g)}(y,z)}{\mu_1(y)\, p(y,z)} \right) \\
&= H(\mu||\mu_1\otimes p) - H(\mu||\mu_1\otimes p^{(g)}) \leq H(\mu||\mu_1\otimes p)
\end{aligned}$$

with equality if and only if $\mu = \mu_1 \otimes p^{(g)}$, which corresponds to the choice

$$g(y,z) = \log\left(\frac{\mu(y,z)}{\mu_1(y)\,p(y,z)}\right).$$

Notice that in this case $U^{(g)} = 0$. This ends the proof for the upper bound.

For the lower bound, similar arguments reduces the problem to the proof of

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}[\nu_n^{(2)} \in B_{(\mu,\eta)}] \geq -H(\mu||\mu_1 \otimes p)$$

for any μ and η small enough. Denote $\mathbb{P}^{(g)}$ and $\mathbb{E}^{(g)}$ the probabilities and expectations corresponding to the modified Markov chain of parameter $g : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$. One has

$$\begin{aligned}
\mathbb{P}[\nu_n^{(2)} \in B_{(\mu,\eta)}] &= \sum_{x_1,\dots,x_n} \pi(x_1) \left(\prod_{i=2}^n p(x_{i-1},x_i)\right) \mathbb{1}_{\nu_x^{(2)} \in B_{(\mu,\eta)}} \\
&\geq C \sum_{x_1,\dots,x_n} \pi^{(g)}(x_1) \left(\prod_{i=2}^n p^{(g)}(x_{i-1},x_i)\right) \mathrm{e}^{-\left(\sum_{i=1}^n g(x_{i-1},x_i) - U^{(g)}(x_{i-1})\right)} \mathbb{1}_{\nu_x^{(2)} \in B_{(\mu,\eta)}} \\
&\geq C\, \mathbb{E}^{(g)}\left[\mathrm{e}^{-\left(\sum_{i=1}^n g(x_{i-1},x_i) - U^{(g)}(x_{i-1})\right)} \mathbb{1}_{\nu_x^{(2)} \in B_{(\mu,\eta)}}\right],
\end{aligned}$$

and one can choose η small enough so that within the ball $B_{(\mu,\eta)}$,

$$\frac{1}{n}\sum_{i=1}^n g(x_{i-1},x_i) - U^{(g)}(x_{i-1}) \leq \int (g(y,z) - U^{(g)}(y))\,\mu(dy,dz) + \varepsilon$$

for a fixed arbitrary $\varepsilon > 0$. One obtains then

$$\mathbb{P}[\nu_n^{(2)} \in B_{(\mu,\eta)}] \geq C\,\mathrm{e}^{-n\left(\int (g(y,z) - U^{(g)}(y))\,\mu(dy,dz) + \varepsilon\right)}\, \mathbb{P}^{(g)}[\nu_n^{(2)} \in B_{(\mu,\eta)}],$$

so it suffices to show the following result: for any μ with $\mu_1 = \mu_2$, if g is chosen to maximize the functional $\int (g(y,z) - U^{(g)}(y))\,\mu(dy,dz)$, then

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}^{(g)}[\nu_n^{(2)} \in B_{(\mu,\eta)}] \geq 0,$$

or, by taking the complementary event,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}^{(g)}[\nu_n^{(2)} \in (B_{(\mu,\eta)})^{\mathrm{c}}] < 0.$$

However, the upper bound part of our Theorem ensures that this limsup is smaller than

$$-\inf_{\substack{\nu \in (B_{(\mu,\eta)})^{\mathrm{c}} \\ \nu_1 = \nu_2}} H(\nu||\nu_1 \otimes p^{(g)}).$$

Therefore, by compactness and continuity of the functional $\nu \mapsto H(\nu||\nu_1 \otimes p^{(g)})$, it suffices to show that $H(\nu||\nu_1 \otimes p^{(g)}) > 0$ on the complementary of $B_{(\mu,\eta)}$, or, equivalently, that $H(\nu||\nu_1 \otimes p^{(g)}) = 0$ implies $\nu = \mu$ for a law with $\nu_1 = \nu_2$ and in $\mathscr{M}^1(\mathfrak{X} \times \mathfrak{X})$.

However, if $H(\nu||\nu_1 \otimes p^{(g)}) = 0$, then

$$\nu(x,y) = (\nu_1 \otimes p^{(g)})(x,y) = \nu_1(x)p^{(g)}(x,y) = \nu_1(x)\,\mathrm{e}^{g(x,y)}\,p(x,y) = \frac{\nu_1(x)}{\mu_1(x)}\,\mu(x,y).$$

Taking the sum over x, we conclude that

$$\nu_1(y) = \nu_2(y) = \sum_{x\in\mathfrak{X}} \frac{\nu_1(x)}{\mu_1(x)}\,\mu(x,y) = (\nu_1 \times q)(y)$$

where q is the stochastic matrix $q(x,y) = \frac{\mu(x,y)}{\mu_1(x)}$. However, this stochastic matrix has for invariant law μ_1, so $\nu_1 = \mu_1$ and $\nu = \mu$. This ends the proof of the lower bound. □

It turns out that Theorem 3.13 is stronger than the general form of Sanov's theorem 3.12: indeed, we shall see in the next chapter how to deduce from a large deviation principle on empirical pair measures a large deviation principle on empirical measures, by using the map

$$\begin{aligned} \mathscr{M}^1(\mathfrak{X}\times\mathfrak{X}) &\to \mathscr{M}^1(\mathfrak{X}) \\ \nu &\mapsto \nu_1 \quad \text{(marginal law)}; \end{aligned}$$

see Proposition 4.6 and the example given just after. On the other hand, the proofs of Theorems 3.3, 3.12 and 3.13 have exhibited the main characteristics of a large deviation principle:

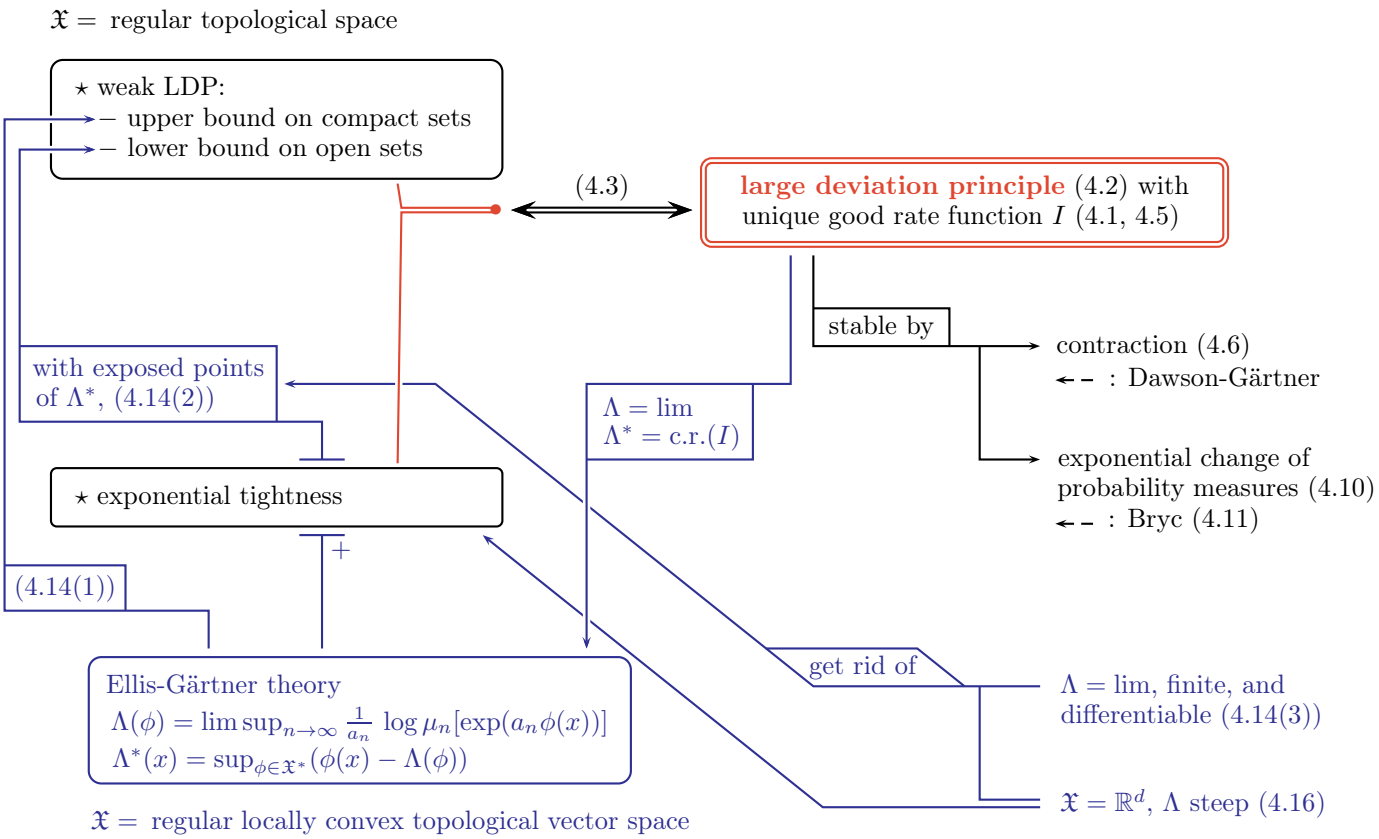
- For sequences of random variables that converge in probability, one expects the probabilities of rare events to decrease exponentially fast, with an upper bound on
$$\limsup_{n\to\infty} s_n \,\log \mathbb{P}[X_n \in F]$$
for F closed subset of the space of states, and lower bounds on
$$\liminf_{n\to\infty} s_n \,\log \mathbb{P}[X_n \in U]$$
for U open subset of the space of states.
- The actual rates of decay are given by a functional which most of the times can be computed by application of Chernov's inequality; the same inequality gives the upper bound (possibly after some topological work).
- The lower bound is then obtained by combining a method of exponential change of measures and the proof of the upper bound.

Chapter 4 is devoted to the task of finding the correct framework in order to generalize all these arguments.

CHAPTER 4

Principles of large deviations

This last Chapter is organized as follows. In Section 4.1, we define large deviation principles in the most general topological setting, and we detail how to use continuous maps or exponential changes of measures in order to translate these LDP from one space to another. In §4.2, we restrict ourselves to the case of locally convex topological vector spaces, and we give a fairly general theorem that contains both Cramér's and Sanov's results, and that is the main criterion in order to prove LDP in a functional setting. Finally, in §4.3, we apply this theorem to many situations: Cramér's large deviations in $\mathbb{R}^d$, refinements of Sanov's theorem, Poissonian approximations, *etc.* We conclude by an analysis of the Brownian paths in the setting of large deviations, thereby proving Schilder's theorem and stating the law of the iterated logarithm.



REMARK. The main difficulty here is probably to remember at each step the hypotheses underlying the results and the chains of implications (this is difficult even to the author). To simplify a bit the reading, we have summarized the chapter by a scheme. One studies a sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ on a general regular topological space $\mathfrak{X}$ (in black), and then more precisely on a regular and locally convex topological vector space $\mathfrak{X}$ (in blue).

4.1. Topological setting and transformations

4.1.1. Rate functions and large deviation principles. Though the theory of convergence of random variables has been developed in Chapters 1 and 2 on polish spaces, it is convenient here to start with a slightly more general setting, so in the following $\mathfrak{X}$ stand for a topological Hausdorff space (Hausdorff means that disjoint can be separated by open sets). We want to deal with sequences, or more generally families of random variables such that asymptotically, their laws are exponentially concentrated, in a fashion similar to Theorems 3.3, 3.12 and 3.13. Therefore, we fix a family $(\mu_\varepsilon)_{\varepsilon>0}$ in $\mathscr{M}^1(\mathfrak{X})$, its asymptotics being understood as the limiting behavior when ε goes to zero. One recovers the case of sequences $(X_n)_{n\in\mathbb{N}}$ of $\mathfrak{X}$-valued random variables by setting

$$\mu_\varepsilon = \text{law of } X_{\lfloor \frac{1}{\varepsilon} \rfloor}.$$

Then, one expects a large deviation principle to be a statement of the kind

$$-\inf_{u\in B^\circ} I(u) \leq \liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(B) \leq \limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(B) \leq -\inf_{u\in \overline{B}} I(u)$$

for any measurable set B (this is equivalent to the analogue statement with an upper bound for closed set and a lower bound for open sets). However, one might want to impose a few restrictions on the kind of functions I allowed in such a statement. Continuity is a little too strong requirement, as can be seen from Cramér's theorem in the case of bounded variables: in this case one has indeed $I = +\infty$ outside the convex hull of the support of the law.

DEFINITION 4.1. *A **rate function** $I : \mathfrak{X} \to [0, +\infty]$ is a lower semi-continuous function,* i.e.*, for all level $\alpha \geq 0$, the level set*

$$I_{\leq\alpha} = \{x \in \mathfrak{X},\ \ I(x) \leq \alpha\}$$

is a closed subset. A good rate function is a rate function whose level sets are compact for all $\alpha \in \mathbb{R}_+$.

