Lecture Notes on Probability Theory
Chapters 7 to 9

Fall 2021 course at ETH, Wendelin Werner

CHAPTER 7

Convergence of probability measures, characteristic functions and the central limit theorem

7.1. Definition of weak convergence

We are now going to study a very different type of questions than in the previous chapters. So far, we have mostly been working in a given probability space $(\Omega, \mathcal{A}, P)$ and looking at sequences of random variables defined in this space, and at their convergence when $n \to \infty$. In all cases (almost sure convergence, convergence in probability, L^p convergence), this notion was implying convergence in probability, i.e., that the *realizations* of the random variables X_n and X were likely to be close when n was large.

We are now going instead going to look at sequences of probability measures defined on a given space, and discuss their convergence.

While (as we will see in a moment), convergence of probability measures are interesting (and useful to study) on more general metric spaces, we will focus here mostly on sequences of probability measures on $\mathbb{R}$ (we will then briefly discuss probability measures on $\mathbb{R}^d$) endowed with the Borel σ-field.

Let us first start with the main concept that we will play with in this chapter:

Definition 7.1.1 (Weak convergence). *We say that a sequence of probability measures $(P_n)_{n\geq 1}$ on a metric space (E, d) (endowed with its Borel σ-field) converges weakly to a probability measure P (on this same space), if for any continuous bounded function from E to $\mathbb{R}$,*

$$\lim_{n\to\infty} \int f(x) dP_n(x) = \int f(x) dP(x).$$

Remark 7.1.2. *There will be no notion of strong convergence of probability measures. This notion of convergence will essentially the only one that we will be discussing!*

Remark 7.1.3. *We will only be dealing with the cases $E = \mathbb{R}$ and $E = \mathbb{R}^d$ in these lectures, even if many of the results that we will be discussing can actually extended to the case of separable complete metric spaces.*

Remark 7.1.4. *The use of continuous functions here is quite natural. Indeed, one wants for instance to say that the sequence of Dirac masses at $1/n$ converges weakly to the Dirac mass at 0, and this indeed holds because $f(1/n) \to f(0)$ when f is continuous.*

Remark 7.1.5. *There exist equivalent descriptions of weak convergence. For instance, it is possible to show that a sequence of probability measures $(P_n)_{n\geq 1}$ in $\mathbb{R}^d$ converges weakly to a probability measure P if and only if for all open set O in $\mathbb{R}^d$, $\limsup_{n\to\infty} P_n(O) \leq P(O)$. Passing to the complement, this is also equivalent to saying that for all closed set F, $\liminf_{n\to\infty} P_n(F) \geq P(F)$. But we won't need or use such equivalent characterizations in these lectures.*

In the next few sections, we will be focusing only on the cases of probability measures on $\mathbb{R}$. We will come back to the case of $\mathbb{R}^d$ in the final section of this chapter.

7.2. Weak convergence and distribution functions

Recall that a probability measure P on $\mathbb{R}$ can be characterized by its distribution $F_P(x) = P((-\infty, x])$, which is right-continuous, non-decreasing and satisfies $\lim_{-\infty} F = 0$ and $\lim_{+\infty} F = 1$, and that conversely, any function F with these properties in the distribution function of some probability measure on $\mathbb{R}$. Since a point of continuity of F corresponds necessarily to a positive jump from $F(x-)$ to $F(x)$ (and the interval $(F(x-), F(x))$ contains rational numbers), the set of discontinuity points of F is at most countable.

Clearly, the example of Dirac masses at $1/n$ for which all $F_{\delta_{1/n}}(0) = 0$ and $F(\delta_0) = 1$, shows that if P_n converges weakly to P, then one does not necessarily have that $F_{P_n}(x) \to F_P(x)$ for all x. However:

Proposition 7.2.1. *Suppose that P_n is a sequence of probability measures on $\mathbb{R}$, that P is a probability measure on $\mathbb{R}$, and let F_n (resp. F) denote the distribution functions of P_n and P respectively. Then, P_n converges weakly to P if and only if for every point of continuity x of F, $\lim_{n\to\infty} F_n(x) = F(x)$.*

Proof. Suppose first that P_n converges weakly to P, and let x_0 be a point of continuity of F. Our goal to show that $F_n(x_0) \to F(x_0)$ as $n \to \infty$.

For each $\varepsilon > 0$, let U_ε be the continuous function that is equal to 1 on $(\infty, x_0]$, is equal to 0 on $[x_0 + \varepsilon, \infty)$ and is linear on $[x_0, x_0 + \varepsilon]$. Since $1_{(\infty,x_0]} \le U_\varepsilon \le 1_{(-\infty,x_0+\varepsilon]}$, we get that for all $n \ge 1$,

$$F_n(x_0) \le \int U_\varepsilon(x) dP_n(x) \text{ and } \int U_\varepsilon(x) dP(x) \le F(x_0 + \varepsilon).$$

The weak convergence statement applied to the continuous function U_ε therefore shows that

$$\limsup_{n\to\infty} F_n(x_0) \le \limsup_{n\to\infty} \int U_\varepsilon(x) dP_n(x) = \int U_\varepsilon(x) dP(x) \le F(x_0 + \varepsilon).$$

But since this is true for all ε and by right-continuity of F, we get that $\limsup_{n\to\infty} F_n(x_0) \le F(x_0)$.

For each $\varepsilon > 0$, let now V_ε be the continuous function that is equal to 1 on $(\infty, x_0 - \varepsilon]$, is equal to 0 on $[x_0, \infty)$ and is linear on $[x_0 - \varepsilon, x_0]$. The weak convergence applied to the continuous function V_ε then shows just as above that $\liminf_{n\to\infty} F_n(x_0) \ge F(x_0 - \varepsilon)$. But since this is true for all ε and by left-continuity of F at x_0, we get that $\liminf_{n\to\infty} F_n(x_0) \ge F(x_0)$.

We can therefore conclude that $F_n(x_0) \to F(x_0)$ as $n \to \infty$.

Let us now conversely assume that F_n converges to F at every point of continuity of F and let us prove the weak convergence. Let f be a bounded continuous function from $\mathbb{R}$ into $\mathbb{R}$. It will be convenient to use the construction of P out of the Lebesgue measure on $[0, 1]$ that shows that

$$\int_0^1 f(Y(t)) dt = \int f(x) dP(x)$$

where

$$Y(t) = \sup\{y \ : \ F(y) < t\}$$

and the same property for P_n and F_n. We therefore need to show that

$$\int_0^1 f(Y_n(t)) dt \to \int_0^1 f(Y(t)) dt,$$

as $n \to \infty$. By dominated convergence and continuity of f, it suffices to prove that for almost every t with respect to the Lebesgue measure, $Y_n(t) \to Y(t)$ as $n \to \infty$.

Let us prove this in a rather pedestrian way:

- Let us first suppose that $t \in (0,1) \setminus F(\mathbb{R})$, so that $t \in (F(x-), F(x))$ for some point of discontinuity x of F. Then one can find an increasing sequence x_l and a decreasing sequence x_l' that both tend to x, and such that all x_l and x_l' are points of continuity of F (the set of discontinuity points being at most countable). Applying the fact that $F_n(x_l)$ and $F_n(x_l')$ tend to $F(x_l)$ and to $F(x_l')$ respectively, shows readily that for each fixed l, one has $Y_n(t) \in [x_l, x_l']$ for all large enough n, which shows that $Y_n(t) \to x = Y(t)$.
- Let us first denote by $\mathcal{T}$ the set of t such that $t = F(x_1) = F(x_2)$ for two different values of x_1 and x_2. This means that F is constant and equal to t on the interval (x_1, x_2) that contains (many) rational numbers, so that we conclude that $\mathcal{T}$ is at most countable.
- Let us denote by $\mathcal{T}'$ the set of t such that there exists a point of discontinuity x of F such that $t \in \{F(x-), F(x)\}$. Since the set of discontinuity points of F is at most countable, the set $\mathcal{T}'$ is at most countable as well.
- Suppose finally that $t \in (0,1)$ is such that $t = F(x)$ for some $x \in R$ and that $t \notin \mathcal{T} \cup \mathcal{T}'$. Since $t \notin \mathcal{T}$, we see that F is strictly increasing at x (meaning that $F(x') < F(x) < F(x'')$ whenever $x' < x < x''$). Since $t \in \mathcal{T}'$, we know that F is continuous at x.

  Then for all $\varepsilon > 0$, using the strict increase, the continuity of F at x and the density of points of continuity of F, we can find other points of continuity x_- and x_+ of F in $(x-\varepsilon, x)$ and $(x, x+\varepsilon)$ respectively such that

  $$F(x-\varepsilon) < F(x_-) < F(x) < F(x_+) < F(x+\varepsilon).$$

  Actually, since the union of the set of discontinuity points of F_n is also at most countable, we can choose such x_- and x_+ so that they are continuity points for each F_n. Since $F_n(x_-)$ and $F_n(x_+)$ converge to $F(x_-)$ and $F(x_+)$ respectively, we get that for all large enough n, $F(x-\varepsilon) < F_n(x_-) < F(x) = t < F_n(x_+) < F(x+\varepsilon)$, so that in particular (since Y_n is non-decreasing)

  $$x_- = Y_n(F_n(x_-)) \leq Y_n(t) \leq Y_n(F_n(x_+)) = x_+.$$

  So, for all $\varepsilon > 0$, wee see that for all large enough n, $Y_n(t) \in (x-\varepsilon, x+\varepsilon)$.

Hence we can conclude that for almost every t (with respect to the Lebesgue measure on $(0,1)$), $Y_n(t) \to Y(t)$, from which the result follows. □

7.3. Tightness and compactness

Let us consider a sequence $(P_n)_{n\geq 1}$ of probability measures in $\mathbb{R}$. The question that we are going to address is under which conditions there exists a subsequence $n_k \to \infty$, such that P_{n_k} converges weakly as $k \to \infty$.

REMARK 7.3.1. *The first example to have in mind is that when P_n is the Dirac mass at n, then it won't be possible to find a probability measure P such that a subsequence P_{n_k} converges weakly to P. Indeed, for any x_0 and any continuous bounded non-negative function f that is equal to 1 on $(-\infty, x_0]$ and to 0 on $[x_0+1, \infty)$, one would have*

$$P((-\infty, x_0]) \leq \int f(x)dP(x) = \lim_{n\to\infty} \int f(x)dP_{n_k}(x) = f(n_k) \to 0$$

as $k \to \infty$, so that (letting then x_0 tend to infinity), $P(\mathbb{R}) = 0$, which is a contradiction.