EXAMPLE. Consider any non-negative, lower semi-continuous and strictly convex function on $\mathbb{R}$ (*i.e.*, the slopes are strictly monotone with respect to the two points between which they are computed), that attains its minimum on $\mathbb{R}$. Such a function is a always a good rate function, and more generally, a non-negative lower semi-continuous convex function is still a rate function. In particular, the Legendre-Fenchel transform of the cumulant generating function of a real random variable X is a rate function. A sufficient

criterion to obtain a good rate function in this setting is to require that 0 lies in the interior of $\mathcal{D}(\Lambda_X)$. Indeed, one has then $\mathbb{E}[\mathrm{e}^{tX}] < \infty$ for $t = \pm t_0$ with $t_0 > 0$; and so,

$$\Lambda_X^*(u) = \sup_{t\in\mathbb{R}}(ut - \Lambda_X(t)) \geq |u|\, t_0 - \max(\Lambda_X(t_0), \Lambda_X(-t_0)).$$

So, $\liminf_{|u|\to\infty} \frac{\Lambda_X^*(u)}{|u|} \geq t_0$, which clearly ensures the boundedness of every level set.

DEFINITION 4.2. *A family of laws $(\mu_\varepsilon)_{\varepsilon>0}$ on $\mathfrak{X}$ follows a* ***large deviation principle*** *(in short LDP) with rate function I if*

$$-\inf_{u\in B^\circ} I(u) \leq \liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(B) \leq \limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(B) \leq -\inf_{u\in\overline{B}} I(u)$$

for any Borel set $B \subset \mathfrak{X}$.

EXAMPLE. A way to restate Theorem 3.3 is: the mean of i.i.d. random variables such that $0 \in \mathcal{D}(\Lambda_X)^\circ$ satisfies a LDP with good rate function Λ_X^*.

EXAMPLE. Similarly, Sanov's theorem reads as follows: the empirical measure ν_n of an irreducible Markov chain with finite space of states $\mathfrak{X} = [\![1, N]\!]$ satisfies a LDP with good rate function

$$I(\nu) = \sup_{\mu\in\mathscr{M}^1(\mathfrak{X})} \left(\sum_{x\in\mathfrak{X}} \nu(x) \log \frac{\mu(x)}{(\mu\, p)(x)} \right). \tag{4.1}$$

Let us proof that I is indeed a good rate function. An important fact that will prove useful later is the non-obvious identity

$$I(\nu) = J(\nu) = \sup_{\lambda\in\mathbb{R}^N} \big(\langle \lambda \,|\, \nu \rangle - \log r(p(i,j)\, \mathrm{e}^{\lambda_j}) \big), \tag{4.2}$$

where $r(A)$ denotes as before the Frobenius eigenvalue of an irreducible non-negative matrix. We proceed by double inequality to prove it. In one direction, notice that the supremum in Equation (4.1) is attained on positive vectors. If μ is such a fixed vector in $\mathscr{M}^1(\mathfrak{X})$, set $\lambda_j = \log \frac{\mu_j}{(\mu\, p)_j}$, so that

$$p^\lambda(i,j) = p(i,j)\, \mathrm{e}^{\lambda_j} = \frac{\mu_j\, p(i,j)}{\sum_{i=1}^N \mu_i\, p(i,j)}.$$

Notice that $(\mu\, p^\lambda)_j = \mu_j$, so $\mu\, p^\lambda = \mu$ and $r(p^\lambda) = 1$ since μ is a positive vector. Therefore,

$$J(\nu) \geq \langle \lambda \,|\, \nu \rangle = \sum_{j=1}^N \nu_j \log \frac{\mu_j}{(\mu\, p)_j},$$

and since this is true for every μ, $J(\nu) \geq I(\nu)$. In the other direction, fix $\lambda \in \mathbb{R}^N$ and choose a Frobenius eigenvector μ for p^λ. One has

$$\begin{aligned} -\langle \lambda \,|\, \nu \rangle + \sum_{j=1}^N \nu_j \log \frac{\mu_j}{(\mu\, p)_j} &= \sum_{j=1}^N \nu_j \log \frac{\mu_j}{(\mu\, p^\lambda)_j} \\ &= -\sum_{j=1}^N \nu_j \log r(p^\lambda) \\ &= -\log r(p^\lambda), \end{aligned}$$

so $I(\nu) \geq \langle \lambda \,|\, \nu \rangle - \log r(p^\lambda)$. Since this is true for any λ, $I(\nu) \geq J(\nu)$ and (3.3) is proved. Now, I appears as the Legendre-Fenchel transform (in dimension N) of a positive differentiable function, so in particular it is lower semi-continuous. Since the space of states $\mathscr{M}^1(\mathfrak{X})$ is compact, the rate function I is automatically good.

In Definition 4.2, notice that $\mu_\varepsilon(\mathfrak{X}) = 1$ for all ε, so the infimum over $\mathfrak{X}$ of a rate function of a large deviation principle should always be 0. Sometimes, the speed of exponential convergence is not proportional to ε, but to another function of ε, for instance $\phi(\varepsilon) = \varepsilon^2$; one then still speaks of a large deviation principle, but with speed $\phi(\varepsilon)$. On the other hand, the goodness of the rate function ensures that in the upper bound, the infimum is actually a minimum, which sometimes eases the computation of this upper bound. Finally, the definition of a LDP reflects a uniform concentration of the family of measures, therefore, it is no surprise that it is closely related to the exponential equivalent of the notion of tightness studied in Chapter 2.

DEFINITION 4.3. *On a topological space $\mathfrak{X}$, a family of probability measures $(\mu_\varepsilon)_{\varepsilon>0}$ is said* ***exponentially tight*** *if, for every $\alpha < \infty$, there exists a compact set K_α such that*

$$\limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(K_\alpha^{\mathrm{c}}) \leq -\alpha.$$

For a countable family of probability measures $(\mu_n)_{n\in\mathbb{N}}$, this statement is replaced by

$$\limsup_{n \to \infty} \frac{1}{n} \log \mu_n(K_\alpha^{\mathrm{c}}) \leq -\alpha.$$

(1) If $(\mu_n)_{n\in\mathbb{N}}$ satisfies a LDP principle with a good rate function, then it is exponentially tight (notice the hypothesis of a countable family).

(2) Under the assumption of exponential tightness, if the upper bound

$$\limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(K) \leq -\inf_{u \in K} I(u)$$

is satisfied by any compact set K, then it is also satisfied by any closed set F.

(3) Under the same assumption, if the lower bound is satisfied by any open set, then I is a good rate function.

So, if the family satisfies a weak LDP, that is to say a LDP with the upper bound only for compact sets, then the exponential tightness ensures a full LDP and a good rate function.

For the proof we refer to [**Dembo and Zeitouni, 1998**, Lemma 1.2.18 and Exercise 4.1.10].

Actually, the notion of exponential tightness plays with respect to large deviation principles a role extremely similar to the role of tightness w.r.t. weak convergence. Let us just state without proof the following theorem, for which we refer to [**Feng and Kurtz, 2006**, Theorem 3.7].

THEOREM 4.4 (Puhalskii). *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of probability measures on a polish space that is exponentially tight. There is a good rate function I and a subsequence $(\mu_{\phi(n)})_{n\in\mathbb{N}}$ such that a LDP with rate function I holds for the subsequence (with asymptotic speed $\phi(n)$).*

In the remaining of this section, we shall discuss a few details around the definition of a LDP, and we shall give another equivalent "integral" formulation, which is in fact one of the main motivation outside the computation of the speed of convergence; see Theorem 4.10 and 4.11.

4.1.2. Contraction principle and a proof of Sanov's theorem. We start with the following precision of Definition 4.2.

PROPOSITION 4.5. *A rate function I for a large deviation principle is unique.*

PROOF. For convenience we write the proof in a metric space, but it holds in any regular topological space (see the definition at the beginning of §4.1.3). Consider a family $(\mu_\varepsilon)_{\varepsilon>0}$ of probability measures that satifies a LDP with respect to two different rate functions I_1 and I_2. Since I_1 is lower semi-continuous,

$$\lim_{\delta\to 0}\inf_{y\in\overline{B}_{(x,\delta)}} I_1(y) \geq I_1(x).$$

By definition of a large deviation principle,

$$\begin{aligned} -\inf_{y\in\overline{B}_{(x,\delta)}} I_1(y) &\geq \limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(\overline{B}_{(x,\delta)}) \\ &\geq \liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(B_{(x,\delta)}) \\ &\geq -\inf_{y\in B_{(x,\delta)}} I_2(x) \geq -I_2(x). \end{aligned}$$

By taking the limit as δ goes to 0, $-I_1(x) \geq -I_2(x)$, *i.e*, $I_1 \leq I_2$. By symmetry of the roles played by I_1 and I_2, $I_1 = I_2$. □

Another important result is the following ***contraction principle***:

PROPOSITION 4.6 (Contraction principle). *Let $\mathfrak{X}$ and $\mathfrak{W}$ be two topological spaces, $f : \mathfrak{X} \to \mathfrak{W}$ a continuous function. If $(\mu_\varepsilon)_{\varepsilon>0}$ is a family of probability measures on $\mathfrak{X}$ that satisfies a LDP with good rate function I, then $(f_\star\mu_\varepsilon)_{\varepsilon>0}$ satisfies on $\mathfrak{W}$ a LDP with good rate function*

$$f_\star I(w) = \inf\{I(x),\ f(x) = w\}.$$

PROOF. Since $f_\star I$ is defined by taking lower bounds of sets of non-negative real numbers, it is non-negative on $\mathfrak{W}$. On the other hand, by goodness of I, the infimum is always attained in the definition of $f_\star I$, because one takes the infimum on a closed set. Take $\alpha \geq 0$; the level set $(f_\star I)_{\leq \alpha} = \{w,\ f_\star I(w) \leq \alpha\}$ is included in the image by f of the level set $I_{\leq \alpha}$. Indeed, if $f_\star I(w) \leq \alpha$, then there exists x such that $I(x) \leq \alpha$ and $f(x) = w$ (the point of $f^{-1}(\{w\})$ at which the infimum is attained), which exactly means that $w \in f(I_{\leq \alpha})$. The converse inclusion is obvious, so

$$(f_\star I)_{\leq \alpha} = f(I_{\leq \alpha}).$$

Since I is a good rate function, $I_{\leq \alpha}$ is compact, and so is its image by any continuous function, so we have shown that $f_\star I$ was a good rate function.

Take a closed set $F \subset \mathfrak{W}$. One can write

$$\limsup_{\varepsilon \to 0} \varepsilon \log f_\star \mu_\varepsilon(F) = \limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(f^{-1}(F)) \leq - \inf_{u \in f^{-1}(F)} I(u) = - \inf_{w \in F} f_\star I(w)$$

since $f^{-1}(F)$ is closed as the reciprocal image of a closed set by a continuous function. The same argument works for open sets, so the LDP for the image laws is demonstrated. □

Surprisingly, there is a way to go in the opposite direction, namely, if all the images of $(\mu_\varepsilon)_{\varepsilon>0}$ by a "sufficiently large" family $\mathscr{F}$ of continuous maps satisfy a LDP with good rate functions, then the family $(\mu_\varepsilon)_{\varepsilon>0}$ also satisfy a LDP with a good rate function, but w.r.t. the weak topology determined by $\mathscr{F}$ — this is the Dawson-Gärtner theorem, see [**Dembo and Zeitouni, 1998**, Theorem 4.6.1]. Most of the discussion of §4.2 is devoted to a particular case of this principle.

COROLLARY 4.7. *Theorem 3.12 with $\pi_0 = \pi$ (the stationary measure) holds in the general setting of an irreducible Markov chain on a finite set.*

To prove Corollary 4.7, we shall use classical arguments of calculus of variations, which we recall here. Suppose that f and g are real-valued continuously differentiable functions on $\mathbb{R}^N$, and consider the following problem: we want to find the local extremas of $f(x_1, \ldots, x_N)$ under the constraint $g(x_1, \ldots, x_N) = c$, where c is some constant. By introducing the Lagrange function

$$L(\lambda, x_1, \ldots, x_N) = f(x_1, \ldots, x_N) - \lambda(g(x_1, \ldots, x_N) - c)$$

where λ is an additional variable, this problem is translated into the equation $dL = 0$, so a constraint extrema $x_{1,0}, \ldots, x_{N,0}$ satisfies the system of $N+1$ ***Euler-Lagrange equations***

$$\begin{cases} g(x_{1,0}, \ldots, x_{N,0}) = c; \\ df_{(x_{1,0}, \ldots, x_{N,0})} = \lambda_0 \, dg_{(x_{1,0}, \ldots, x_{N,0})}. \end{cases}$$

More generally, if one wants to find the local extremas of $f(x_1, \ldots, x_N)$ under r constraints $g_1(x_1, \ldots, x_N) = c_1$, $g_1(x_2, \ldots, x_N) = c_2$, *etc.*, then one has to solve the system of $N + r$

equations

$$\begin{cases} g_1(x_{1,0},\dots,x_{N,0}) = c_1; \\ \quad\vdots \\ g_r(x_{1,0},\dots,x_{N,0}) = c_r; \\ df_{(x_{1,0},\dots,x_{N,0})} = \lambda_{1,0}\, dg_{1,(x_{1,0},\dots,x_{N,0})} + \cdots + \lambda_{r,0}\, dg_{r,(x_{1,0},\dots,x_{N,0})}. \end{cases}$$

As an application of this principle, we have:

LEMMA 4.8. *For $\nu \in \mathscr{M}^1(\mathfrak{X})$ with positive coordinates, let μ be the maximizer of the function*

$$F(\mu) = \sum_{x\in\mathfrak{X}} \nu(x)\, \log\left(\frac{\mu(x)}{(\mu p)(x)}\right),$$

so that $I(\nu) = F(\mu)$. If

$$\nu^{(2)} = \frac{\mu(x)\, p(x,y)\, \nu(y)}{(\mu p)(y)},$$

then the two marginales of $\nu^{(2)}$ are equal to ν, and $F(\mu) = J(\nu^{(2)})$ with the notations of Theorem 3.13.

PROOF. The constraint of maximization of F is $\sum_{x\in\mathfrak{X}} \mu(x) = 1$, so Euler-Lagrange equation reads as

$$dF_\mu = \sum_{x\in\mathfrak{X}} \left(\frac{\nu(x)}{\mu(x)} - \sum_{y\in\mathfrak{X}} \frac{p(x,y)\nu(y)}{(\mu p)(y)}\right) d\mu_x = \lambda \sum_{x\in\mathfrak{X}} d\mu_x.$$

This means that there is a common constant λ such that for all $x \in \mathfrak{X}$,

$$\lambda = \frac{\nu(x)}{\mu(x)} - \sum_{y\in\mathfrak{X}} \frac{p(x,y)\, \nu(y)}{(\mu p)(y)}.$$

Multiplying by $\mu(x)$ and taking the sum over $x \in \mathfrak{X}$, we get

$$\lambda = \lambda \sum_{x\in\mathfrak{X}} \mu(x) = \sum_{x\in\mathfrak{X}} \nu(x) - \sum_{(x,y)\in\mathfrak{X}^2} \frac{\mu(x)\, p(x,y)\, \nu(y)}{(\mu p)(y)} = 1 - 1 = 0.$$

So, μ is solution of the equations $\nu(x) = \sum_{y\in\mathfrak{X}} \frac{\mu(x)\, p(x,y)\, \nu(y)}{(\mu p)(y)}$ for every x. Equivalently, the first marginale of $\nu^{(2)}$ is ν, and the same is true for the second marginale:

$$\sum_{x\in\mathfrak{X}} \nu^{(2)}(x,y) = \sum_{x\in\mathfrak{X}} \frac{\mu(x)\, p(x,y)\, \nu(y)}{(\mu p)(y)} = \frac{(\mu p)(y)\, \nu(y)}{(\mu p)(y)} = \nu(y).$$

We can then rewrite:

$$I(\nu) = F(\mu) = \sum_x \nu(x)\, \log\left(\frac{\mu(x)}{(\mu p)(x)}\right) = \sum_{x,y} \nu^{(2)}(x,y)\, \log\left(\frac{\mu(x)}{(\mu p)(x)}\right).$$

Notice that

$$\log\left(\frac{\mu(x)}{(\mu p)(x)}\right) = \log\left(\frac{\nu^{(2)}(x,y)}{\nu(x)\,p(x,y)}\right) + \log\left(\frac{\nu(x)\,(\mu p)(y)}{\nu(y)\,(\mu p)(x)}\right).$$

Since $\nu^{(2)}(x,y)$ has same x and y marginales, the expectation of the second term under $\nu^{(2)}(x,y)$ is zero. Therefore,

$$F(\mu) = \sum_{x,y} \nu^{(2)}(x,y)\,\log\left(\frac{\nu^{(2)}(x,y)}{(\nu\otimes p)(x)}\right) = H(\nu^{(2)}||\nu\otimes p) = J(\nu^{(2)}). \qquad \square$$

Similarly, the Lagrange principle leads to:

LEMMA 4.9. *For $\nu \in \mathscr{M}^{(1)}(\mathfrak{X})$ with positive coordinates, let $\nu^{(2)}$ be the pair measure that minimizes $J(\nu^{(2)})$ under the constraints $\nu_1^{(2)} = \nu_2^{(2)} = \nu$. There exists two positive functions γ and $\widetilde{\gamma}$ on $\mathfrak{X}$ such that*

$$\nu^{(2)}(x,y) = \nu(x)\,\gamma(x)\,p(x,y)\,\widetilde{\gamma}(y).$$

One can suppose without loss of generality that $\nu\cdot\gamma = \nu(x)\gamma(x)$ is a probability measure on $\mathfrak{X}$. Then, $J(\nu^{(2)}) = F(\nu\cdot\gamma)$ with the notations of the previous lemma.

PROOF. The $2N$ constraints are $\sum_{y\in\mathfrak{X}}\nu^{(2)}(x,y) = \sum_{y\in\mathfrak{X}}\nu^{(2)}(y,x) = \nu(x)$ for all $x\in\mathfrak{X}$. Therefore, there exists $2N$ constants $\alpha_x, \widetilde{\alpha}_y$ such that if

$$G(\nu^{(2)}) = \sum_{x,y}\nu^{(2)}(x,y)\,\log\left(\frac{\nu^{(2)}(x,y)}{\nu(x)\,p(x,y)}\right),$$

then $dG_{\nu^{(2)}} = \sum_{x,y}(\alpha_x+\widetilde{\alpha}_y)\,d\nu^{(2)}_{x,y}$ at the constraint minimizer. However,

$$dG_{\nu^{(2)}} = \sum_{x,y}\left(1+\log\left(\frac{\nu^{(2)}(x,y)}{\nu(x)\,p(x,y)}\right)\right)\,d\nu^{(2)}_{x,y}.$$

Setting $\gamma(x) = \mathrm{e}^{\alpha_x - 1/2}$ and $\widetilde{\gamma}(y) = \mathrm{e}^{\widetilde{\alpha}_y - 1/2}$, we obtain the first part of the lemma. One can force $\nu\cdot\gamma$ to be a probability measure by replacing γ by $\frac{\gamma}{|\nu\cdot\gamma|}$ and $\widetilde{\gamma}$ by $|\nu\cdot\gamma|\,\widetilde{\gamma}$; this does not change the equation

$$\nu^{(2)}(x,y) = \nu(x)\,\gamma(x)\,p(x,y)\,\widetilde{\gamma}(y).$$

Taking marginales, we get on the one hand $\sum_{y\in\mathfrak{X}}\gamma(x)\,p(x,y)\,\widetilde{\gamma}(y) = 1$, and on the other hand $\sum_{x\in\mathfrak{X}}\nu(x)\,\gamma(x)\,p(x,y)\,\widetilde{\gamma}(y) = \nu(y)$. This means that $p^{\gamma}(x,y) = \gamma(x)\,p(x,y)\,\widetilde{\gamma}(y)$ is

the transition matrix of an irreducible Markov chain with Frobenius eigenvector ν. Then,

$$\begin{aligned} J(\nu^{(2)}) &= \sum_{x,y} \nu(x)\, p^{\gamma}(x,y)\, \log(\gamma(x)\,\widetilde{\gamma}(y)) \\ &= \sum_{x} \nu(x) \log(\gamma(x)) + \sum_{x,y} \nu(x)\, p^{\gamma}(x,y) \log(\widetilde{\gamma}(y)) \\ &= \sum_{x} \nu(x) \log(\gamma(x)) + \sum_{y} \nu(y)\, \log(\widetilde{\gamma}(y)) \\ &= \sum_{x} \nu(x)\, \log(\gamma(x)\,\widetilde{\gamma}(x)). \end{aligned}$$

Finally, $\widetilde{\gamma}(x) = \frac{\nu(x)}{((\nu\cdot\gamma)p)(x)}$, which ends the proof. Then, it can be checked that one can get rid of the hypothesis of positive coordinates for ν. □

Proof of Corollary 4.7. One uses Theorem 3.13 and the contraction principle 4.6. Recall that the empirical pair measure $\nu_n^{(2)}$ has for first marginale law the empirical measure ν_n; and obviously this projection $\mathscr{M}^1(\mathfrak{X}\times\mathfrak{X}) \to \mathscr{M}^1(\mathfrak{X})$ is continuous. The rate function of the LDP satisfied by the empirical pair measure is indeed a good rate function: it is lower semi-continuous as the composition of the functions

$$\nu \mapsto (\nu, \nu_1) \mapsto (\nu, \nu_1 \otimes p) \mapsto H(\nu || \nu_1 \otimes p),$$

and automatically good since $\mathscr{M}^1(\mathfrak{X}\times\mathfrak{X})$ is compact. Therefore, ν_n satisfies a LDP with good rate function

$$T(\nu) = \inf_{\nu_1^{(2)} = \nu_2^{(2)} = \nu} \{H(\nu^{(2)} || \nu \otimes p)\}.$$

However, μ in Lemma 4.8 and $\nu\cdot\gamma$ in Lemma 4.9 satisfy the same equations, so they are equal and

$$T(\nu) = F(\nu\cdot\gamma) = F(\mu) = I(\nu),$$

which ends the proof. It can then be checked that the assumption $\pi_0 = \pi$ (one starts from the stationary law) was only there to ease certain computations on the empirical pair measures, so Theorem 3.12 holds in full generality. □

4.1.3. Varadhan's lemma. Fix a topological space $\mathfrak{X}$ and a family of probability measures $(\mu_\varepsilon)_{\varepsilon>0}$ that satisfies a large deviation principle with good rate function I. For the results of this paragraph to hold, we have to suppose that a point and a disjoint closed subset of $\mathfrak{X}$ can be separated by open subsets; thus, $\mathfrak{X}$ is a ***regular space***, which is a little more than Hausdorff. This hypothesis will be used at the beginning of the proof of Bryc's theorem.

If I attains its minimum at exactly one point x_0, then $\mu_\varepsilon \rightharpoonup \delta_{x_0}$. Indeed, if $(X_\varepsilon)_{\varepsilon>0}$ is a family of random variables representatives of the laws μ_ε, then

$$\mathbb{P}[d(X_\varepsilon, x_0) \geq \eta] = \mu_\varepsilon((B_{(x_0,\eta)})^{\mathrm{c}}) \leq C_{\eta,\delta}\ \exp\left(-\frac{\inf_{u\notin B_{(x_0,\eta)}} I(u) - \delta}{\varepsilon}\right) \quad \text{for any } \delta > 0$$
$$\leq C_{\eta,\delta}\ \exp\left(-\frac{K_\eta}{\varepsilon}\right) \to_{\varepsilon\to 0} 0.$$

In general, one stills expect the laws μ_ε to be concentrated around the points where I is maximal. ***Varadhan's lemma*** precises this statement; it is also a generalization of Laplace's method for functions on the real line.

THEOREM 4.10 (Varadhan). *Fix a continuous function ϕ on $\mathfrak{X}$. One assumes the tail condition*

$$\lim_{M\to\infty} \limsup_{\varepsilon\to 0} \left(\varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\, \mathbf{1}_{\phi(X)\geq M}\right]\right) = -\infty.$$

Then,

$$\lim_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\right] = \sup_{x\in\mathfrak{X}} (\phi(x) - I(x)).$$

REMARK. The tail condition is verified if, for instance, there exists $\gamma > 1$ such that

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\gamma\phi(X)}{\varepsilon}}\right] < +\infty.$$

Indeed, for $M > 0$, consider the random variable $Y_\varepsilon = \exp(\frac{\phi(X_\varepsilon)-M}{\varepsilon})$, where X_ε follows the law μ_ε. Notice that $Y_\varepsilon \geq 1$ if and only if $\phi(X_\varepsilon) \geq M$. Then,

$$\mathrm{e}^{-\frac{M}{\varepsilon}}\, \mathbb{E}\left[\mathrm{e}^{\frac{\phi(X_\varepsilon)}{\varepsilon}}\, \mathbf{1}_{\phi(X_\varepsilon)\geq M}\right] = \mathbb{E}[Y_\varepsilon\, \mathbf{1}_{Y_\varepsilon\geq 1}] \leq \mathbb{E}[(Y^\varepsilon)^\gamma] = \mathrm{e}^{-\frac{\gamma M}{\varepsilon}}\, \mathbb{E}\left[\mathrm{e}^{\frac{\gamma\,\phi(X_\varepsilon)}{\varepsilon}}\right],$$

and therefore,

$$\limsup_{\varepsilon\to 0} \left(\varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\, \mathbf{1}_{\phi(X)\geq M}\right]\right) \leq -(\gamma-1)M + \limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\gamma\phi(X)}{\varepsilon}}\right],$$

which goes to $-\infty$ as the bound M goes to infinity.