To avoid this "loss of mass to infinity" phenomenon, one is led to the following definition:

DEFINITION 7.3.2. *A family of probability measures $(P_i)_{i\in I}$ on a metric space (E, d) is tight if for every $\varepsilon > 0$, one can find a compact set C, such that for all $i \in I$, $P_i(C) \geq 1 - \varepsilon$.*

In the case a family $(P_i)_{i\in I}$ of probability measures on $\mathbb{R}$, this means that:

DEFINITION 7.3.3. *A family of probability measures $(P_i)_{i\in I}$ on a metric space $\mathbb{R}$ is tight if for every $\delta > 0$, one can find $K > 0$, such that for all $i \in I$, $P_i([-K, K]) \geq 1 - \delta$.*

[for the implication in one direction, take $\varepsilon = \delta$, and for the implication in the other direction, take $\delta = \varepsilon/2$].

We are now ready to state the main result of this section:

PROPOSITION 7.3.4. *Any tight sequence of family measures $(P_n)_{n\geq 1}$ on $\mathbb{R}$ has a subsequence that converges weakly.*

REMARK 7.3.5. *This statement in fact holds also for sequences of probability measure in complete separable metric spaces. It is known as Prokhorov's theorem.*

PROOF. We are going to build on the previous description of weak convergence via distribution functions.

Let $(x_q)_{q\geq 1}$ denote any given dense sequence in $\mathbb{R}$. The fact that any sequence in $[0, 1]$ has a subsequential limit, and the "diagonal trick" show that one can find an increasing sequence $(N_k)_{k\geq 1}$ such that for all q, $F_{N_k}(x_q)$ converges as $k \to \infty$. [Define the increasing sequences $(n_k(j))_{j\geq 1}$ by induction over k so that for all fixed k, $F_{n_1\circ\ldots\circ n_k(j)}(x_k)$ converges in $[0, 1]$, and then choose finally $N_k = n_1 \circ \ldots \circ n_k(k)$].

Let $\tilde{F}(x_q) := \lim_{k\to\infty} F_{N_k}(x_q)$ and then define

$$F(x) := \inf\{\tilde{F}(x_q) \ : \ x_q > x\}.$$

The function F is clearly right-continuous and non-decreasing on $\mathbb{R}$. We can also note that for all q, $\tilde{F}(x_q) \leq F(x_q) \leq 1$.

Suppose now that x is a continuity point of F. For all ε, we can find $x_q < x_{q'} < x < x_{q''}$ so that $F(x) - \varepsilon < F(x_q) \leq F(x_{q'}) \leq F(x) \leq F(x_{q''}) < F(x) + \varepsilon$. By the previous convergence result, we then get for all k large enough,

$$F_{N_k}(x) \leq F_{N_k}(x_{q''}) \leq \tilde{F}(x_{q''}) + \varepsilon \leq F(x_{q''}) + \varepsilon \leq F(x) + 2\varepsilon.$$

Similarly, for all large enough k,

$$F_{N_k}(x) \geq F_{N_k}(x_{q'}) \geq \tilde{F}(x_{q'}) - \varepsilon \geq F(x_q) - \varepsilon \geq F(x) - 2\varepsilon.$$

We therefore see that for all continuity point x of F, $F_{N_k}(x) \to F(x)$.

Finally, we note that for all $\varepsilon > 0$, tightness ensures that one can find K so that for all $x < -K$ and all $x' > K$,

$$F_n(x) \leq \varepsilon \text{ and } F_n(x') \geq 1 - \varepsilon.$$

So in particular, we see that if x_0 and x_0' are points of continuity of F in $(-\infty, -K)$ and (K, ∞) respectively, then $F(x_0) \leq \varepsilon$ and $F(x_0') \geq 1 - \varepsilon$, from which it follows $\lim_{-\infty} F = 0$ and $\lim_{+\infty} F = 1$, so that F is indeed the distribution function from some probability measure P.

The previous criterion for weak convergence via distribution functions allows to conclude. □

7.4. Characteristic functions

7.4.1. Definition. A very useful tool to study convergence of probability measures on $\mathbb{R}$ are their characteristic functions:

DEFINITION 7.4.1. *Suppose that P is a probability measure on $\mathbb{R}$. Its characteristic function φ_P is the function from $\mathbb{R}$ into $\mathbb{C}$ defined by*

$$\varphi_P(\theta) = \int e^{i\theta x} dP(x).$$

If P is the law of a random variable X, then $\varphi(\theta) = E[e^{i\theta X}]$. We also call this the *characteristic function of* X and then write $\varphi_X = \varphi_P$.

REMARK 7.4.2. *One can recognize that (possibly up to normalization by a constant), this is nothing else than the Fourier transform of the measure P.*

Here are some obvious properties of φ_P:

- $\varphi(0) = 1$ and $|\varphi(\cdot)| \leq 1$.
- φ is continuous (just use dominated convergence).
- If the law of X and $-X$ are the same, then φ_X is real-valued.
- For all real constant λ, one has $\varphi_{\lambda X}(\theta) = \varphi_X(\lambda\theta)$.

REMARK 7.4.3. *It is easy to see (using the usual "differential under the integral" theorems) that if X is a random variable in L^p for some integer $p \geq 1$, then φ is p-times differentiable and that*

$$\varphi^{(p)}(\theta) = E[(iX)^p \exp(i\theta X)],$$

and in particular,

$$\varphi^{(p)}(0) = i^p E[X^p].$$

But if X is not in L^p, then φ may not be p-times differentiable.

Let us list some examples of characteristic functions that have a nice expression:

1. An important example in what follows is the case of the standard Gaussian distribution, with density $(2\pi)^{-1/2}\exp(-x^2/2)$ on $\mathbb{R}$. What is so special about this distribution is that its characteristic function is $\exp(-\theta^2/2)$. We will comment on how to compute this below.

   A bit more generally, when the law of a random variable X is the standard Gaussian distribution, then for all $\sigma > 0$, we call the law of σX a centered Gaussian distribution with variance σ^2, and we denote it by $\mathcal{N}(0, \sigma^2)$. Its characteristic function is then clearly $\exp(-\theta^2\sigma^2/2)$.
2. The Poisson distribution with parameter $\lambda > 0$, defined on $\mathbb{N}$ by $P(\{n\}) = e^{-\lambda}\lambda^n/n!$. Its characteristic function is then $\exp(\lambda(e^{i\theta} - 1))$.
3. The Cauchy distribution with density $dx/(\pi(1 + x^2))$ on $\mathbb{R}$. Its characteristic function turns out to be
$$\int_{\mathbb{R}} \frac{e^{i\theta x}}{\pi(1 + x^2)} = \exp(-|\theta|)$$

Let us now justify the previous explicit expressions: The computation of the characteristic function of the Poisson distribution is pretty straightforward:

$$\sum_{n\geq 0} e^{-\lambda}\frac{\lambda^n}{n!}e^{i\theta n} = e^{-\lambda}\sum_{n\geq 0}\frac{(\lambda e^{i\theta})^n}{n!} = e^{-\lambda}e^{\lambda\exp(i\theta)}.$$

The computations of characteristic functions of distributions with densities can be a little bit trickier: One trick that works well for both the Gaussian and the Cauchy distribution is to use contour integrals and the residue theorem. For the standard Gaussian distribution, one first notices that

$$E[e^{i\theta X}] = (2\pi)^{-1/2} \int_{\mathbb{R}} e^{i\theta x - x^2/2} dx = (2\pi)^{-1/2} e^{-\theta^2/2} \int_{\mathbb{R}} e^{-(x-i\theta)^2/2} dx,$$

and then using the fact that the contour integral of $\exp(-z^2/2)dz$ over the boundary of the rectangle $[-R, R] \times [0, \theta]$ is 0 and then letting $R \to \infty$, one gets readily that

$$\int_{\mathbb{R}} e^{-(x-i\theta)^2/2} dx = \int_{\mathbb{R}} e^{-x^2/2} dx$$

which allows to conclude.

For the Cauchy distribution, by symmetry (since we know the characteristic function is even), we note that it suffices to consider the case where θ is positive. We then choose R large and to look at the contour integral of $\exp(i\theta z)dz/(\pi(1+z^2))$ along the boundary of the upper semi-disk of radius R. The residue theorem tells us readily that the value is $\exp(i^2\theta) = \exp(-\theta)$. Letting $R \to \infty$ (and noting that the contribution of the contour integral on the real axis converges to $\varphi(\theta)$ while the contribution of integral over the half-circle tends to 0), we can conclude.

7.4.2. Inversion formula. An important property is that:

Proposition 7.4.4. *If two probability measures on $\mathbb{R}$ have the same characteristic function, then they are the same measures.*

In fact, it is possible to explicitely reconstruct the distribution function F of a probability measure out of its characteristic function:

Proposition 7.4.5 (Inversion formula). *If F and φ are respectively the distribution function and the characteristic function of a probability measure P, then for all $a < b$,*

$$\lim_{T \to +\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ia\theta} - e^{-ib\theta}}{i\theta} \varphi(\theta) d\theta = F^{\#}(b) - F^{\#}(a),$$

where $F^{\#}(x) = (F(x) + F(x-))/2$.

This second proposition implies indeed the first one – letting $a \to -\infty$ shows that one can recover $\tilde{F}$ from φ, and then, since the discontinuity points of F and $F^{\#}$ are the same, it is easy to recover F at all continuity points, and therefore F – and finally we conclude because F determines P.