PROOF OF THEOREM 4.10. Fix $x \in \mathfrak{X}$ and $\delta > 0$. Since ϕ is continuous there is an open neighborhood U of x such that $\inf_{u\in U} \phi(u) \geq \phi(x) - \delta$. Then,

$$\liminf_{\varepsilon\to 0} \varepsilon \log \mathbb{E}\left[\mathrm{e}^{\frac{\phi(X_\varepsilon)}{\varepsilon}}\right] \geq \liminf_{\varepsilon\to 0} \varepsilon \log \mathbb{E}\left[\mathrm{e}^{\frac{\phi(X_\varepsilon)}{\varepsilon}}\, \mathbf{1}_U(X_\varepsilon)\right] \geq \inf_{u\in U} \phi(u) + \liminf_{\varepsilon} \varepsilon \log \mu_\varepsilon[U]$$
$$\geq \inf_{u\in U} \phi(u) - \inf_{u\in U} I(u) \geq \phi(x) - I(x) - \delta.$$

Since this is true for all δ one can replace the right-hand side by $\phi(x) - I(x)$, and since this is true for all $x \in \mathfrak{X}$, one can even take $\sup_{x\in\mathfrak{X}}(\phi(x) - I(x))$.

Conversely, suppose first ϕ bounded from above by some constant M. Since I is a good rate function, for any level α, $I_{\leq\alpha}$ is a compact set and can therefore be covered by a finite number of open sets $B_{(x_i,\varepsilon)}$ such that

$$\left(\inf_{u\in\overline{B}_{(x_i,\varepsilon)}} I(u)\right) \geq I(x_i)-\delta \qquad ; \qquad \left(\sup_{u\in\overline{B}_{(x_i,\varepsilon)}} \phi(u)\right) \leq \phi(x_i)+\delta$$

where $\delta>0$ is fixed and arbitrary small. Then,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{e}^{\frac{\phi(X_\varepsilon)}{\varepsilon}}\right] &\leq \sum_{i=1}^{r} \mathbb{E}\left[\mathrm{e}^{\frac{\phi(X_\varepsilon)}{\varepsilon}} \mathbf{1}_{X_\varepsilon\in B_{(x_i,\varepsilon)}}\right] + \mathrm{e}^{\frac{M}{\varepsilon}}\,\mu_\varepsilon((I_{\leq\alpha})^{\mathrm{c}}) \\
&\leq \sum_{i=1}^{r} \mathrm{e}^{\frac{\phi(x_i)+\delta}{\varepsilon}} \mu_\varepsilon(\overline{B}_{(x_i,\varepsilon)}) + \mathrm{e}^{\frac{M}{\varepsilon}}\,\mu_\varepsilon((I_{\leq\alpha})^{\mathrm{c}}) \\
&\leq \left(\sum_{i=1}^{r} C_i\,\mathrm{e}^{\frac{\phi(x_i)-I(x_i)+2\delta}{\varepsilon}}\right) + C\,\mathrm{e}^{\frac{M-\alpha}{\varepsilon}},
\end{aligned}$$

so,

$$\begin{aligned}
\limsup_{\varepsilon\to 0} \varepsilon \log \mathbb{E}\left[\mathrm{e}^{\frac{\phi(X_\varepsilon)}{\varepsilon}}\right] &\leq \max\left(\max\left\{\phi(x_i)-I(x_i)+2\delta,\ i\in[\![1,r]\!]\right\}, M-\alpha\right) \\
&\leq \max\left(\sup_{x\in\mathfrak{X}}(\phi(x)-I(x)+2\delta), M-\alpha\right).
\end{aligned}$$

Since this is true for all δ (with α fixed), a correct bound is then

$$\max\left(\sup_{x\in\mathfrak{X}}(\phi(x)-I(x)), M-\alpha\right),$$

and making α going to infinity,

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mathbb{E}\left[\mathrm{e}^{\frac{\phi(X_\varepsilon)}{\varepsilon}}\right] \leq \sup_{x\in\mathfrak{X}}(\phi(x)-I(x)),$$

which ends the proof in the bounded case. The tail condition allows precisely to get rid of this hypothesis. □

Remark. A statement equivalent to Varadhan's lemma 4.10 is the following result on exponential changes of probability measures and LDP. Suppose that $(\mu_\varepsilon)_{\varepsilon>0}$ follows a LDP with good rate function I, and consider a bounded continuous function ϕ. Then, the family of probability measures

$$d\nu_\varepsilon^\phi = \frac{\mathrm{e}^{\frac{\phi(x)}{\varepsilon}}}{\mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\right]}\,d\mu_\varepsilon$$

satisfies an LDP with good rate function

$$I^\phi(x) = I(x)-\phi(x)-\inf_{y\in\mathfrak{X}}(I(y)-\phi(y)).$$

Indeed, by modifying just a little the proof of Varadhan's theorem, one sees that

$$\begin{aligned} -\inf_{u\in\overline{B}}(I(u)-\phi(u)) &\geq \limsup_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\,\mathbf{1}_{X\in B}\right] \\ &\geq \liminf_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\,\mathbf{1}_{X\in B}\right] \\ &\geq -\inf_{u\in B^{\mathrm{o}}}(I(u)-\phi(u)) \end{aligned}$$

and the normalization constant $\mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\right]$ of the probability measure is obtained by looking at the case $B=\mathfrak{X}$.

From the previous remark, we see that large deviation principles are conserved by exponential changes of measures w.r.t. bounded continuous functions. Moreover, Varadhan's lemma yields in each case the asymptotics of the normalization constants, also known as ***partition functions***; these quantities play an important role in statistical mechanics. The following theorem, due to Bryc, proves the converse: if the renormalized partition functions $\varepsilon\log\mu_\varepsilon\left[\mathrm{e}^{\phi(X)/\varepsilon}\right]$ all have a limit when ε goes to infinity and ϕ is a bounded continuous function on $\mathfrak{X}$, then one has a LDP.

THEOREM 4.11 (Bryc). *For $\phi\in\mathscr{C}_b(\mathfrak{X})$, denote*

$$\Lambda_\phi=\lim_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\right],$$

provided it exists. If $(\mu_\varepsilon)_{\varepsilon>0}$ is exponentially tight and if Λ_ϕ exists for all $\phi\in\mathscr{C}_b(\mathfrak{X})$, then the family satisfies a LDP with good rate function

$$I(x)=\sup_{\phi\in\mathscr{C}_b(\mathfrak{X})}(\phi(x)-\Lambda_\phi).$$

PROOF. For the lower bound, fix an open set $U\subset\mathfrak{X}$ and $u\in U$. The topological assumptions on $\mathfrak{X}$ ensure that one can construct a continuous function such that ϕ takes its values in $[0,1]$, $\phi(x)=1$ and $\phi(u)=0$ outside U (see [**Lang, 1993**]). Set then $f_m(u)=m(\phi(u)-1)$; this is a continuous bounded non-positive function. Since

$$\mu_\varepsilon\left[\mathrm{e}^{\frac{f_m(u)}{\varepsilon}}\right]\leq\mathrm{e}^{-\frac{m}{\varepsilon}}\,\mu_\varepsilon(U^{\mathrm{c}})+\mu_\varepsilon(U)\leq\mathrm{e}^{-\frac{m}{\varepsilon}}+\mu_\varepsilon(U),$$

one has

$$\max\left(\liminf_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon(U),-m\right)\geq\Lambda_{f_m}=-(f_m(x)-\Lambda_{f_m})\geq-\sup_{f\in\mathscr{C}_b(\mathfrak{X})}(f(x)-\Lambda_f)=-I(x).$$

Making m go to infinity, and then taking the supremum over U, one concludes that

$$\liminf_{\varepsilon\to 0}\varepsilon\log\mu_\varepsilon(U)\geq-\inf_{u\in U}I(u).$$

For the upper bound, fix $\eta>0$, and $x\in\mathfrak{X}$. Denote $I^\eta=\min(I-\eta,\frac{1}{\eta})$. There is a function $\phi\in\mathscr{C}_b(\mathfrak{X})$ such that $\phi(x)-\Lambda_\phi\geq I^\eta(x)$ — the transformation $I\to I^\eta$ is meant

to take care of the cases $I(x) < +\infty$ and $I(x) = +\infty$ simultaneously. By continuity of ϕ, one can then choose a neighborhood A_x of x such that

$$\forall y \in A_x, \;\; \phi(y) - \phi(x) \geq -\eta.$$

By Chebyshev's inequality,

$$\mu_\varepsilon[A_x] \leq \mathbb{E}[\mathrm{e}^{\frac{\phi(X_\varepsilon)-\phi(x)}{\varepsilon}}] \exp\left(-\inf_{y\in A_x} \phi(y) - \phi(x)\right)$$
$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon[A_x] \leq \eta - \phi(x) + \Lambda(\phi) \leq \eta - I^\eta(x).$$

By taking finite open coverings, one deduces from this that for any (relatively) compact set K,

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon[K] \leq \eta - \inf_{x\in K} I^\eta(x),$$

and since this is true for any η, $\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon[K] \leq -\inf_{x\in K} I(x)$. Thus, one has a weak LDP, and by the assumption of exponential tightness, a full LDP with I a good rate function, *cf.* Definition 4.3. □

Bryc's theorem is already a powerful criterion to get a large deviation principle in a general polish space, but unfortunately, the space of test functions $\mathscr{C}_b(\mathfrak{X})$ can be absolutely enormous, which makes the computation of the rate function quite difficult. The Ellis-Gärtner theorem, studied in the next section, reduces a lot the size of the space of test functions against which the asymptotics of Laplace transform have to be computed.

4.2. Ellis-Gärtner theorem

To get general principle of large deviations, we need to add a few topological assumptions on the space $\mathfrak{X}$ supporting the probability measures $(\mu_\varepsilon)_{\varepsilon>0}$. Thus, in the following, we assume that $\mathfrak{X}$ is a ***topological vector space that is regular and locally convex***, *i.e.*,

(i) $\mathfrak{X}$ is a vector space and a regular topological space, such that the vector space laws (addition, multiplication by a scalar) are continuous with respect to the topology;

(ii) every point of $\mathfrak{X}$ has a basis of neighborhoods that are convex subsets of $\mathfrak{X}$.

Recall that a convex subset $C \subset \mathfrak{X}$ is a subset with

$$\forall a, b \in C, \;\; \forall \lambda \in (0,1), \;\; \lambda a + (1-\lambda) b \in C;$$

the general theory of convex bodies is detailed in [**Schneider, 1993**, **Hörmander, 1994**]. Every normed vector space satisfies the previous hypotheses, and $\mathfrak{X} = \mathscr{C}(\mathbb{R}_+)$ also satisfies these hypotheses. The vector space of continuous linear functionals $\Phi : \mathfrak{X} \to \mathbb{R}$ will be denoted $\mathfrak{X}^*$.

4.2.1. Legendre-Fenchel transforms in infinite-dimensional spaces. The dual space $\mathfrak{X}^*$ appears in the generalization of the Legendre-Fenchel transform, defined as follows: if $\Lambda : \mathfrak{X}^* \to (-\infty, +\infty]$ is a function not equal everywhere to $+\infty$, then its Legendre-Fenchel transform is

$$\begin{aligned} \Lambda^* : \mathfrak{X} &\to (-\infty, +\infty] \\ x &\mapsto \sup_{\phi \in \mathfrak{X}^*} (\phi(x) - \Lambda(\phi)). \end{aligned}$$

When $\mathfrak{X} = \mathbb{R}$ this definition corresponds clearly to the one given in §3.1.

THEOREM 4.12. *Let $(\mu_\varepsilon)_{\varepsilon>0}$ be a family of probability laws on $\mathfrak{X}$, and denote*

$$\Lambda(\phi) = \limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}} \right]$$

for $\phi \in \mathfrak{X}^*$. The function Λ is convex, and its Legendre-Fenchel transform is a convex rate function. Moreover, for all compact subsets $K \subset \mathfrak{X}$,

$$\limsup_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(K) \leq - \inf_{u \in K} \Lambda^*(u).$$

PROOF. The convexity follows from the convexity of each function $\phi \mapsto \varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}} \right]$, which can be proved as follows. Take two linear forms ϕ and ψ, and consider the map

$$\begin{aligned} f : [0,1] &\to \mathbb{R} \\ t &\mapsto \varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{t\phi(X) + (1-t)\psi(X)}{\varepsilon}} \right]. \end{aligned}$$

If $f(0)$ and $f(1)$ are finite, then for any t, by Hölder's inequality,

$$f(t) \leq \varepsilon \log \left(\mu_\varepsilon \left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}} \right]^t \mu_\varepsilon \left[\mathrm{e}^{\frac{\psi(X)}{\varepsilon}} \right]^{1-t} \right) = t\, f(0) + (1-t) f(1).$$

This is also obviously true if $f(0) = +\infty$ or if $f(1) = +\infty$, so the convexity is shown. Next, as a supremum of affine continuous functions, a Legendre-Fenchel transform is always convex and semi-continuous, and by looking at $\phi = 0$, one sees that Λ^* is always non-negative, whence a rate function. Finally, the upper bound is proved exactly the same way as the upper bound in Bryc's theorem 4.11. □

As a consequence of this theorem, one has a good candidate for the rate function leading a large deviation principle: the Legendre-Fenchel transform of the limit of the quantities $\varepsilon \log \mu_\varepsilon[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}]$. The next result shows that if a large deviation principle with good rate function I, then $I = \Lambda^*$ in many situations.