Proof of the inversion formula. Noting that for any two reals $|\exp(iu) - \exp(iv)| \leq |v - u|$ (the length of the cord is smaller than the length of the arc joining $\exp(iu)$ and $\exp(iv)$), we see that

$$\int_{-T}^{T} \int_{\mathbb{R}} d\theta dP(x) \left| \frac{e^{i(x-a)\theta} - e^{i(x-b)\theta}}{i\theta} \right| \leq 2T(b-a) < \infty,$$

so that we can apply Fubini's theorem and invert the order of integration in x and θ, and we get that

$$\int_{-T}^{T} \frac{e^{-ia\theta} - e^{-ib\theta}}{i\theta} \varphi(\theta) d\theta = \int_{\mathbb{R}} dP(x) \left(\int_{-T}^{T} \frac{e^{i(x-a)\theta} - e^{i(x-b)\theta}}{i\theta} d\theta \right).$$

Since cos is even and sin is odd, we get that this quantity is equal to

$$2 \int_{\mathbb{R}} dP(x) \left(\int_0^T \frac{\sin(x-a)\theta - \sin(x-b)\theta}{\theta} d\theta \right).$$

We can use now the classical fact (known as the Dirichlet integral – it can be proved using elementary analysis in a number of ways) that

$$\lim_{T\to\infty}\int_0^T \frac{\sin u}{u}du = \frac{\pi}{2}$$

(note in particular also that it is bounded uniformly with respect to T). Since for positive λ,

$$\int_0^{\lambda T}\frac{\sin u}{u}du = \int_0^T \frac{\sin \lambda u}{u}du,$$

the limit as $T \to \infty$ of

$$\int_0^T \frac{\sin(x-a)\theta - \sin(x-b)\theta}{\theta}d\theta$$

will depend only on the position of x relatively to a and b: It is $\pi/2 - \pi/2 = 0$ if $x \notin [a,b]$, it is $\pi/2 + \pi/2 = \pi$ if $x \in (a,b)$ and it is $\pi/2 + 0 = \pi/2$ if $x \in \{a,b\}$. We can then finally use the dominated convergence theorem (the integrand being bounded uniformly) to see get the desired result. □

To conclude this section, we can state and prove the following fact, which can be useful to have in our toolbox:

Proposition 7.4.6. *If $\int |\varphi(\theta)|d\theta < \infty$, then the law P has a continuous density f with respect to the Lebesgue measure on $\mathbb{R}$, and*

$$f(x) = \frac{1}{2\pi}\int e^{-i\theta x}\varphi(\theta)d\theta.$$

Proof. The proof is essentially just a direct application of the inversion formula, using the dominated convergence theorem a couple of times. Recall that

$$F^{\#}(b) - F^{\#}(a) = \lim_{T\to+\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ia\theta} - e^{-ib\theta}}{i\theta}\varphi(\theta)d\theta.$$

Since $\int |\varphi(\theta)|d\theta < \infty$, we can simply write

$$F^{\#}(b) - \tilde{F}^{\#}(a) = \frac{1}{2\pi}\int_{\mathbb{R}}\frac{e^{-ia\theta} - e^{-ib\theta}}{i\theta}\varphi(\theta)d\theta.$$

By dominated convergence again, we see that the right-hand side is continuous with respect to b on (a,∞) (for any a), from which it follows that $F^{\#}$ is continuous in $\mathbb{R}$, which in turn implies that F is actually continuous and that $F = F^{\#}$.

Then, we can look at

$$\frac{F(b)-F(a)}{b-a} = \frac{1}{2\pi}\int_{\mathbb{R}}\frac{e^{-ia\theta} - e^{-ib\theta}}{i\theta(b-a)}\varphi(\theta)d\theta.$$

By dominated convergence again (we leave the details as an exercise), letting $b \to a+$ or $a \to b-$, we can then easily check that F is differentiable on $\mathbb{R}$ and that its derivative is indeed the continuous function $a \mapsto (2\pi)^{-1}\int_{\mathbb{R}} e^{-i\theta a}\varphi(\theta)d\theta$. □

7.4.3. Characteristic functions and independence. An important and simple observation is the following:

PROPOSITION 7.4.7. *If $X_1, \ldots, X_n$ are independent random variables defined on the same probability space, then*

$$\varphi_{X_1+\ldots+X_n}(\theta) = \prod_{j=1}^{n} \varphi_{X_j}(\theta).$$

PROOF. This is just due to the fact that the random variables $\exp(i\theta X_1), \ldots, \exp(i\theta X_n)$ are independent (and bounded) random variables with values in $\mathbb{C}$ (one can write $\exp(iy) = \cos(y) + i\sin(y)$ and expand the product if one does not feel at ease with the complex multiplications here). □

This can be a very useful tool to actually determine the law of the sum of some independent random variables, in particular in the case of random variables with a density where the computations via density functions can be cumbersome. Basically, if one knows that characteristic functions of $X_1, \ldots, X_n$, then the previous result gives us automatically the characteristic function of $X_1 + \cdots + X_n$, and if one happens to recognize that characteristic function as that of a known law, then (since characteristic functions determine the law), one knows the law of $X_1 + \cdots + X_n$.

For example: *If X and Y are two independent centered Gaussian random variables with respective variances σ_X^2 and σ_Y^2, then $X + Y$ is a centered Gaussian random variable with variance $\sigma_X^2 + \sigma_Y^2$.* Indeed,

$$\varphi_{X+Y}(\theta) = \varphi_X(\theta)\varphi_Y(\theta) = \exp(-\theta^2(\sigma_X^2 + \sigma_Y^2))$$

which is the characteristic function of a centered Gaussian variable with variance $\sigma_X^2 + \sigma_Y^2$ and we can conclude noting that the characteristic function characterizes the law.

Similarly, one gets that if X and Y are independent random variables with standard Cauchy distributions, then the law of $(X + Y)/2$ is also a standard Cauchy distribution.

A similar argument can be used show that if X and Y are independent Poisson random variables with respective parameters λ_Y and λ_Y, then $X + Y$ is a Poisson random variable with parameter $\lambda_X + \lambda_Y$. But this fact could have been derived immediately looking at $P(X + Y = n) = \sum_{j=0}^{n} P(X = j)P(Y = n - j)$.

7.5. Weak convergence via characteristic functions

Clearly, since (for all fixed θ), $x \mapsto e^{i\theta x}$ is a bounded continuous function, when P_n converges weakly to P, then for all $\theta \in \mathbb{R}$, one has $\lim_{n\to\infty} \varphi_n(\theta) = \varphi(\theta)$. We will now discuss results in the other direction: Does the convergence of φ_n suffice to obtain weak convergence?

Proposition 7.5.1 (Lévy's theorem). *Let $(P_n)_{n\geq 1}$ denote a sequence of probability measures on $\mathbb{R}$, such that for all $\theta \in \mathbb{R}$, the sequence $\varphi_n(\theta)$ converges to some number $\psi(\theta)$. Assume furthermore that $\theta \mapsto \psi(\theta)$ is continuous at $\theta = 0$. Then, ψ is the characteristic function of a probability measure P, and P_n does converge weakly to P.*

Proof. The first key step is to show that the continuity at 0 of ψ implies that the sequence P_n is tight. For this, we choose $\varepsilon > 0$, and we want to K so that for all n, $P_n([-K, K]) \geq 1 - \varepsilon$.

First, using the continuity at 0 of ψ, we can find $\delta > 0$ so that for all $\theta \in [-\delta, \delta]$, one has $|\psi(\theta) - 1| < \varepsilon/4$, and therefore $|\psi(\theta) + \psi(-\theta) - 2| < \varepsilon/2$.

We note that for each n, $\varphi_n(\theta) + \varphi_n(-\theta) = \int 2\cos(\theta x) dP_n$ is real-valued and takes its values in $[-2, 2]$, so the same is true for its limit $\psi(\theta) + \psi(-\theta)$. By dominated convergence,

$$\int_0^\delta (2 - \varphi_n(\theta) - \varphi_n(-\theta)) d\theta \to \int_0^\delta (2 - \psi(\theta) - \psi(-\theta)) d\theta \leq \delta\varepsilon/2$$

as $\delta \to 0$, so that there exists n_0 such that for all $n \geq n_0$,

$$\int_0^\delta (2 - \varphi_n(\theta) - \varphi_n(-\theta)) d\theta \leq \delta\varepsilon.$$

But on the other hand, by Fubini (since the absolute value of the integrand is bounded here and all measures involved are finite)

$$\int_0^\delta (2 - \varphi_n(\theta) - \varphi_n(-\theta)) d\theta = \int_{\mathbb{R}} \left(\int_0^\delta (2 - 2\cos(\theta x)) d\theta \right) dP_n(x) = \int_{\mathbb{R}} (2\delta - 2\frac{\sin(\delta x)}{x}) dP_n(x).$$

Dividing everything by 2δ, we therefore get that for all $n \geq n_0$,

$$\int_{\mathbb{R}} (1 - \frac{\sin(\delta x)}{\delta x}) dP_n(x) \leq \varepsilon/2.$$

Note that the function $y \mapsto \sin(\delta y)/(\delta y)$ tends to 1 when $y \to 0+$ or $y \to 0-$ and is otherwise in $(-1, 1)$ on $\mathbb{R} \setminus \{0\}$ (simply because $|\sin(u)| \leq |u|$ as the derivative of sin is in $[-1, 1]$). In particular, when $|y| > 2/\delta$, it is in $(-1/2, 1/2)$ and therefore $1 - \sin(\delta y)/(\delta y) \geq 1/2$. Hence, for all $n \geq n_0$

$$P_n(\mathbb{R} \setminus [-2/\delta, 2/\delta)]) \leq 2 \int_{|x|>2/\delta} (1 - \frac{\sin(\delta x)}{\delta x}) dP_n(x) \leq 2 \int_{\mathbb{R}} (1 - \frac{\sin(\delta x)}{\delta x}) dP_n(x) \leq \varepsilon.$$

Finally, we need to obtain a bound for the probability measures P_n for $n \leq n_0$: We can choose $K \geq 2/\delta$ such that for all $n = 1, \ldots, n_0$, $P_n([-K, K]) \geq 1 - \varepsilon$, and (since $K > 2/\delta$), we get that for all $n \geq 1$, $P_n([-K, K]) \geq 1 - \varepsilon$. which completes the proof of the fact that $(P_n)_{n\geq 1}$ is tight.

Next, since $(P_n)_{n\geq 1}$ is tight, we can extract a subsequence such that P_{n_k} converges weakly to some probability measure P, and then φ_{n_k} converges to the characteristic function of P. Since φ_n does converge also to ψ, we conclude that ψ is the characteristic function of P.