PROPOSITION 4.13. *Suppose that $(\mu_\varepsilon)_{\varepsilon>0}$ satisfies a LDP with good rate function I, and that $\Lambda(\phi) < +\infty$ for all $\phi \in \mathfrak{X}^*$. Then, $\Lambda(\phi)$ is in fact a limit,*

$$\Lambda(\phi) = \sup_{x \in \mathfrak{X}} (\phi(x) - I(x)),$$

and Λ^* is the largest convex rate function smaller than I (in particular, $I = \Lambda^*$ if I is convex).

PROOF. We admit the following result, which is the infinite-dimensional generalization of the Legendre-Fenchel duality of Definition 3.2: if f be a lower semi-continuous function on $\mathfrak{X}$, and if

$$g(\phi) = \sup_{x\in\mathfrak{X}} (\phi(x) - f(x)),$$

then the Legendre-Fenchel transform of g is the largest convex function smaller than f (in particular $f = g^*$ if f is supposed convex). The proof is analog to the one given for functions on $\mathbb{R}$, only one has to use Hahn-Banach theorem to construct certain affine functions.

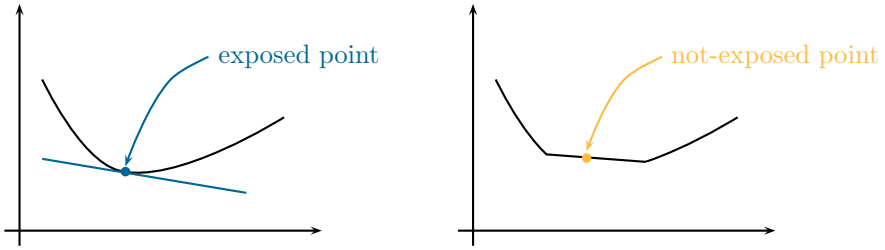
If the assumptions of the proposition are verified, then for any $\phi \in \mathfrak{X}^*$, $\Lambda(\gamma\phi) < \infty$ for some $\gamma > 1$, which allows to apply Varadhan's theorem 4.10:

$$\Lambda(\phi) = \lim_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\right] = \sup_{x\in\mathfrak{X}} (\phi(x) - I(x)).$$

Legendre-Fenchel duality ensures then that Λ^* is the convex regularization of I. □

4.2.2. Exposed points of Λ^* and the general criterion. We still need sufficient conditions for Λ^* governing a full LDP. This is what is provided by Ellis-Gärtner theorem, which consists mainly on topological assumptions on Λ. Call ***exposed point*** x of Λ^* a point such that there exists an "exposing hyperplane" $\phi \in \mathfrak{X}^*$ with

$$\phi(x) - \Lambda^*(x) > \phi(z) - \Lambda^*(z) \quad \text{for all } z \neq x.$$



THEOREM 4.14 (Ellis-Gärtner, Baldi). *Consider an exponentially tight family of probability measures $(\mu_\varepsilon)_{\varepsilon>0}$ on $\mathfrak{X}$.*

(1) For any closed subset $F \subset \mathfrak{X}$,

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(F) \leq -\inf_{u\in F} \Lambda^*(u).$$

(2) Denote $\mathfrak{E}$ the set of exposed points of Λ^* for which some exposing hyperplane ϕ satisfies $\Lambda(\phi) = \lim_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon[\mathrm{e}^{\phi(X)/\varepsilon}] < +\infty$, and $\Lambda(\gamma\phi) < +\infty$ for some $\gamma > 1$. For any open subset $U \subset \mathfrak{X}$,

$$\liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(U) \geq -\inf_{u\in U\cap\mathfrak{E}} \Lambda^*(u).$$

(3) *Suppose that Λ exists as a limit everywhere, is finite-valued and is differentiable in the Gateaux sense. Then,*

$$\liminf_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon(U) \geq - \inf_{u\in U} \Lambda^*(u).$$

for all $u \in U$, i.e., *a LDP holds with good rate function Λ^*.*

PROOF. (1) From Theorem 4.12, one gets an upper bound over compact subsets $K \subset \mathfrak{X}$, and from 4.3, this upper bound extends to arbitrary closed subsets, because one has exponential tightness.

(2) One can assume $\Lambda > -\infty$ everywhere, since otherwise $\Lambda^* = +\infty$ everywhere and the lower bound is trivial. Fix an open set U, an exposed point $x \in U \cap \mathfrak{E}$, and an exposing hyperplane ϕ. By continuity of the linear form ϕ, for $\delta > 0$ sufficiently small, there is an open neighborhood $U_{(x,\eta)} \subset U$ of x such that $|\phi(y - x)| \leq \eta$ for every $y \in U_{(x,\eta)}$ — note that $\mathfrak{X}$ is not supposed to be a metric space, so one cannot take balls. On the other hand, since $\Lambda(\phi) < +\infty$, $\log \mu_\varepsilon[\mathrm{e}^{\phi(X)/\varepsilon}] < +\infty$ for ε small enough. We introduce the modified probability measures

$$\nu_\varepsilon^\phi(dx) = \frac{\mathrm{e}^{\frac{\phi(x)}{\varepsilon}}}{\mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(x)}{\varepsilon}}\right]}\, \mu_\varepsilon(dx).$$

This is the same idea as in the proof of Theorem 3.3. One then writes:

$$\begin{aligned}
\varepsilon \log \mu_\varepsilon(U) \geq \varepsilon \log \mu_\varepsilon(U_{(x,\eta)}) &= \varepsilon \log \int_{U_{(x,\eta)}} \frac{\mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(x)}{\varepsilon}}\right]}{\mathrm{e}^{\frac{\phi(y)}{\varepsilon}}}\, \nu_\varepsilon^\phi(dy) \\
&\geq \varepsilon \log \mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\right] - \phi(x) - \eta + \varepsilon \log \nu_\varepsilon^\phi(U_{(x,\eta)}) \\
&\geq \Lambda(\phi) - \phi(x) - 2\eta + \varepsilon \log \nu_\varepsilon^\phi(U_{(x,\eta)}) \\
&\geq -\Lambda^*(x) - 2\eta + \varepsilon \log \nu_\varepsilon^\phi(U_{(x,\eta)})
\end{aligned}$$

for ε sufficiently small. The lower bound follows now from the following law of large numbers for ν_ε^ϕ:

$$\nu_\varepsilon^\phi(U_{(x,\eta)}) \to 1 \quad \text{as } \varepsilon \to 0.$$

To prove it, fix a family of compact sets K_α as in Definition 4.3, and consider the compact set $U_{(x,\eta)}^{\mathrm{c}} \cap K_\alpha$. We want to apply Theorem 4.12 to the family of measures $(\nu_\varepsilon^\phi)_{\varepsilon>0}$. For $\theta \in \mathfrak{X}^*$, one has

$$\varepsilon \log \nu_\varepsilon^\phi\left[\mathrm{e}^{\frac{\theta(X)}{\varepsilon}}\right] = \varepsilon \log \mu_\varepsilon\left[\mathrm{e}^{\frac{(\phi+\theta)(X)}{\varepsilon}}\right] - \varepsilon \log \mu_\varepsilon\left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}}\right] \simeq \Lambda(\phi+\theta) - \Lambda(\phi),$$

and therefore,

$$\limsup_{\varepsilon\to 0} \varepsilon \log \nu_\varepsilon^\phi(U_{(x,\eta)}^{\mathrm{c}} \cap K_\alpha) \leq - \inf_{y\in U_{(x,\eta)}\cap K_\alpha} g(y),$$

where g is the Legendre-Fenchel transform of the new translated asymptotic cumulant generating function $f(\theta) = \Lambda(\phi + \theta) - \Lambda(\phi)$. Let us compute g:

$$\begin{aligned} g(y) &= \sup_{\theta \in \mathfrak{X}^*} (\theta(y) - \Lambda(\phi + \theta) + \Lambda(\phi)) \\ &= \Lambda^*(y) + \Lambda(\phi) - \phi(y) \\ &\geq \Lambda^*(y) - \Lambda^*(x) - \phi(y - x). \end{aligned}$$

Since ϕ is an exposing hyperplane for x and the convex function Λ^*, the right-hand side is strictly positive for $y \neq x$. So,

$$\limsup_{\varepsilon \to 0} \varepsilon \, \log \nu_\varepsilon^\phi(U_{(x,\eta)}^{\mathrm{c}} \cap K_\alpha) < 0. \tag{4.3}$$

We shall prove in a moment that

$$\limsup_{\varepsilon \to 0} \varepsilon \, \log \nu_\varepsilon^\phi(K_\alpha^{\mathrm{c}}) < 0 \tag{4.4}$$

for α large enough. Taken together, Equations (4.3) and (4.4) ensure that $\limsup_{\varepsilon \to 0} \varepsilon \, \log \nu_\varepsilon^\phi(U_{(x,\eta)}^{\mathrm{c}}) < 0$, and therefore, $\nu_\varepsilon^\phi(U_{(x,\eta)}^{\mathrm{c}}) \to 0$ and $\nu_\varepsilon^\phi(U_{(x,\eta)}) \to 1$.

The inequality (4.4) is obtained again by splitting K_α^{c} in two halves. Set $H_\beta = \{x \in \mathfrak{X} \mid \phi(x) < \beta\}$. For every $\gamma > 1$, by Chernov's inequality,

$$\begin{aligned} \varepsilon \log \nu_\varepsilon^\phi(H_\beta^{\mathrm{c}}) &= \varepsilon \log \int_{H_\beta^{\mathrm{c}}} \nu_\varepsilon^\phi(dx) \\ &\leq \varepsilon \log \int_{\mathfrak{X}} \mathrm{e}^{\frac{(\gamma-1)\phi(x)}{\varepsilon}} \, \nu_\varepsilon^\phi(dx) - \beta(\gamma - 1) \\ &\leq \varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\gamma \phi(X)}{\varepsilon}} \right] - \varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}} \right] - \beta(\gamma - 1). \end{aligned}$$

By hypothesis, there exists $\gamma > 1$ such that this upper bound is finite. Then, for β large enough, the upper bound becomes asymptotically negative:

$$\limsup_{\varepsilon \to 0} \varepsilon \, \log \nu_\varepsilon^\phi(H_\beta^{\mathrm{c}}) < 0. \tag{4.5}$$

On the other hand,

$$\varepsilon \log \nu_\varepsilon^\phi(K_\alpha^{\mathrm{c}} \cap H_\beta) \leq \beta - \varepsilon \log \mu_\varepsilon \left[\mathrm{e}^{\frac{\phi(X)}{\varepsilon}} \right] + \varepsilon \log \nu_\varepsilon(K_\alpha^{\mathrm{c}}),$$

which gives asymptotically

$$\limsup_{\varepsilon \to 0} \varepsilon \, \log \nu_\varepsilon^\phi(K_\alpha^{\mathrm{c}} \cap H_\beta) < \beta - \Lambda(\phi) - \alpha < 0 \quad \text{for } \alpha \text{ large enough.} \tag{4.6}$$

Combining (4.5) and (4.6) leads to Equation (4.4), and then to the law of large numbers for $\nu_\varepsilon^\phi(U_{(x,\eta)})$. As a consequence, for every $\eta > 0$, $\liminf_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(U) \geq -\Lambda^*(x) - 2\eta$, and finally

$$\liminf_{\varepsilon \to 0} \varepsilon \log \mu_\varepsilon(U) \geq - \inf_{x \in U \cap \mathfrak{E}} \Lambda^*(x).$$

(3) Recall that in a topological vector space, a function F is called ***differentiable in the Gateaux sense*** at x if for every direction y, there is a limit to the quotient
$$\frac{F(x+\varepsilon y)-F(x)}{\varepsilon}$$
as ε goes to 0 (the limit does not need to be linear or continuous w.r.t. y, as opposed to Frechet derivatives). These hypotheses on Λ ensure that one can approximate any point $x \in U$ by exposed points $x_n \to x$, with $\Lambda^*(x_n) \to \Lambda^*(x)$. Hence,
$$\inf_{x\in U\cap \mathfrak{E}} \Lambda^*(x) = \inf_{x\in U} \Lambda^*(x),$$
so a weak LDP is proved, and by exponential tightness this is a full LDP with a good rate function. However, the proofs of the approximation result are quite involved, and in particular they rely on advanced convex analysis; so again we refer to [Dembo and Zeitouni, 1998, Chapter 4]. □

4.3. Applications of the Ellis-Gärtner theory

In this Section, we shall use Theorem 4.14 in many situations, and of course we shall need each time to show that the sequence of measures considered is exponentially tight. Note that once Λ and Λ^* have been computed, if one expects a full LDP with good rate function Λ^*, then the level sets of Λ^* are compact and are therefore natural candidates to check the exponential tightness. However, this technique requires one to show in advance that Λ^* is good, which might be quite difficult; and then to prove *a priori* a statement like
$$\limsup_{\varepsilon\to 0} \frac{1}{\varepsilon} \log \mu_\varepsilon[((\Lambda^*)_{\leq\alpha})^c] \leq -\alpha,$$
which is not obvious, or even true in general (notice that this is not at all the same statement as Proposition 4.12). Thus, it is usually easier to prove the exponential tightness by *ad hoc* arguments. In particular, Theorem 4.16 and the remark after Theorem 4.15 provide such criterions.