If P_n did not converge weakly to P, then it means that for some bounded continuous f_0, and some subsequence m_k, $\int f_0(x) dP_{m_k}(x)$ remains bounded away from $\int f_0(x) dP(x)$. But by tightness, we can find a sub-subsequence that converges weakly to some probability measure Q, and by the same argument as before, we conclude that the characteristic function of Q has to be ψ. But since the characteristic functions determine the probability measure, we get that $Q = P$, which leads to a contradiction (since $\int f_0(x) dP_{m_k}(x)$ does then not stay bounded away from $\int f_0(x) dP(x)$). □

7.6. A first consequence: The Central Limit Theorem

Suppose now that $(X_n)_{n\geq 1}$ is a sequence of independent identically distributed random variables that are in L^2. We suppose the $E[X_j] = 0$ and we denote the variance of X_1 by σ^2 (here $\sigma^2 = E[X_1^2]$ since $E[X_1] = 0$). We assume that $\sigma^2 \neq 0$, i.e., that $P(X_1 = 0) \neq 1$. We write $S_n := X_1 + \cdots + X_n$.

We know from the law of large numbers that S_n/n tends to 0 almost surely when $n \to \infty$. It is a natural question to ask what the actual order of magnitude of S_n is when n is large. A first indication that it will be of the order of $\sqrt{n}$ is that

$$E[S_n^2] = E[X_1^2] + \cdots + E[X_n^2] = n\sigma^2.$$

The central limit theorem gives a rather precise answer here:

THEOREM 7.6.1 (Central limit theorem). *The law of $S_n/(\sigma\sqrt{n})$ converges weakly to a standard Gaussian distribution $\mathcal{N}(0,1)$.*

In the same line of thought as Remark 7.4.3, we can establish the following lemma:

LEMMA 7.6.2. *Suppose that X is a random variable in L^2 such that $E[X] = 0$ and $E[X^2] = \sigma^2$. Then, as $\theta \to 0$,*

$$\varphi_X(\theta) = 1 - \frac{\sigma^2\theta^2}{2} + o(\theta^2).$$

PROOF OF THE LEMMA. To prove this, we can first note that for all $x \in \mathbb{R}$, Taylor's formula with integral rest (using the fact that the second and third derivatives of $\exp(iz)$ are $i^2\exp(iz)$ and $i^3\exp(iz)$ which have a modulus bounded by 1) gives readily

$$|\exp(ix) - (1+ix)| \leq \int_0^x ydy = |x|^2/2 \text{ and } |\exp(ix) - (1 + ix - x^2/2)| \leq \int_0^{|x|} y^2dy/2 = |z|^3/6,$$

so that for all $x \in \mathbb{R}$,

$$|\exp(iz) - (1 + iz - z^2/2)| \leq \min(|z|^2/2 + |z|^2/2, |z|^3/6).$$

Hence, we readily get that

$$|\varphi_X(\theta) - (1 - \frac{\sigma^2\theta^2}{2})| \leq \theta^2 E[\min(X^2, \theta X^3)].$$

By dominated convergence (since $E[X^2] < \infty$), we have $E[\min(X^2, \theta X^3)] \to 0$ as $\theta \to 0$, which concludes the proof. □

PROOF OF THE CLT. Let $Y_n = S_n/(\sigma\sqrt{n})$. Let φ_n denote the characteristic function of Y_n, and let φ denote the characteristic function of X_1. By Lévy's characterization theorem, we see that it is sufficient to prove that for all θ, as $n \to \infty$, $\varphi_n(\theta) \to \exp(-\theta^2/2$, which is characteristic function of the standard Gaussian.

By the previous remark, we know that

$$\varphi(\theta) = 1 - \frac{\sigma^2\theta^2}{2} + o(\theta^2)$$

as $\theta \to 0$. But the definition of Y_n and the independence between $X_1, \ldots, X_n$ then shows that

$$\varphi_n(\theta) = \varphi(\theta/(\sigma\sqrt{n}))^n = (1 - \frac{\theta^2}{2n} + o(\theta^2/n))^n.$$

Letting $n \to \infty$, we therefore get that for every given $\theta \in \mathbb{R}$,

$$\log\varphi_n(\theta) = n\log(1 - \frac{\theta^2}{2n} + o(\theta^2/n)) = -n \times \frac{\theta^2}{2n} + o(1) \to -\frac{\theta^2}{2}$$

and therefore that indeed, $\varphi_n(\theta) \to \exp(-\theta^2/2)$ as $n \to \infty$. □

Remark 7.6.3. *If $(Y_n)_{n \geq 1}$ is any sequence of independent identically distributed (but not necessarily centered) random variables in L^2, we can apply the previous result to $X_n = Y_n - m$ where $m = E[Y_1]$, and we then see that the law of of $(Y_1 + \cdots + Y_n - nm)/\sqrt{n}$ converges to a centered Gaussian with variance Var(Y_1) (with the convention that a centered Gaussian with variance 0 is the Dirac mass at 0).*

Remark 7.6.4. *We can note that the condition that X is in L^2 actually implies readily that for any $\varepsilon > 0$,*

$$P(\max(|X_1|, \ldots, |X_n|) \geq \varepsilon\sqrt{n}) \leq nP(X_1^2 \geq \varepsilon^2 n) \leq \varepsilon^{-2} E[X_1^2 1_{X_1^2 \geq \varepsilon^2 n}] \to 0$$

as $n \to \infty$ (by dominated convergence). So, one can heuristically say that (with high probability when n is large), none of the individual terms X_j for $j \leq n$ will be of the same order of magnitude than $X_1 + \cdots + X_n$ – which is $\sqrt{n}$). This will contrast the results in the next section.

7.7. Another consequence of Lévy's theorem: The symmetric stable distributions

Lévy's theorem can for instance be used to show the *existence* of distributions on $\mathbb{R}$ with nice properties: Let us recall that the characteristic function of the standard Gaussian distribution is $\exp(-\theta^2/2)$ and that the characteristic function of the Cauchy distribution is $\exp(-|\theta|)$. By considering multiples of these random variables, we see that $\exp(-c|\theta|^\alpha)$ is the characteristic function of some probability measure for all $c > 0$ when $\alpha = 1$ and $\alpha = 2$. More generally:

Definition 7.7.1. *We say that if for $\alpha > 0$ and $c > 0$, $\exp(-c|\theta|^\alpha)$ is the characteristic function of some probability measure P, then P is a* symmetric stable distribution of index *α.*

We can note that if P is a symmetric stable then:

(1) The law P is symmetric with respect to 0 (i.e., if the law of a random variable X is P, then the law of $-X$ is P as well – this is just because the characteristic function is even).
(2) For any $N \geq 2$, there exists $A(N) > 0$ such that if $X_1, \dots, X_N$ are independent random variables with law P, then the law of $X_1 + \cdots + X_N$ is the same as that of $A(N)X_1$. This is a rather remarkable property.

The second fact is due to the fact that

$$\varphi_{X_1+\cdots+X_N}(\theta) = \varphi_{X_1}(\theta)^N = \varphi_{X_1/N^{1/\alpha}}(\theta)$$

so that $X_1 + \cdots + X_N$ has the same law as $N^{1/\alpha}X_1$, so that $A(N) = N^{1/\alpha}$.

A first observation is that symmetric stable laws with index greater than 2 can not exist. To see this one can first check that if some random variable Z is not in L^2 (so that $E[Z^2] = \infty$), then necessarily (by Fatou's lemma for instance),

$$\frac{2 - \varphi_Z(\theta) - \varphi_Z(-\theta)}{\theta^2} = \int \frac{2 - 2\cos(\theta x)}{\theta^2} dP(x) \to \infty$$

as $\theta \to 0$ (noting that the integrand is non-negative and tends to x^2 as $\theta \to 0$). But if $\varphi_Z = \exp(-|\theta|^\alpha)$ for $\alpha > 2$, then the limit is finite, so that the random variable has to be in L^2 and has a finite variance σ^2 (which is not zero because X is symmetric and not equal to δ_0) and it also has zero expectation (by symmetry). By the argument used in the previous section, this would then imply that when $\theta \to 0$, $\varphi(\theta) = 1 - \sigma^2\theta^2/2 + o(\theta^2)$ which contradicts the fact that $\varphi(\theta) - 1 \sim c|\theta|^\alpha$ as $\theta \to 0$. (Another way to see the contradiction once one knows that the law has a finite variance σ^2 is to say that on the one hand, the variance of the sum of two independent copies of would be $2\sigma^2$, and on the other hand, the variance of $2^{1/\alpha}X$ is $2^{2/\alpha}\sigma^2$, which is not possible if $\alpha > 2$).

The main goal of the present section is to show using Lévy's theorem that:

Proposition 7.7.2. *Symmetric stable distributions of indey α exist for every $\alpha \in (0, 2]$.*

Proof. We already know that this is true when $\alpha = 2$ (just consider a Gaussian centered random variable). Let us consider $\alpha \in (0, 2)$ and a sequence $(Y_n)_{n\geq 1}$ of independent identically distributed random variables with density on $\mathbb{R}$ given by

$$1_{\{|x|>C\}}|x|^{-\alpha-1},$$

where C is chosen so that the integral of this function is 1.

We can note that the characteristic function φ_Y of this law is then

$$\varphi_Y(\theta) = \int_C^\infty (e^{it\theta} + e^{-it\theta})\frac{dt}{t^{\alpha+1}} = 2|\theta|^\alpha \int_{C|\theta|}^\infty \cos(v)\frac{dv}{v^{\alpha+1}}.$$

Note that the definition of C shows that

$$2|\theta|^\alpha \int_{C|\theta|}^\infty \frac{dv}{v^{\alpha+1}} = 2\int_C^\infty \frac{x}{x^{\alpha+1}} = 1.$$

When $\theta \to 0$, we therefore see that

$$1 - \varphi_Y(\theta) = 2|\theta|^\alpha \int_{C|\theta|}^\infty (1-\cos(v)) \frac{dv}{v^{\alpha+1}} \sim c_\alpha |\theta|^\alpha$$

where

$$c_\alpha := 2 \int_0^\infty (1-\cos u) \frac{du}{u^{\alpha+1}}$$

(note that $1 - \cos u \sim u^2/2$ when $u \to 0$, so that integral indeed converges near the origin).

Now we let for each n,

$$X_n := n^{-1/\alpha}(Y_1 + \cdots + Y_n).$$

The characteristic function of X_n is then

$$\varphi_{X_n}(\theta) = \left(\varphi_Y(\theta/n^{1/\alpha})\right)^n = \left(1 - c_\alpha \frac{|\theta|^\alpha}{n} + o(\frac{|\theta|^\alpha}{n})\right)^n.$$

Hence, as $n \to \infty$, we get that for all θ,

$$\varphi_{X_n}(\theta) \to \exp(-c_\alpha |\theta|^\alpha).$$

The function $\theta \mapsto \exp(-c_\alpha|\theta|^\alpha)$ is continuous at 0, so we can apply Lévy's theorem: It is indeed the characteristic function of some distribution in $\mathbb{R}$. By scaling, we can get rid of the constant c_α, so that $\exp(-|\theta|^\alpha)$ is also the characteristic function of some probability distribution. □

Remark 7.7.3. *This time, we can note that for any positive y,*

$$P(\max(|Y_1|, \ldots, |Y_n|) \le yn^{1/\alpha}) = P(|Y_1| \le yn^{1/\alpha})^n \le (1 - cy^{-\alpha}/n)^n \to \exp(-cy^{-\alpha})$$

for some constant c. In particular, the limsup of these probabilities is strictly smaller than 1. This shows that (with high probability when n is large), the largest of the values $|Y_j|$ for $j \le n$, will be of the same order of magnitude than $Y_1 + \cdots + Y_n$, which contrasts with the case of the sums of i.i.d. variables that are in L^2 (as discussed at the end of the previous section).

Let us now briefly comment on the following fact. Suppose that P is a probability measure that satisfies the properties (1) and (2) stated just after the definition of symmteric stable distribution. In particular, when X_1 and X_2 are independent and both have the law P, then for some $A \ge 0$, the law of $X_1 + X_2$ is identical to the law of AX_1. If $A > 0$, there exists then $\alpha > 0$ such that $A = 2^{1/\alpha}$. The characteristic function of P is real-valued, even and continuous, and satisfies

$$\varphi(\theta) \times \varphi(\theta) = \varphi(2^{1/\alpha}\theta)$$

for all $\theta > 0$. This means that $\Phi(x) := \log(\varphi(x^{1/\alpha}))$ satisfies $\Phi(2x) = 2\Phi(x)$ and is continuous on $[0, \infty)$. It is actually possible to show that Φ is necessarily linear i.e., that:

Proposition 7.7.4. *If a probability measure P other than the Dirac mass at the origin satisfies properties (1) and (2), then there exists $\alpha > 0$ and $c > 0$, such that $\varphi_P(\theta) = \exp(-c|\theta|^\alpha)$.*

In other words, the symmetric stable laws with $\alpha \in (0, 2]$ and the Dirac mass at the origin are the only distributions that satisfy the two properties (1) and (2).

Rough idea of the proof. A first step in the proof is to check that necessarily, for every n, one has $A(n) = n^{1/\alpha}$. The same argument as for $n = 2$ then shows that for any $n \ge 1$, $\Phi(nx) = n\Phi(x)$ for any n, which in turn is then sufficient to conclude that Φ has to be linear. □

Remark 7.7.5. *It is in fact possible (but we will not do this here, to characterize the probability measures that satisfy only (2) (i.e., one removes the symmetry assumption).*

7.8. Weak convergence vs. almost sure convergence of random variables

We now make some comments on the relation between convergence of random variables and convergence of their laws. We have deliberately postponed these comments to this later part of the chapter to avoid any confusion.

Let us comment on the proof of Proposition 7.2.1. We have shown that if P_n converges weakly to P, then F_n converges to F at every point of continuity of F. But then, we can use the second part of the proof, where we have constructed on the space $(0, 1)$ endowed with the Lebesgue measure, a sequence of random variables (Y_n) and a random variable Y such that Y_n converges almost surely to Y, and the law of Y_n is P_n (for each n), while the law of Y is P. Hence, this proves the following statement:

Proposition 7.8.1 (Skorokhod's representation theorem). *If a sequence of probability measures $(P_n)_{n\geq 1}$ on $\mathbb{R}$ converges weakly to a probability measure P, then it is possible to construct a probability space and in this probability space, a sequence of random variables $(Y_n)_{n\geq 1}$ (so that for each $n \geq 1$, the law of Y_n is P_n) and a random variable X with law P, such that $Y_n \to Y$ almost surely.*

On the other hand, if a sequence of random variables $(Z_n)_{n\geq 1}$ converges almost surely to some random variable Z as $n \to \infty$, then for any bounded continuous function f, $f(Z_n) \to f(Z)$ almost surely, so that by dominated convergence, the law of Z_n converges weakly to the law of Z.

But the previous proposition is very far from a converse to this last fact. If $(X_n)_{n\geq 1}$ is a sequence of random variables on the same probability space, such that the law of X_n converges weakly to the law of another random variable X in that space, this does not imply at all that X_n converges in probability or almost surely to some random variable. In many important examples (for instance in the central limit theorem), this will *not* be the case. The simplest counterexample is for instance when $(X_n)_{n\geq 1}$ is a sequence of independent identically distributed random variables (when the law is not a Dirac mass).

There is just one special case where such a converse statement holds: When X_n converges in law to the Dirac mass at some point a, then X_n converges in probability to a (see the final exercise of Exercise sheet 10).

Remark 7.8.2. *When $(X_n)_{n\geq 1}$ is a sequence of random variables such that the law of X_n converges weakly as $n \to \infty$, one sometimes says "X_n converges in distribution".*

7.9. Extensions to random vectors (survey)

Let us now explain how to generalize the previous statements (that we have obtained for probability measures on $\mathbb{R}$ in this chapter) to the case of probability measures on $\mathbb{R}^d$. We will not give full details here, this is more a survey type section – we will in particular not explain at all how to derive the generalization of Prokhorov's theorem. The content of this section will *not* be discussed in the exam.

(1) (Laws in $\mathbb{R}^d$, basics). A probability measure on $\mathbb{R}^d$ corresponds to the law of a random vector $(X_1, \ldots, X_d)$, or equivalently to the joint law of d random variables. We recall that a probability measure P on $\mathbb{R}^d$ is fully determined by its values on a π-system that generates the Borel σ-field on $\mathbb{R}^d$. In particular, this shows that P is fully determined by the knowledge of

$$F(a_1, \ldots, a_d) := P((-\infty, a_1] \times \cdots \times (-\infty, a_d])$$

for $a_1, \ldots, a_d \in \mathbb{R}$. Mind that for $d \geq 2$, we de not have derived a result that says which functions F do actually correspond to a probability measure P.

(2) (Covariance matrix). When P is the law of a random vector $(X_1, \ldots, X_d)$ that is in L^2 (i.e., each of the X_j for $j \leq d$ is in L^2), then one can define its covariance matrix $S = (S_{k,l})_{k,l \leq d}$ by $S_{k,l} := E[X_k X_l] - E[X_k]E[X_l]$. It is easy to check that the covariance matrix of X is the same as the covariance matrix of $(\tilde{X}_1, \ldots, \tilde{X}_d) := (X_1 - E[X_1], \ldots, X_d - E[X_d])$ because

$$E[(X_k - E[X_k])(X_l - E[X_l])] = E[X_k X_l] - (2-1)E[X_k]E[X_l].$$

A covariance matrix is always symmetric (this is obvious from its definition) and positive because for each $\lambda_1, \ldots, \lambda_d$,

$$\sum_{k,l} \lambda_k \lambda_l S_{k,l} = E[(\sum_{j=1}^{d} \lambda_j \tilde{X}_j)^2] \geq 0.$$

Conversely, if R is a symmetric $d \times d$ positive matrix (i.e., such that for all $\lambda_1, \ldots, \lambda_d$, $\sum_{k,l} \lambda_l \lambda_k R_{j,k} \geq 0$, then it is easy to see that it is the covariance matrix of some random vector. Indeed, one can first recall that it is possible to find a $d \times d$ matrix C such that $C \times C^t = R$, and then one chooses a collection $(N_1, \ldots, N_d)$ of independent random variables in L^2 with $E[N_j] = 0$ and $E[(N_j)^2] = 1$ for all $j \leq d$. One can then check that the vector $(X_1, \ldots, X_d) := C \times (N_1, \ldots, N_d)$ has R as covariance matrix.

(3) (Characteristic functions). Next, we can define the characteristic function $\varphi = \varphi_{X_1, \ldots, X_n}$ of such a random vector. It is this time the function defined on $\mathbb{R}^d$ by

$$\varphi(\theta_1, \ldots, \theta_d) := E[\exp(i(\theta_1 X_1 + \ldots + \theta_d X_d))].$$

We can note that by our result on characteristic functions for real-valued random variables, the knowledge of $\varphi_{X_1, \ldots, X_d}$ is the same as the knowledge of the law of $\alpha_1 X_1 + \ldots + \alpha_d X_d$ for each given $\alpha_1, \ldots, \alpha_d$ in $\mathbb{R}$. [Indeed, the knowledge of $E[\exp(i\theta(\alpha_1 X_1 + \ldots + \alpha_d X_d))]$ for each θ and $\alpha_1, \ldots, \alpha_d$ is the same as the knowledge of $\varphi(\theta_1, \ldots, \theta_d)$ for all $\theta_1, \ldots, \theta_d$.]