4.3.1. Cramér's theorem in $\mathbb{R}^d$. A first application of the Ellis-Gärtner theory is a generalization of Theorem 3.3 to a multi-dimensional setting. Fix a dimension d and a random variable X with values in $\mathbb{R}^d$. We consider the mean Z_n of n independent copies of X. Recall that when $d = 1$, a sufficient condition to obtain a LDP with good rate function Λ_X^* is to ask for 0 being in the interior of the domain of the cumulant generating function Λ_X.

THEOREM 4.15. *Denote* $\Lambda_X(x) = \log \mathbb{E}[\mathrm{e}^{\langle x \,|\, X\rangle}]$, *and* $\Lambda_X^*(x) = \sup_{y\in\mathbb{R}^d}(\langle y \,|\, x\rangle - \Lambda(y))$. *If* $\mathcal{D}(\Lambda_X) = \mathbb{R}^d$, *then* Z_n *satisifies a LDP on* $\mathbb{R}^d$ *with good rate function* Λ_X^*.

REMARK. In fact, this LDP still holds if one only assumes that 0 lies in the interior of $\mathcal{D}(\Lambda_X) \subset \mathbb{R}^d$. But the proof does not follow readily from Ellis-Gärtner theorem, and is not a direct modification of the arguments of §3.1; it requires an argument of subadditivity, *cf.* [Dembo and Zeitouni, 1998, §6.1].

PROOF. On a finite-dimensional space, linear forms are represented by scalar products $\langle y \,|\, \cdot \rangle$. Since $\frac{1}{n} \log \mathbb{E}[\mathrm{e}^{\langle nx \,|\, Z_n \rangle}] = \log \mathbb{E}[\mathrm{e}^{\langle x \,|\, X \rangle}] = \Lambda_X(x)$, which is by hypothesis finite everywhere and therefore differentiable, the Ellis-Gärtner theorem will apply with $\Lambda = \Lambda_X$, and the only think to check is the exponential tightness of the family of measures. Denote μ_n the law of Z_n, and $K_\rho = [-\rho, \rho]^d$. The complementary of K_ρ is included into the union of the half spaces

$$H^{\pm}_{\rho,i} = \{x \in \mathbb{R}^d \mid \pm x_i \geq \rho\},$$

and for these sets one can use the one-dimensional Cramér theorem:

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n[H^{\pm}_{\rho,i}] \leq - \inf_{\pm x_i \in [\rho, +\infty)} \Lambda^*_{X_i}(x_i).$$

By the one-dimensional case, the limit of the right-hand side as ρ goes to infinity is $-\infty$, for each i. This proves the exponential tightness. □

4.3.2. Poissonian approximations. To see more of the full strengh of Ellis-Gärtner theorem, one can look at sums of independent random variables that are not equidistributed. Fix a sequence of parameters $p_k \in (0, 1)$, with

$$\sum_{k=1}^{\infty} p_k = +\infty \quad \text{and} \quad \sum_{k=1}^{\infty} (p_k)^2 < +\infty.$$

Notice that the second condition implies that $p_k \to 0$. We then consider the sum of independent Bernoulli variables

$$S_n = \sum_{k=1}^{n} \mathcal{B}(p_k), \quad \text{with } \mathbb{P}[\mathcal{B}(p_k) = 1] = p_k \text{ and } \mathbb{P}[\mathcal{B}(p_k) = 0] = 1 - p_k.$$

It is well-known that sums of Bernoulli variables with parameters going to 0 approximate Poisson variables. Here, one has indeed

$$\begin{aligned} \mathbb{E}\big[\mathrm{e}^{tS_n}\big] &= \prod_{k=1}^{n} \mathbb{E}\big[\mathrm{e}^{t\mathcal{B}(p_k)}\big] = \prod_{k=1}^{n} (1 + p_k\,(\mathrm{e}^t - 1)) \\ &= \mathrm{e}^{\left(\sum_{k=1}^n p_k\right)(\mathrm{e}^t - 1)} \prod_{k=1}^{n} (1 + p_k\,(\mathrm{e}^t - 1))\,\mathrm{e}^{-p_k\,(\mathrm{e}^t - 1)}. \end{aligned}$$

Denote $a_n = \sum_{k=1}^n p_k$, and $\psi_n(t) = \mathbb{E}[\mathrm{e}^{tX_n}]\,\mathrm{e}^{-a_n\,(\mathrm{e}^t - 1)}$; $\psi_n(t)$ is the product in the last term of the previous list of equations. For any t,

$$\log \psi_n(t) = \sum_{k=1}^{n} \log(1 + p_k\,(\mathrm{e}^t - 1)) - p_k(\mathrm{e}^t - 1) = -\sum_{k=1}^{n} \frac{(p_k\,(\mathrm{e}^t - 1))^2}{2}\,(1 + o(1))$$

is convergent since $\sum_{k=1}^{\infty} (p_k)^2$ is supposed finite. Therefore, if $X_n = \frac{S_n}{a_n}$ and μ_n is the law of X_n, then

$$\lim_{n\to\infty} \frac{1}{a_n} \log \mu_n\big[\mathrm{e}^{a_n t\, X}\big] = \lim_{n\to\infty} \frac{1}{a_n} \log \mathbb{E}\big[\mathrm{e}^{t\, S_n}\big] = \mathrm{e}^t - 1 = \Lambda(t).$$

To apply Theorem 4.14, one still has to show the exponential tightness of the sequence of measures $(\mu_n)_{n\in\mathbb{N}}$. In finite dimension, it becomes useful to have a criterion on the limit Λ that automatically ensures this. We shall admit the following:

THEOREM 4.16. *Suppose that $\Lambda : \mathbb{R}^d \to \mathbb{R} \sqcup \{+\infty\}$ is **steep**, which means that it is differentiable on the non-empty open set $\mathcal{D}(\Lambda)^{\circ}$, and that for every point b in the boundary of this open set,*

$$\lim_{x\to b} |\nabla\Lambda(x)| = +\infty.$$

Notice that this is automatically true if $\mathcal{D}(\Lambda) = \mathbb{R}^d$. Then, if

$$\lim_{n\to\infty} \frac{1}{a_n} \log \mu_n[\mathrm{e}^{a_n t X}] = \Lambda(t)$$

for a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ with laws μ_n, the family of laws is exponentially tight and Ellis-Gärtner Theorem 4.14 applies: the random variables $(X_n)_{n\in\mathbb{N}}$ satisfy a LDP with speed $(a_n)_{n\in\mathbb{N}}$ and rate function Λ^*.

In our example, $\mathcal{D}(\Lambda) = \mathbb{R}$, so steepness is automatic and Ellis-Gärtner theorem applies: the laws μ_n of the rescaled sums of Bernoulli variables satisfy a LDP with speed a_n and good rate function

$$\Lambda^*(x) = \sup_{t\in\mathbb{R}} (tx - (\mathrm{e}^t - 1)) = \begin{cases} +\infty & \text{if } x < 0; \\ x\log x - x + 1 & \text{if } x \geq 0. \end{cases}$$

More concretely, one obtains

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}[S_n \geq (1+\varepsilon)a_n] = (1+\varepsilon)\log(1+\varepsilon) - \varepsilon;$$
$$\forall \varepsilon \in (0,1), \quad \lim_{n\to\infty} \frac{1}{a_n} \log \mathbb{P}[S_n \leq (1-\varepsilon)a_n] = (1+\varepsilon)\log(1+\varepsilon) - \varepsilon.$$

Notice that these two statements are not at all implied by the convergence in law towards a Poisson random variable $\frac{S_n}{a_n} \to \mathcal{P}$. In this case one can actually compute the exact asymptotics of $\mathbb{P}[S_n \geq (1+\varepsilon)a_n]$ (instead of the logarithm), but this follows from difficult arguments of harmonic analysis on the characteristic functions of the r.v. S_n.

4.3.3. Ellis-Gärtner theory and moderate deviations. Another advantage of the Ellis-Gärtner theory is its flexibility in comparison to the previous results, because it allows one to choose the rate a_n of the fluctuations that one wants to study. In particular, it leads to a complete description of the fluctuations of sums of i.i.d. real random variables, in a fashion similar to what was described in §3.1.3. Hence, fix a law μ of Laplace transform Φ defined on a neighborhood of 0, with mean $\Phi'(0) = 0$ and variance $\sigma^2 = \Phi''(0)$. We fix an exponent $\alpha \in [\frac{1}{2}, 1]$, and consider the laws μ_n of the means $M_n = \frac{1}{n}\sum_{k=1}^n X_k$ of a sequence of i.i.d. random variables with law μ. One has:

$$\frac{1}{n^\alpha} \log \mu_n \left[\mathrm{e}^{n^\alpha M_n}\right] = n^{1-\alpha} \log \Phi(n^{\alpha-1}x) \to \begin{cases} \Lambda(x) = \log\Phi(x) & \text{if } \alpha = 1; \\ \frac{\sigma^2 x^2}{2} & \text{if } \frac{1}{2} \leq \alpha < 1. \end{cases}$$

In both cases, Ellis-Gärtner theorem applies, so one recovers Cramér's theorem 3.3 when $\alpha = 1$, and the moderate deviations of Proposition 3.6 when $\alpha < 1$.

4.3.4. A new proof of Sanov's theorem. By using Ellis-Gärtner theorem, one can give a quick new proof of Sanov's theorem, and in full generality (*e.g* without assumption on the initial distribution). Indeed, if $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the random empirical measure of a Markov chain with irreducible transition matrix $(p(i,j))_{1\leq i,j\leq N}$, then its Laplace transform is

$$\begin{aligned}\Phi_n(n\lambda \in \mathbb{R}^N) &= \mathbb{E}\left[\exp\left(\sum_{i=1}^n \lambda_{X_i}\right)\right] \\ &= \sum_{a_0,a_1,\dots,a_n} \pi_0(a_0)\, p(a_0,a_1)\mathrm{e}^{\lambda_{a_1}}\, p(a_1,a_2)\mathrm{e}^{\lambda_{a_2}}\cdots p(a_{n-1},a_n)\mathrm{e}^{\lambda_{a_n}} \\ &= \sum_{a,b=1}^N \pi_0(a)\big((p(i,j)\,\mathrm{e}^{\lambda_j})^n\big)(a,b).\end{aligned}$$

Let $r(\lambda)$ be the Perron-Frobenius of the matrix $(p(i,j)\mathrm{e}^{\lambda_j})_{1\leq i,j\leq N}$, and π^λ the corresponding eigenvector. Since it has positive coordinates, one can estimate the previous quantity by multiples of the same sum, but starting from π^λ instead of π_0. It follows that

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\Phi_n(n\lambda) &= \lim_{n\to\infty}\frac{1}{n}\log\left(\sum_{a,b=1}^N \pi^\lambda(a)\big((p(i,j)\,\mathrm{e}^{\lambda_j})^n\big)(a,b)\right) \\ &= \lim_{n\to\infty}\frac{1}{n}\log\left(\sum_{b=1}^N (r(\lambda))^n\,\pi^\lambda(b)\right) = \log r(\lambda).\end{aligned}$$

But then we have seen that the Legendre-Fenchel transform of $\log r(\lambda)$ was $I(\nu)$, the generalization of relative entropy involved in Sanov's theorem. Since the space of probability measure is compact, exponential tightness is automatic and Theorem 4.14 gives back immediately Sanov's theorem 3.12.

4.3.5. Fine analysis of Brownian paths. To conclude this chapter and the lecture, we are going to study in detail one of the simplest example of large deviations in a infinite-dimensional setting. Consider a Brownian motion $W : \mathbb{R}_+ \to \mathbb{R}$, as defined by Theorem 2.9. If $\mathscr{C}(\mathbb{R}_+,\mathbb{R})$ is endowed with the topology of uniform convergence on every compact, then the law of W is the unique probability measure on this space for which $W(0) = 0$ almost surely, and such that for every times $t_1 \leq t_2 \leq \cdots \leq t_n$, the vector

$$(W_{t_1}, W_{t_2} - W_{t_1}, \dots, W_{t_n} - W_{t_{n-1}})$$

is a vector of independent centered Gaussian random variables with variances

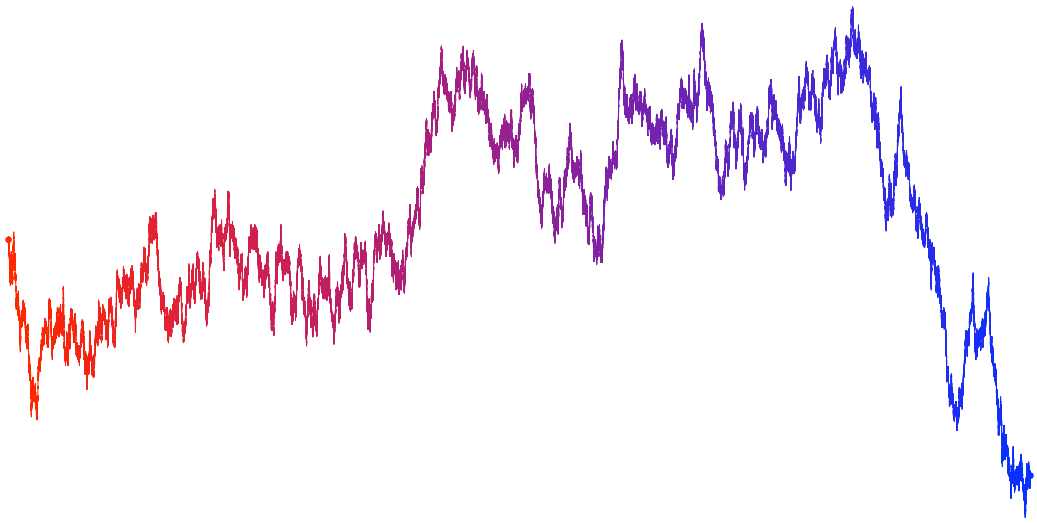
$$(t_1, t_2 - t_1, \dots, t_n - t_{n-1}).$$

The picture on the next page represents (an approximation of) a Brownian motion on a finite interval of time. Let us then give a list of well-known properties of the Brownian motions — we refer to [**Revuz and Yor, 2004**, Chapter 1] for proofs.