(4) (Inversion Formula). Actually, it turns out that this is also equivalent to the knowledge of the actual law of $(X_1, \ldots, X_d)$. This can be viewed as a consequence of the following fact (since it shows that a given function can be the characteristic function of at most one probability measure):

Proposition 7.9.1 (Inversion formula). *For all box $B = [a_1, b_1] \times \cdots \times [a_d, b_d]$ such that $P(\partial B) = 0$, one has*

$$\lim_{T \to +\infty} \frac{1}{(2\pi)^d} \int_{[-T,T]^d} \varphi(\theta_1, \dots, \theta_d) \left[\prod_{j=1}^{d} \frac{(e^{-ia_j\theta_j} - e^{-ib_j\theta_j})}{i\theta_j} \right] d\theta_1 \dots d\theta_d = P(B).$$

Proof. By Fubini,

$$\int_{[-T,T]^d} \varphi(\theta_1, \dots, \theta_d) \left[\prod_{j=1}^{d} \frac{(e^{-ia_j\theta_j} - e^{-ib_j\theta_j})}{i\theta_j} \right] d\theta_1 \dots d\theta_d$$
$$= \int dP(x_1, \dots, dx_d) \left[\prod_{j=1}^{d} \int_{-T}^{T} \frac{(e^{-ia_j\theta_j} - e^{-ib_j\theta_j})e^{i\theta_j x_j}}{i\theta_j} d\theta_j \right]$$

But we have seen that

$$\int_{-T}^{T} \frac{(e^{-ia_j\theta_j} - e^{-ib_j\theta_j})e^{i\theta_j x_j}}{i\theta_j} d\theta_j \to \pi(21_{(a_j,b_j)} + 1_{\{a_j,b_j\}}),$$

and we can then (as in the $d = 1$ case) conclude readily using the dominated convergence theorem. □

By letting each a_j tend to $-\infty$, we see that φ determines $P((-\infty, b_1] \times \ldots \times (-\infty, b_d])$ for all $b_1, \dots, b_d$ for which $P[X_j = b_j] = 0$ (for each $j \leq d$), which in turn (for instance using the right-continuity of F in each of the d variables separately) determines F everywhere, and therefore P.

(5) (Gaussian vectors). Suppose that $Z_1, \dots, Z_d$ are d independent standard Gaussian random variables (i.e., centered and with variance 1). Then, we know that for all $\beta_1, \dots, \beta_d$,

$$E[\exp(i(\beta_1 Z_1 + \ldots + \beta_d Z_d))] = \exp(-(\sum_{j=1}^{d} \beta_j^2)/2).$$

Let us consider any $d \times d$ matrix $C = (c_{k,l})_{k,l \leq d}$ and define the vector $X = (X_1, \dots, X_d) := Z \times C$ (using matrix product notation).

We can note as above that the covariance matrix of X is $S := C \times (C^t)$ (in matrix product notation). We can furthermore see that for all $\theta_1, \dots, \theta_d$,

$$E[\exp(i(\theta_1 X_1 + \ldots + \theta_d X_d))] = \exp(-\frac{1}{2} \sum_{k,l} \theta_l S_{k,l} \theta_k).$$

Since any positive symmetric $d \times d$ matrix S can be written in the form $C \times (C^t)$ for some matrix C, we conclude that for any such S, the function

$$\exp(-\frac{1}{2} \sum_{k,l} \theta_l S_{k,l} \theta_k)$$

is the characteristic function of some random vector. The law of this random vector is called the *centered Gaussian law with covariance matrix S*.

(6) (Tightness and Prokhorov's theorem). Let us now briefly survey without proof the results regarding density functions and tightness in $\mathbb{R}^d$. It is possible to show that for any tight sequence of probability measures on $\mathbb{R}^d$, one can construct a subsequential limit. This can be done by applying the diagonal trick in some form (for instance, one can apply it for $F_n(x_l)$ where x_l is a dense collection of points in $\mathbb{R}^d$. Some work is needed though to check

that the limiting function can in turn define a distribution function of some probability measure that P_{N_k} would converge to. This is where some really new insight is needed compared to the one-dimensional case, because one can not use the explicit construction of a probability measure with a given distribution function out of the Lebesgue measure on $(0,1)$. The standard procedure is there actually to revisit how probability measures are constructed starting from "outer measures" (i.e., to provide a construction that is reminiscent to the construction of the Lebesgue measure itself – note that we used the existence of the Lebesgue measure in the study of the one-dimensional case).

(7) (Central Limit Theorem). We now finally get to the central limit theorem in $\mathbb{R}^d$.

THEOREM 7.9.2 (Central limit theorem in $\mathbb{R}^d$). *Suppose that $(V_n)_{n\geq 1}$ is a sequence of independent identically distributed random vectors in $\mathbb{R}^d$. If we write $V_1 = (V_{1,1}, \dots, V_{1,d})$, we assume that each $V_{1,j}$ for $j \leq d$ is in L^2 and that $E[V_{1,j}] = 0$, and we denote by $S = (S_{k,l})_{k,l\leq d}$ the $d \times d$ covariance matrix of this vector, so that $S_{k,l} = E[V_{1,k}V_{1,l}]$.*

Then, as $n \to \infty$, the law of $(V_1 + \cdots + V_n)/\sqrt{n}$ converges weakly to the Gaussian centered law with covariance matrix S.

OUTLINE OF THE PROOF. We define P_n the law of $Y_n := S_n/\sqrt{n} = (V_1+\cdots+V_n)/\sqrt{n}$. We can apply our analysis for real-valued random variables to each of the d coordinates of $S_n/\sqrt{n}$, which readily shows that for all $\varepsilon > 0$, if we choose K large enough, then for all n, each of the d coordinates of $S_n/\sqrt{n}$ will be in $[-K, K]$ with probability at least $1 - \varepsilon$. This therefore shows that $S_n/\sqrt{n}$ will be in $[-K, K]^d$ with probability at least $1 - d\varepsilon$. The sequence $(P_n)_{n\geq 1}$ is therefore tight.

Nest, suppose that for some subsequence $n_k \to \infty$, P_{n_k} converges weakly to some probability measure P. Applying the one-dimensional Central Limit Theorem to each given linear combination of the d coordinates of Y_n shows that for each given $\alpha_1, \dots, \alpha_d$, the law of $\alpha_1 Y_n^1 + \ldots + \alpha_d Y_n^d$ converges to a centered Gaussian distribution with covariance given by $\sum_{k,l\leq d} \alpha_k \alpha_l E[V_{1,k}V_{1,l}]$. This in turn show that the limiting probability measure P has to be the Gaussian centered law in $\mathbb{R}^d$ with covariance matrix S.

We can then conclude exactly as in the one-dimensional case, using the fact that a tight sequence of probability measures with just one possible accumulation point (i.e., there is only one P that can be the limit of any subsequential limit), does necessarily converge to this accumulation point. So, the sequence of probability measure P_n does indeed converge weakly to P. □

CHAPTER 8

A glimpse of large deviation theory

8.1. General observations, warm-up

In probability theory, the object of study is often a random system with many random inputs, and one is usually interested in its "typical" behaviour when the number of inputs is very large. This is what we have done so far in these lectures.

Large deviation theory is devoted to the study of "atypical events" for such systems. The main goal is then to answer the following two questions: What is the decay rate (with respect to the size of the system) of the probability of this atypical event (i.e., how fast does it go to 0)? And then, what can one say about the conditional law of the system given this atypical event?

Of course, such issues arises in many important applications, so that the answers are actually very useful!

In this mini-chapter, we will consider the simplest of examples, i.e. the behaviour of the sum of independent identically distributed random variables $(X_n)_{n\geq 1}$ of random variables with law P. We will be interested in the case where $E[|X_1|] < \infty$ (and as we will see, much stronger conditions will actually be needed in our study), and we want to study features of the random variable $S_n := X_1 + \cdots + X_n$. The law of large numbers says that as $n \to \infty$, $S_n/n \to E[X_1]$ almost surely, and the Central Limit Theorem gives further information when $E[(X_1)^2] < \infty$. We now want to understand the asymptotic behaviour when $n \to \infty$ of

$$P[S_n > an],$$

when $a > E[X_1]$ – and then the conditional law of S_n (or of $(X_1, \ldots, X_n)$) when one conditions on $\{S_n > na\}$.

We start with some little warm-up observations:

- A first trivial observation is that if $P[X_1 > a] \in (0,1)$ (note that otherwise $P[S_n > an] = 0$), then
$$P[S_n > an] \geq P[X_1 > a, \ldots, X_n > a] = P[X_1 > a]^n.$$
So this probability can not decay faster than exponentially (w.r. to n).
- When the random variables X_n are Gaussian, then everything is easy and pretty explicit, because S_n is Gaussian as well: For instance, when $(X_n)_{n\geq 1}$ is an i.i.d. sequence of standard Gaussian random variables, then, S_n is a centered Gaussian with variance n. From this, it follows that when $a > 0$ and $n \to \infty$,
$$P[S_n > na] = \frac{1}{\sqrt{2\pi}} \int_{a\sqrt{n}}^{\infty} e^{-y^2/2} dy \sim \frac{1}{a\sqrt{2\pi n}} e^{-na^2/2}.$$
So, we see that the probability that $S_n > an$ decays (at first order) exponentially in n.
- Some further remarks in this Gaussian case: A similar computation shows that
$$P[S_n \in (an, an + n^{\epsilon})] \sim P[S_n > an]$$
as soon as $\epsilon > 0$. In other words, when one conditions on $S_n/n > a$, the conditional probability that $S_n/n \in (a, a + n^{-\beta})$ goes to 1 as soon as $\beta < 1$ [It is actually easy

to see via the same computation that the conditional law of $(S_n - an)$ converges to an exponential random variable as $n \to \infty$. So, the conditional law of S_n/n is indeed pretty much concentrated around a].

- If instead of Gaussian random variables, one considers a sequence of symmetric i.i.d. random variables (X_n) with the property that
$$P[X_1 > x] > x^{-\alpha}$$
as $x \to \infty$ for some given (potentially large) α. Then, it is clear that
$$P[S_n > an] \geq P[X_n > an]P[S_{n-1} > 0] \geq (an)^{-\alpha}/2.$$
So, this probability does not decay exponentially in such cases. Similarly, if $P[X_1 > x] \sim \exp(-\sqrt{n})$ (note that all moments of X are finite in that case), the probability that $S_n > an$ will not decay exponentially either.

8.2. Cramér's Theorem

We now work in a more general case, and will set the stage for Cramér's theorem: We suppose that the law of X is not a Dirac mass (in which case things would be trivial) and that
$$E[\exp(tX_1)] < \infty$$
for all real t. This condition can actually be easily relaxed into the condition that this holds for t in a neighborhood of 0, but it is very important to stress that some pretty strong control on the tail distribution of X is needed when one wants to derive large deviation estimates. Existence of moments is for instance not sufficient (as pointed out at the end of the previous section).