(1) Invariance by scaling: if $(W_s)_{s\in\mathbb{R}_+}$ is a BM, then so is $(\sqrt{\varepsilon}\, W_{s/\varepsilon})_{s\in\mathbb{R}^+}$ for all $\varepsilon > 0$.

(2) Invariance by time-translation: if $(W_s)_{s\in\mathbb{R}_+}$ is a BM, then for every time $t \geq 0$, $(W_{t+s} - W_t)_{s\in\mathbb{R}_+}$ is a BM independent of $(W_s)_{0\leq s\leq t}$. This is even true for an optional random time $t = \tau$ (assuming τ is almost surely finite).

(3) Invariance by time-inversion: if $(W_s)_{s\in\mathbb{R}_+}$ is a BM, then so is $(s\, W_{1/s})_{s\in\mathbb{R}^+}$.



(4) Invariance by space-reflexion: if $(W_s)_{s\in\mathbb{R}_+}$ is a BM and τ is a finite optional time, denote

$$W_s^\tau = \begin{cases} W_s & \text{if } s < \tau, \\ 2W_\tau - W_s & \text{if } s \geq \tau \end{cases}$$

the BM reflected at time τ. It is again a Brownian motion.

(5) Invariance by time-reflexion: if $(W_s)_{s\in[0,1]}$ is a BM (with finite horizon), then so is $(W_1 - W_{1-s})_{s\in[0,1]}$.

(6) The function $t \mapsto W_t$ is almost surely nowhere locally α-Hölder for $\alpha > \frac{1}{2}$, and it is locally α-Hölder for $\alpha < \frac{1}{2}$.

(7) One has $\lim_{t\to\infty} \frac{|W_t|}{t^\alpha} = 0$ almost surely for every $\alpha > \frac{1}{2}$, and in particular for $\alpha = 1$.

The last properties allows one to consider a BM as a probability measure on a slightly smaller space than $\mathscr{C}(\mathbb{R}_+, \mathbb{R})$, namely,

$$\mathfrak{X} = \left\{ W \in \mathscr{C}(\mathbb{R}_+, \mathbb{R}) \mid W(0) = 0, \ \lim_{t\to\infty} \frac{|W(t)|}{t} = 0 \right\},$$

which is a Banach space for the norm

$$\|W\|_{\mathfrak{X}} = \sup_{t\in\mathbb{R}_+} \frac{|W(t)|}{1+t}.$$

Notice that the topology of $\mathfrak{X}$ is just the restriction of the topology of $\mathscr{C}(\mathbb{R}_+, \mathbb{R})$, but this time it corresponds to a norm (and therefore a topology which is regular and locally convex).

LEMMA 4.17. *The topological dual of $\mathfrak{X}$ is the Banach space $\mathfrak{X}^*$ of finite signed Borel measures on $\mathbb{R}_+$ with*

$$\lambda(\{0\}) = 0 \qquad ; \qquad \|\lambda\|_{\mathfrak{X}^*} = \int_{\mathbb{R}^+} (1+s)\,|\lambda|(ds) < +\infty.$$

PROOF. Notice that $W_t \mapsto \frac{W_t}{1+t}$ is an isometry between $\mathfrak{X}$ and the space of bounded continuous functions on $\mathbb{R}_+$ that vanish at 0 and $+\infty$, endowed with the sup norm. This space itself can be identified with

$$\mathfrak{Y} = \{f \in \mathscr{C}([0,1]) \mid f(0) = f(1) = 0\}.$$

by using the map $s \mapsto \frac{2}{\pi}\arctan(s)$ to send $[0,+\infty]$ to $[0,1]$. By Riesz' representation theorem, the dual of $\mathscr{C}([0,1])$ is the set of signed measures on $[0,1]$, and the dual of $\mathfrak{Y}$ is the set of signed measures with $\lambda(\{0\}) = \lambda(\{1\}) = 0$. Going back in the chain of isometries, we have shown the lemma. Notice that $\|\cdot\|_{\mathfrak{X}^*}$ is indeed the dual norm of $\|\cdot\|_{\mathfrak{X}}$, that is

$$\|\lambda\|_{\mathfrak{X}^*} = \sup_{W\in\mathfrak{X}} \frac{|\lambda(W)|}{\|W\|_{\mathfrak{X}}}.$$

□

We want to prove a LDP for the family of probability measures $(\mu_\varepsilon)_{\varepsilon>0}$ of the scaled BM $(\sqrt{\varepsilon}\,W_s)_{s\in\mathbb{R}_+}$. Obviously $\mu_\varepsilon \to \delta_0$, the law of the constant function equal to 0, so this will be a way to understand the "large values" of a Brownian motion. By scaling invariance, this is also a way to understand the asymptotic behavior in time. In view of Theorem 4.14, the first thing to do is to compute the log-Laplace transform of μ.

LEMMA 4.18. *For every $\lambda \in \mathfrak{X}^*$,*

$$\log \mathbb{E}\big[\mathrm{e}^{\lambda(W)}\big] = \log\left(\int_{\mathfrak{X}} \exp\left(\int_0^\infty \omega(t)\,\lambda(dt)\right)\mu(d\omega)\right) = \frac{1}{2}\iint_{(\mathbb{R}_+)^2} (s\wedge t)\,\lambda(ds)\lambda(dt).$$

PROOF. Under Wiener's law μ, $(W_t)_{t\in\mathbb{R}^+}$ is a Gaussian process with covariance form $C(s,t) = s\wedge t$, that is to say that every finite vector $(W_{t_1},\dots,W_{t_n})$ is a centered Gaussian vector with covariance matrix $(t_i \wedge t_j)_{1\le i,j\le n}$. In particular, every linear combination $\lambda_1 W_{t_1} + \lambda_2 W_{t_2} + \cdots + \lambda_n W_{t_n}$ is a Gaussian random variable with mean 0 and variance

$$\sum_{i,j=1}^n \lambda_i\lambda_j(t_i\wedge t_j).$$

This formula can be extended by density to any linear form of the path $(W_t)_{t\in\mathbb{R}_+}$ which is in the dual space $\mathfrak{X}^*$; so, under μ, $\lambda(\omega) = \int_0^\infty \omega(t)\,\lambda(dt)$ is a Gaussian random variable with variance

$$\iint_{(\mathbb{R}_+)^2} (s\wedge t)\,\lambda(ds)\lambda(dt),$$

which is indeed finite. □

Thus, we have computed the log-Laplace transform $\Lambda(\lambda)$ of Wiener's law, and it follows that

$$\lim_{\varepsilon\to 0}\left(\frac{1}{\varepsilon}\log\mu_\varepsilon(\lambda)\right)=\lim_{\varepsilon\to 0}\frac{\Lambda(\sqrt{\varepsilon}\lambda)}{\varepsilon}=\Lambda(\lambda).$$

One can therefore expect a LDP for $(\mu_\varepsilon)_{\varepsilon>0}$ with rate function the Legendre-Fenchel transform of Λ. Notice that for non-atomic measures λ (which are dense in $\mathfrak{X}^*$),

$$\begin{aligned}\Lambda(\lambda)&=\frac{1}{2}\iint_{(\mathbb{R}_+)^2}(s\wedge t)\,\lambda(ds)\lambda(dt)=\iint_{(\mathbb{R}^+)^2}\mathbf{1}_{s\le t}\,s\,\lambda(ds)\lambda(dt)\\&=\iint_{(\mathbb{R}^+)^2}\mathbf{1}_{s\le t}\,s\,F_\lambda'(s)F_\lambda'(t)\,ds\,dt\qquad\text{with } F_\lambda(t)=\int_t^\infty\lambda(ds)\\&=-\int_{\mathbb{R}^+}s\;F_\lambda'(s)\,F_\lambda(s)\,ds=-\frac{1}{2}\int_0^\infty s\,d((F_\lambda(s))^2)=\frac{1}{2}\int_0^\infty(F_\lambda(s))^2\,ds.\end{aligned}$$

This identity leads to the following definition:

Definition 4.19. *The* ***Cameron-Martin space*** $\mathscr{H}^1(\mathbb{R}_+)$ *is the set of locally absolutely continuous functions on* $\mathbb{R}_+$ *that vanishes at* 0*, and with derivative in* $\mathscr{L}^2(\mathbb{R}_+)$.

An absolutely continuous function on a compact interval $[a,b]$ is a function f such that for every $\varepsilon>0$, there exists some $\eta>0$ with the property that if $([x_k,y_k])_{k\le K}$ are disjoint intervals in $[a,b]$ of total length $\sum_{k\le K}y_k-x_k\le\eta$, then

$$\sum_{k\le K}|f(y_k)-f(x_k)|\le\varepsilon.$$

This is quite stronger than uniform continuity, and it is equivalent to the existence of a derivative f' almost everywhere and that is integrable on $[a,b]$. So,

$$f(x)=f(a)+\int_a^x f'(s)\,ds\qquad\text{with } f'\in\mathscr{L}^1([a,b]).$$

Then, the Cameron-Martin space is the space of functions with derivative that exists almost everywhere and is square integrable on $\mathbb{R}_+$. It is endowed with the Hilbert space topology given by

$$\|f\|_{\mathscr{H}^1}=\|f'\|_{\mathscr{L}^2}=\sqrt{\int_0^\infty|f'(s)|^2\,ds}$$

Notice that $\mathscr{H}^1(\mathbb{R}_+)\subset\mathfrak{X}$: indeed, elements of $\mathscr{H}^1(\mathbb{R}_+)$ are indeed continuous functions that vanish at 0, and

$$\frac{|f(t)|}{1+t}=\frac{1}{1+t}\left|\int_0^t f'(s)\,ds\right|\le\frac{\sqrt{t}}{1+t}\,\|f'\|_{\mathscr{L}^2}\to_{t\to\infty}0.$$

Remark. The Cameron-Martin space is involved in exponential change of measures for Brownian motions. Namely, if $(W_t)_{t\in\mathbb{R}_+}$ is a BM under the law μ, then the law μ^θ

of a drifted BM $(W_t + \theta_t)_{t\in\mathbb{R}_+}$ is absolutely continuous with respect to μ if and only if $\theta \in \mathscr{H}^1(\mathbb{R}_+)$, with Radon-Nikodym derivative

$$\frac{d\mu^\theta}{d\mu}(\omega) = \exp\left(\int_{\mathbb{R}} \theta'(s)\,\omega(s)\,ds - \frac{1}{2}\|\theta\|^2_{\mathscr{H}^1}\right).$$

PROPOSITION 4.20. *The Legendre-Fenchel transform of Λ is*

$$\Lambda^*(\omega) = \begin{cases} \frac{1}{2}\|\omega\|^2_{\mathscr{H}^1} & \text{if } \omega \in \mathscr{H}^1(\mathbb{R}_+); \\ +\infty & \text{otherwise.} \end{cases}$$

PROOF. Suppose first that $\omega \in \mathscr{H}^1(\mathbb{R}_+)$. For λ non-atomic, one has

$$\begin{aligned}\lambda(\omega) - \Lambda(\lambda) &= \int_{\mathbb{R}_+} \omega(s)\,\lambda(ds) - \frac{1}{2}\int_{\mathbb{R}_+} (F_\lambda(s))^2\,ds \\ &= \int_{\mathbb{R}_+} \left(\omega'(s)\,F_\lambda(s) - \frac{1}{2}(F_\lambda(s))^2\right) ds\end{aligned}$$

which leads by density into $\mathfrak{X}^*$ to the identity

$$\sup_{\lambda\in\mathfrak{X}^*} (\lambda(\omega) - \Lambda(\lambda)) = \sup_{F\in\mathscr{L}^2(\mathbb{R}_+)} \left(\langle \omega' \mid F\rangle_{\mathscr{L}^2} - \frac{1}{2}\langle F \mid F\rangle_{\mathscr{L}^2}\right) = \frac{1}{2}\langle \omega' \mid \omega'\rangle_{\mathscr{L}^2} = \frac{1}{2}\|\omega\|^2_{\mathscr{H}^1}.$$

Conversely, suppose $\Lambda^*(\omega) < +\infty$. For any smooth path with support compact $\widetilde{\omega}$, one defines a linear form in $\mathfrak{X}^*$ by $\lambda_{\widetilde{\omega}}([t,+\infty)) = \widetilde{\omega}(t)$. One has

$$\begin{aligned}K = \Lambda^*(\omega) &\geq \int_0^\infty \omega(s)\,\lambda_{\widetilde{\omega}}(ds) - \frac{1}{2}\int_0^\infty (\widetilde{\omega}(s))^2\,ds \\ &\geq -\int_0^\infty \omega(s)\,\widetilde{\omega}'(s)\,ds - \frac{1}{2}\int_0^\infty (\widetilde{\omega}(s))^2\,ds\end{aligned}$$

Therefore, the linear map on the set of smooth paths with compact supports

$$\widetilde{\omega} \mapsto -\int_0^\infty \omega(s)\,\widetilde{\omega}'(s)\,ds$$

is continuous w.r.t. the $\mathscr{L}^2$-norm, so it can be represented by a unique $\omega' \in \mathscr{L}^2(\mathbb{R}_+)$:

$$\forall \widetilde{\omega} \text{ smooth and compactly supported, } \; -\int_0^\infty \omega(s)\,\widetilde{\omega}'(s)\,ds = \int_0^\infty \omega'(s)\,\widetilde{\omega}(s)\,ds.$$

From there it is immediate that $\omega(t) = \int_0^t \omega'(s)\,ds$, and by construction the derivative is indeed in $\mathscr{L}^2(\mathbb{R}_+)$, so $\omega \in \mathscr{H}^1(\mathbb{R}_+)$. □

To obtain the LDP, it suffices now to prove the exponential tightness: one can then apply Ellis-Gärtner theorem, including the third part, since Λ exists everywhere and is finite and differentiable. The standard proof of this exponential tightness for laws of Brownian motions relies on the following clever trick:

LEMMA 4.21 (Fernique). *Let μ be a measure on a topological vector space $\mathfrak{X}$, and $F : \mathfrak{X} \to [0, +\infty]$ a measurable functional which satisfies*

$$\forall \alpha \in \mathbb{R},\ \forall x \in \mathfrak{X},\ F(\alpha x) = |\alpha|\, F(x) \qquad ; \qquad \forall (x, y) \in \mathfrak{X},\ F(x + y) \leq F(x) + F(y).$$

Suppose that $\mu \otimes \mu$ is invariant by the rotation

$$(x, y) \mapsto \left(\frac{x - y}{\sqrt{2}}, \frac{x + y}{\sqrt{2}} \right).$$

If F is finite μ-almost surely, then there exists an $\alpha > 0$ such that

$$\int_{\mathfrak{X}} \exp(\alpha (F(x))^2)\, \mu(dx) < \infty.$$

PROOF. Let $s < t$ be two positive real numbers. One has

$$\begin{aligned}
&\mu[F(x) \leq s]\, \mu[F(x) \geq t] \\
&= \mu^{\otimes 2}[F(x) \leq s \text{ and } F(y) \geq t] \\
&= \mu^{\otimes 2}[F(x - y) \leq \sqrt{2}\, s \text{ and } F(x + y) \geq \sqrt{2}\, t] \quad \text{by invariance of } \mu \text{ by rotation;} \\
&\leq \mu^{\otimes 2}[|F(x) - F(y)| \leq \sqrt{2}\, s \text{ and } F(x) + F(y) \geq \sqrt{2}\, t] \quad \text{by subadditivity;} \\
&\leq \mu^{\otimes 2}\left[F(x) \geq \frac{t - s}{\sqrt{2}} \text{ and } F(y) \geq \frac{t - s}{\sqrt{2}} \right] = \left(\mu \left[F(x) \geq \frac{t - s}{\sqrt{2}} \right] \right)^2.
\end{aligned}$$

Fix s such that $\mu[F(x) \leq s] > \frac{1}{2}$, and define a sequence $(t_n)_{n \in \mathbb{N}}$ by

$$t_0 = s \qquad ; \qquad t_n = s + \sqrt{2}\, t_{n-1} = \left(\frac{\sqrt{2}^{n+1} - 1}{\sqrt{2} - 1} \right) s.$$

The previous computation shows that

$$\frac{\mu[F(x) \geq t_n]}{\mu[F(x) \leq t_0]} \leq \left(\frac{\mu[F(x) \geq t_{n-1}]}{\mu[F(x) \leq t_0]} \right)^2 \leq \left(\frac{\mu[F(x) \geq t_0]}{\mu[F(x) \leq t_0]} \right)^{2^n} = \theta^{2^n}$$

with $\theta < 1$. Then, with $t_{n-1} = 0$,

$$\begin{aligned}
\int_{\mathfrak{X}} \exp(\alpha (F(x))^2) \mu(dx) &\leq \sum_{n=0}^{\infty} \mu[t_n \geq F(x) \geq t_{n-1}]\, \exp(\alpha (t_n)^2) \\
&\leq \exp(\alpha s^2) + \sum_{n=0}^{\infty} \mu[F(x) \geq t_n]\, \exp(\alpha (t_{n+1})^2) \\
&\leq K_1^{\alpha} + \sum_{n=0}^{\infty} 2\, \exp(2^n \log \theta + \alpha K_2\, 2^{\frac{n}{2}})
\end{aligned}$$

for some constants $K_1 = \exp(s^2)$ and $K_2 = 2s/(\sqrt{2} - 1)$. Since $\log \theta < 0$ the proof is done. □

LEMMA 4.22. *Fernique's theorem applies with μ the law of the Brownian motion, and*

$$F(\omega) = \sup_{t\geq 1} \frac{|\omega(t)|}{t^{3/4}} + \sum_{n=1}^{\infty} \frac{1}{2^n} \sup_{0\leq s<t\leq n} \frac{|\omega(t)-\omega(s)|}{|t-s|^{1/4}}.$$

Moreover, the level sets of this function are relatively compact in $\mathfrak{X}$.

PROOF. The second part of the lemma is an immediate consequence of Arzela-Ascoli Theorem 2.5, since functions f with bounded $F(f) \leq B$ are uniformly bounded on every compact $[0,n]$ by $B\,n^{3/4}$, and also equicontinuous on every compact $[0,n]$, with local δ-module of uniform continuity smaller than $2^n\,B\,\delta^{1/4}$. For the first part of the lemma, one has mainly to check that $F(\omega) < \infty$ almost surely under Wiener's law, since the subadditivity of each term in F is obvious, as well as the rotation-invariance of μ (a two-dimensional Brownian motion is indeed invariant by any rotation in $\mathrm{SO}(2,\mathbb{R})$). However,

$$\frac{|\omega(t)|}{t^{3/4}} \to_{t\to\infty} 0$$

almost surely, so it is bounded. On the other hand, since ω is almost surely $\frac{1}{4}$-Hölder on $[0,1]$, by using Fernique's theorem, the random variables

$$\sup_{0\leq s<t\leq n} \frac{|\omega(t)-\omega(s)|}{|t-s|^{1/4}}$$

have moments of all order, and even well-defined Laplace transforms for their squares. In particular, one can consider for any n

$$C(n) = \int_{\mathfrak{X}} \sup_{0\leq s<t\leq n} \left(\frac{|\omega(t)-\omega(s)|}{|t-s|^{1/4}}\right)^8 \mu(d\omega),$$

and by invariance of the BM by scaling, $C(n) = C(1)\,n^2$. It follows that the second term in $F(w)$ has norm in $\mathscr{L}^8(\mathbb{R})$ smaller than

$$\sum_{n=1}^{\infty} \left(\frac{C(1)\,n^2}{2^n}\right)^{1/8} < +\infty,$$

so in particular it is finite almost surely. □

COROLLARY 4.23. *The laws $(\mu_\varepsilon)_{\varepsilon>0}$ of the rescaled Brownian motions are exponentially tight.*

PROOF. By Fernique's theorem, there is some $\alpha > 0$ such that $\mu[\mathrm{e}^{\alpha F^2}] = K < \infty$, and therefore,

$$\mu_\varepsilon[(F_{\leq B})^{\mathrm{c}}] = \mu[(F_{\leq B/\sqrt{\varepsilon}})^{\mathrm{c}}] = \mu\left[F^2 > \frac{B^2}{\varepsilon}\right] \leq \frac{\mu[\mathrm{e}^{\alpha F^2}]}{\mathrm{e}^{\frac{\alpha B^2}{\varepsilon}}} \leq K\,\mathrm{e}^{-\frac{\alpha B^2}{\varepsilon}};$$

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mu_\varepsilon[(F_{\leq B})^{\mathrm{c}}] \leq -\alpha\,B^2.$$

Moreover, the previous lemma ensures that the level sets $F_{\leq B}$ are relatively compact. □

THEOREM 4.24 (Schilder). *The laws $(\mu_\varepsilon)_{\varepsilon>0}$ of the rescaled Brownian motions satisfy a LDP with good rate function Λ^* given by Proposition 4.20.*

REMARK. It might seem a bit counter-intuitive that the rate function of the LDP for BM is finite only on paths with derivative almost everywhere and that is square integrable, although Brownian motions themselves are nowhere derivable. However, since $\mathscr{H}^1(\mathbb{R}_+)$ is dense in $\mathfrak{X}$ for the topology of uniform convergence on compact subsets, one obtains in the LDP meaningful upper and lower bounds in most cases. Intuitively, Schilder's theorem ensures that when one rescales a Brownian motion to zero, the last paths that one observes are close to paths in the Cameron-Martin space, and with rate of exponential decay given by the functional $\|\cdot\|_{\mathscr{H}^1}$. The next theorem, which ends the lecture, makes this much more concrete.

THEOREM 4.25 (Strassen). *Denote*

$$X^{(n)}(t) = \frac{W(nt)}{\sqrt{2n\,\log\log n}},$$

where W is a standard Brownian motion. Almost surely, the sequence $(X^{(n)})_{n\in\mathbb{N}}$ is relatively compact in $\mathfrak{X}$, and its closure (the limits of convergent subsequences of $(X^{(n)})_{n\in\mathbb{N}}$) is exactly the unit ball of the Cameron-Martin space:

$$\left(X^{\phi(n)}(\omega) \to_{n\to\infty} X \quad\Longleftrightarrow\quad X \in \mathscr{H}^1(\mathbb{R}^+) \text{ and } \int_0^\infty |X'(s)|^2\,ds \le 1\right) \quad \text{a.s. in } \omega.$$

We refer to [**Deuschel and Stroock, 1989**, Chapter 1] for a proof of this result, where most computations are made starting from Schilder's theorem. Strassen law has an important consequence known as the ***law of the iterated logarithm***: given a Brownian motion,

$$\limsup_{t\to\infty} \frac{W_t}{\sqrt{2t\,\log\log t}} = 1 \quad \text{almost surely.}$$

Indeed, denote B the unit ball of the Cameron-Martin space

$$\mathscr{H}([0,1]) = \left\{ f \mid f(0) = 0 \text{ and } \int_0^1 (f'(s))^2\,ds < \infty \right\}.$$

From Strassen theorem, it suffices to compute $\sup_{X\in B} X(1)$. However, simple calculus of variations shows that the maximiser X of $X(1)$ under the constraint $\int_0^1 (X'(s))^2\,ds = 1$ is the affine function $X(t) = t$, so the maximal value is indeed 1.

Bibliography

[Billingsley, 1999] Billingsley, P. (1999). Convergence of Probability Measures. 2nd edition, Wiley.

[Dembo and Zeitouni, 1998] Dembo, A. and Zeitouni, O. (1998). Large Deviations Techniques and Applications, vol. 38, of Stochastic Modelling and Applied Probability. 2nd edition, Springer-Verlag.

[Deuschel and Stroock, 1989] Deuschel, J.-D. and Stroock, D. W. (1989). Large Deviations, vol. 137, of Pure and Applied Mathematics. Academic Press.

[Dunford and Schwartz, 1988] Dunford, N. and Schwartz, J. T. (1988). Linear Operators. Wiley.

[Feng and Kurtz, 2006] Feng, J. and Kurtz, T. G. (2006). Large Deviations for Stochastic Processes, vol. 131, of Mathematical Surveys and Monographs. American Mathematical Society.

[Hörmander, 1994] Hörmander, L. (1994). Notions of convexity, vol. 127, of Progress in Mathematics. Birkhäuser.

[Kallenberg, 2001] Kallenberg, O. (2001). Foundations of Modern Probability. Probability and Its Applications, 2nd edition, Springer-Verlag.

[Lang, 1993] Lang, S. (1993). Real and Functional Analysis, vol. 142, of Graduate Texts in Mathematics. Springer-Verlag.

[Meyer, 2000] Meyer, C. (2000). Matrix analysis and applied linear algebra. SIAM.

[Revuz and Yor, 2004] Revuz, D. and Yor, M. (2004). Continuous martingales and Brownian motions, vol. 293, of Grundlehren der mathematischen Wissenschaften. 3rd edition, Springer-Verlag.

[Schneider, 1993] Schneider, R. (1993). Convex Bodies: The Brunn-Minkowski Theory, vol. 44, of Encyclopaedia of Mathematics and Its Applications. Cambridge University Press.