We can note that the function $\varphi(t) := E[\exp(tX)]$ is then C^∞ and strictly convex, and that for all k and t, the k-th derivative of φ at t is $E[X^k \exp(tX)]$. In particular, if a is chosen so that $a > E[X_1]$ and $P[X_1 > a] > 0$, then the function $t \mapsto e^{-at}\varphi(t)$ tends to ∞ when $t \to \infty$ and has a strictly negative derivative at $t = 0$: There therefore exists a unique $t^* = t^*(a)$ such that
$$e^{-at^*}\varphi(t^*) = \min_{t \geq 0} e^{-at}\varphi(t).$$
Note that this quantity $U(a)$ is in $(0, 1)$. If we write down what it means for the derivative of $e^{-at}\varphi(t)$ to be equal to 0 at t^*, we see that
$$E[(X - a)e^{t^*X}] = 0.$$

Let us also define $I(a) = \log(1/U((a))$, so that $U(a) = e^{-I(a)}$.

Theorem 8.2.1 (Cramér's theorem). *Under the above condition,* $\lim_{n\to\infty} P[S_n > an]^{1/n} = U(a)$.

In other words,
$$P[S_n > an] = U(a)^{n+o(n)} = e^{-nI(a)+o(n)}$$
as $n \to \infty$. So, at first order, this probability decays exponentially. However, this statement is of course weaker than saying that $P[S_n > an] = U(a)^{n+O(1)}$ say (indeed, in the Gaussian example in the warm-up section, there is an additional power-law term).

Remark 8.2.2. *The general intuition behind such an exponential decay for the probability is somehow that the optimal strategy in order to achieve $S_n > an$ is that "each X_j" contributed a little bit, and tilts itself a little bit, in order to favor taking larger values. The term $U(a)$ then represents in some sense the "cost" of this tilt for each j, so that $U(a)^n$ is then the "total cost". This will be apparent in the proof of the lower bound.*

Proof of the upper bound. The proof of the upper bound for $P[S_n > an]$ is rather straightforward – just as in many cases in large deviation settings. Here, it is just a consequence of Markov's inequality:

$$P[S_n - an > 0] \le E[e^{t^*(S_n - an)}] = E[e^{t^*(X_1 - a)}]^n = U(a)^n.$$

This is straightforward, and shows that the term $U(a)^{o(n)}$ in Cramér's theorem is actually always smaller than 1. □

Deriving the lower bound is a bit trickier, but also quite instructive:

Proof of the lower bound. In order to get a lower bound on $P[S_n > an]$, one has to construct a somewhat explicit "strategy" in order to construct a subset A of the event $\{S_n > an\}$, and to evaluate the probability of this event A. One general very useful idea here is to use a *change of measure.*

Before explaining a strategy that will provide a good lower bound, let us first explain this idea in the case of the simplest possible strategy, which is to choose $A := \{X_1 > a, \dots, X_n > a\} \subset \{S_n > an\}$. Then, we can view the "conditioning on A" as a two-step operation. First, one conditions each X_j to be greater than a (which "costs" a probability $P[X_1 > a]$ for each of them). Then, we note that for the conditional probability measure, one always has $S_n > a$. Of course, the obtained lower bound $P[X_1 > a]^n$ for $P[S_n > na]$ is not good enough here.

The idea will be instead of brutally making the cut-off for X_1 at a, to re-weight the probability distribution of X_1 to favor the larger values. More specifically, we define a probability measure Q on $\mathbb{R}$ (and we denote by Y the random variable with this distribution) in such a way that for any measurable bounded real-valued function f,

$$E[f(Y)] = \frac{1}{Z} E[f(X) \exp(t^* X)]$$

where $Z := E[\exp(t^* X)]$. In other words, $Q(dx) = P_X(dx) e^{t^* x}/Z$ — i.e., the law of Y is obtained from the law of X by giving a bonus to the large values of X and by penalizing the negative ones. Of course, all this is possible thanks to the condition on the law of X. Note that it also implies that all the moments of Y are finite.

Furthermore, we can note that

$$E[Y] = E[X e^{t^* X}]/E[e^{t^* X}] = a,$$

and that the variance σ_Y^2 of Y in finite and positive (because Y is not a constant random variable). In particular, this means that one can apply the usual central limit to a sequence of i.i.d. random variables with law Q, so that

$$\frac{Y_1 + \ldots + Y_n - na}{\sigma\sqrt{n}}$$

converges in distribution to a standard Gaussian random variable. In particular, we see that

$$\lambda_n := P[Y_1 + \ldots + Y_n \in (an, an + \sqrt{n}]]$$

converges to a positive constant λ.

We are now ready to derive our lower bound for $P[S_n > na]$:

$$\begin{aligned} P[S_n > na] &= E[1_{X_1 + \ldots + X_n > na}] \\ &= Z^n E[1_{Y_1 + \ldots + Y_n > a} e^{-t^*(Y_1 + \ldots + Y_n)}] \\ &\ge Z^n E[1_{Y_1 + \ldots + Y_n \in (na, na + \sqrt{n}]} e^{-t^*(na + \sqrt{n})}] \\ &= E[\exp(t^*(X - a))]^n \times \exp(-t^* \sqrt{n}) \times P[Y_1 + \ldots Y_n \in (na, na + \sqrt{n}]]. \end{aligned}$$

Hence,

$$P[S_n > na]^{1/n} \geq U(a) \times \exp(-t^*/\sqrt{n}) \times \lambda_n^{1/n}$$

and the right-hand side clearly converges to $U(a)$ as $n \to \infty$. □

CHAPTER 9

A glimpse of random walks in $\mathbb{Z}^d$

In the present chapter, we will study the simple random walk in $\mathbb{Z}^d$, which is arguably one of the simplest processes of all, and will derive some of its basic properties. In particular, we shall see that:

(1) Simple random walks in $\mathbb{Z}$ and $\mathbb{Z}^2$ are "recurrent", meaning here that almost surely, they will visit each point on the lattice infinitely often, while simple random walk in $\mathbb{Z}^d$ for $d \geq 3$ are "transient" meaning that they almost surely tend to infinity.

(2) There is a close relation between simple random walks, discrete harmonic functions and some martingales.

Both these type of considerations can be generalized to another class of random processes called Markov chains, but we will only very briefly comment on this.

9.1. Number of return times

Suppose that $d \geq 1$ and consider a sequence of identically distributed random vectors $(\varepsilon_n)_{n\geq 1}$ where the distribution of ε_1 is uniform on the set of $2d$ neighbors of the origin in $\mathbb{Z}^d$. In other words,

$$P(\varepsilon_1 = (1, 0, \dots, 0)) = P(\varepsilon_1 = (-1, 0, \dots, 0)) = P(\varepsilon_1 = (0, 1, 0 \dots, 0)) = \dots = 1/(2d).$$

We then define $Z_0 = O = (0, \dots, 0)$ and $Z_n = \varepsilon_1 + \dots + \varepsilon_n$ for $n \geq 1$. The process $Z = (Z_n)_{n\geq 0}$ is called the simple random walk in $\mathbb{Z}^d$.

We will be interested in the total number N of positive times at which Z is back at the origin. In particular, we want to know whether N is finite or infinite.

For each given $n_0 \geq 1$, we define the process $(Z_n^{[n_0]} := (Z_{n_0+n} - Z_{n_0})_{n\geq 0}$, which is heuristically speaking the random walk rebooted at time n_0. Since $Z_n^{[n_0]} = \varepsilon_{n_0+1} + \ldots + \varepsilon_{n_0+n}$ when $n \geq 1$, it is clear that $(Z_n^{[n_0]})_{n\geq 0}$ has the same law as $(Z_n)_{n\geq 0}$ and that $(Z_n^{[n_0]})_{n\geq 0}$ is independent of $(Z_0, Z_1, \dots, Z_{n_0})$. So, if we denote by $N^{[n_0]}$ the number of returns at O by $Z^{[n_0]}$, then this random variable has the same law as N, and it is independent of $(Z_0, Z_1, \dots, Z_{n_0})$.

Let T be the first time (if it exists) at which Z_n returns to the origin – if this never happens, we write $T = \infty$. We can note that if T is finite, then it is actually even. Furthermore, for any even $n_0 = 2p_0$, $P(T = n_0)$ is positive (the random walk can move p_0 times in one direction and then back).

We note that for any even $n_0 = 2p_0 \geq 2$, the event $\{T = n_0\}$ is measurable with respect to $\sigma(Z_0, \dots, Z_{n_0})$. It is therefore independent of $N^{[n_0]}$. On the other hand, when $T = n_0$, we know that $N = N^{[n_0]} + 1$. We therefore conclude that for all n_0 and all $k \geq 1$,

$$P(N \geq k) = \sum_{p_0=1}^{\infty} P(T = 2p_0 \text{ and } N^{[2p_o]} \geq k-1) = \sum_{p_0=1}^{\infty} P(T = 2p_0) P(N^{[2p_o]} \geq k-1)$$

$$= \sum_{p_0=1}^{\infty} P(T = 2p_0) P(N \geq k-1) = P(N \geq k-1) P(T < \infty).$$

By induction, we therefore see that for all $k \geq 1$,

$$P(T \geq k) = P(T < \infty)^k.$$

So, they are only two options:

- Either $P(T < \infty) = 1$, and then $N = \infty$ almost surely.
- Either $P(T = \infty) > 0$, and then $N < \infty$ almost surely, and the law of N is geometric with
$$P(N = k) = P(T = \infty) \times P(T < \infty)^k$$
for all $k \geq 0$.

In particular, we see that either $E[N] = \infty$, and then $N = \infty$ almost surely, or $E[N] < \infty$ and then $N < \infty$ almost surely. This will be useful because

$$E[N] = \sum_{p_0 \geq 1} P(Z_{2p_0} = O)$$

and it will be possible to check (see the next section) for which dimensions this is finite or infinite.

Let us make one more comment. Consider any $x \in \mathbb{Z}^d \setminus O$. Define now S the first time (if it exists) at which Z visits x, and let N_x denote the total number of visits of x by Z.

- If we are in the configuration where Z almost surely returns to the origin a finite number of times, then: (a) Either $S = \infty$ and therefore $N_x = 0$, or (b) For some $n_0 \geq 1$, $S = n_0$, but then, since $N^{[n_0]} < \infty$ almost surely for all n_0, we get that $N_x < \infty$ almost surely.
- If we are in the configuration where Z almost surely returns to 0 infinitely many times, if we denote by q the probability that $S < \infty$, we can note that either Z visited x before T (which happens with positive probability), and if not, then the conditional probability that it will visit x after T is still q. Hence,
$$q = P(S < T) + (1 - P(S < T))q,$$
which implies that $q = 1$. In other words, $S < \infty$ almost surely. Furthermore, the number of returns at x by Z after S has the same law as N, so that $N_x = \infty$ almost surely.

We can summarize all the above as follows:

Proposition 9.1.1. *For a simple random walk in $\mathbb{Z}^d$:*

- *Either $\sum_{n \geq 1} P(Z_n = O) = \infty$, and then almost surely, the random walk Z visits each site x in $\mathbb{Z}^d$ infinitely often. We then say that the simple random walk in $\mathbb{Z}^d$ is recurrent.*
- *Or $\sum_{n \geq 1} P(Z_n = O) < \infty$, and then almost surely, each $x \in \mathbb{Z}^d$ is visited only finitely many times, so that $\|Z_n\| \to \infty$ as $n \to \infty$. We then say that the simple random walk in $\mathbb{Z}^d$ is transient.*

Remark 9.1.2. *We can note that it is possible to formalize the previous ideas quite nicely using the formalism of stopping times. One can consider the filtration $(\mathcal{F}_n := \sigma(Z_0, \dots, Z_n))_{n \geq 0}$, and note that T and S are in fact stopping times for this filtration. In fact, for any finite stopping time τ, it is easy to see that $Z^{[\tau]} = (Z_{\tau+n} - Z_\tau)_{n \geq 0}$ is a simple random walk that is independent of $\mathcal{F}_\tau$. This is called the* strong Markov property of simple random walk.

9.2. Recurrence in $\mathbb{Z}^2$, transience in $\mathbb{Z}^3$

We are now going to prove the following result:

Proposition 9.2.1. *Simple random walk in $\mathbb{Z}^d$ is recurrent when $d = 1$ and $d = 2$, and it is transient when $d \geq 3$.*

Proof. We can first start with the following remark about the simple random walk $(S_n)_{n\geq 0}$ on $\mathbb{Z}$. By the binomial formula and Stirling's formula, we have that

$$P(S_{2n} = 0) = 2^{-2n}\frac{(2n)!}{n!n!} \sim \frac{1}{\sqrt{2\pi n}}$$

as $n \to \infty$. Since $\sum_{n\geq 1} n^{-1/2} = \infty$, we get that simple random walk in $\mathbb{Z}$ is recurrent (note that we could also have deduced this from the fact that in this case, $\limsup_{n\to\infty} S_n = +\infty$ and $\liminf_{n\to\infty} S_n = -\infty$ almost surely). Note that this estimate implies also the existence of a constant C such that this probability is bounded by $C/\sqrt{n}$ for all $n \geq 1$.

- To show the recurrence of simple random walk in $\mathbb{Z}^2$, we note that if we consider two independent one-dimensional random walks S^1 and S^2, and looks at
$$Z_n := \left(\frac{S_n^1 + S_n^2}{2}, \frac{S_n^1 - S_n^2}{2}\right),$$
one gets a simple random walk on $\mathbb{Z}^2$. Furthermore, $Z_n = O$ if and only if $S_n^1 = S_n^2 = 0$. It follows that if $(Z_n)_{n\geq 0}$ is a simple random walk in $\mathbb{Z}^2$, then
$$P(Z_{2n} = O) = P(S_{2n}^1 = 0 \text{ and } S_{2n}^2 = 0) = P(S_{2n}^1 = 0) \times P(S_{2n}^2) = 0) \sim \frac{1}{2\pi n}$$
as $n \to \infty$, and therefore
$$\sum_{n\geq 1} P(S_n = O) = \infty.$$
The simple random walk in $\mathbb{Z}^2$ is therefore recurrent.
- When $d \geq 3$, we can introduce the following additional random variables. For each n, we denote by U_n^1, U_n^2, U_n^3 the respective number of jumps be the first, second and third coordinate of the random walk Z up to time n. So, $U_n^1 + U_n^2 + U_n^3 = n$, and each U_n^j has the law of the sum of n independent random Bernoulli random variables with parameter 1/3.

  On the one hand, it is easy to see that, conditionally on $\{U_n^1 = u_1, U_n^2 = u_2, U_n^3 = u_3\}$ (note that for $u_1+u_2+u_3 = n$, this event has a positive probability), the three coordinates of Z_n are independent, and have respectively the laws of S_{u_1}, S_{u_2} and S_{u_3}. In particular, we see that if $\min(u_1, u_2, u_3) \geq n/4$, then
$$P(Z_n = O|U_n^1 = u_1, U_n^2 = u_2, U_n^3 = u_3) \leq (C/\sqrt{n/4}))^3 \leq 8C^3 n^{-3/2},$$
so that
$$P(Z_n = O|\min(U_n^1, U_n^2, U_n^3) \geq n/4) \leq 8C^3 n^{-3/2}.$$
On the other hand, elementary large deviations (this is the easy upper bound in Cramér's theorem) tell us that $P(U_n^1 < n/4)$ decays exponentially in n, i.e., that there exist constants C_1 and C_2, such that this probability is bounded by $C_1 \exp(-C_2 n)$.

  Putting the pieces together, we get that for all $n \geq 1$,
$$\begin{aligned} P(Z_n = O) &\leq P(\min(U_n^1, U_n^2, U_n^3) < n/4) + P(Z_n = O|\min(U_n^1, U_n^2, U_n^3) \geq n/4) \\ &\leq 3C_1 \exp(-C_2 n) + 8C^3 n^{-3/2}. \end{aligned}$$

This last expression being summable in n, we conclude that the simple random walk in 3 dimensions is transient.

- When $d > 3$, we could use the same argument as for $d = 3$ (just replacing the $n/4$ cut-off by $n/(d+1)$ for instance) to see that $P(Z_n = O) \leq C_d/n^{d/2}$ for some constant C_d, so that the simple random walk is again transient. Another simpler argument is to note that the projection of a d-dimensional random walk on the first three coordinates is just a three-dimensional random walk with the property that it can stay still at each step with probability $(d-3)/d$, and to use the transience of the 3-dimensional random walk.

□

9.3. Discrete harmonic functions and martingales

Suppose that D is a subset of $\mathbb{Z}^d$, and denote by ∂D, the set of points in $\mathbb{Z}^d \setminus D$ that are at distance 1 of D, and $\overline{D} = D \cup \partial D$.

A function H from $\overline{D}$ into $\mathbb{R}$ is said to be (discrete) harmonic in D if for all $x \in D$, $H(x)$ is equal to the mean values of H on the $2d$ neighbors of x.

We are very quickly going to point out a basic relation between harmonic functions and random walks.

Suppose that H is a function defined in $\overline{D}$ that is harmonic in D. Suppose furthermore that H is bounded (this is an important assumption here – note that this is always the case if D is finite).

Let us now fix $x_0 \in D$, and denote by $(Z_n)_{n\geq 0}$ a simple random walk in $\mathbb{Z}^d$ that is started from $Z_0 = x_0$. To recall this dependence on the starting point x_0, it is customary to write P_{x_0} for this probability measure (and E_{x_0} for the corresponding expectations).

We let T be the first time (if it exists – otherwise, we let $T = \infty$) when Z exits D (i.e., when it is in ∂D). We let $(\mathcal{F}_n)_{n\geq 0}$ denote the filtration generated by the random walk.

Lemma 9.3.1. *The following statements hold:*

- *The process $(H(Z_{\min(n,T)}))_{n\geq 0}$ is then a martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$.*
- *It converges almost surely to $H(Z_T)$ as $n \to \infty$.*
- *One has $H(x_0) = E_{x_0}[H(Z_T)]$.*

Proof. When $T > n$, then the conditional expectation of the increment $H(Z_{n+1}) - H(Z_n)$ given $\mathcal{F}_n$ is the difference between the mean value of H on the neighbors of Z_n and $H(Z_n)$, which is equal to 0 since H is harmonic. When $T \leq n$, then clearly $H(Z_{\min(n+1,T)}) = H(Z_{\min(n,T)})$. We therefore conclude that $(M_n)_{n\geq 0}$ is indeed a martingale.

It converges almost surely to M_T (for instance because the limsup of the first coordinate of Z_n is $+\infty$ as $n \to \infty$ so that $T < \infty$ almost surely) – and M_n converges also in L^1 to M_T because H (and therefore M_n) is bounded. □

Definition 9.3.2. *Suppose now that D is a subset of $\mathbb{Z}^d$ as above and that f is some given function from ∂D into $\mathbb{R}$. We say that H is a solution to the discrete Dirichlet problem in D with boundary values given by f, if it is a function from $\overline{D}$ into $\mathbb{R}$, such that (a) H is discrete harmonic in D, and (b) $H = f$ on ∂D.*

We can now easily deduce the following result from the previous lemma:

Proposition 9.3.3. *When D is bounded, then the solution to the discrete Dirichlet problem (when D is bounded) exists, it is unique, and it is given by the function $H(x) = E_x[H(Z_T)]$.*

Remark 9.3.4. *There actually exists numerous ways to prove this. We can note that the Dirichlet problem is a simple linear system with $\#D$ equations (one per site of D) and $\#D$ unknown (the values of H at each site of D) – so that one just has to check that the corresponding matrix is invertible, which can be done in a number of ways.*

Proof. Because of the lemma, it suffices to check that the function $E_x[H(Z_T)]$ is indeed a solution to the Dirichlet problem. It is clearly equal to f on ∂D, so that it remains to check that it is harmonic in D. But this follows immediately by applying the Markov property at time 1 to the simple random walk started from x. □

Remark 9.3.5. *So, with this particular example, we see how simple random walks and harmonic functions are very naturally related to martingales. This observation can be extended in the continuum setting and will related Brownian motion to harmonic functions in the continuum (this is one of the main topics of Brownian motion and stochastic calculus lectures).*

REMARK 9.3.6 (Extension to Markov Chains). *Simple random walks are just one very special example of a more general class of processes in discrete time, for which one can make a similar relation to martingales.*

Suppose that one is given a state space $\mathcal{S}$ that is at most countable. Suppose that we are given a function Q from $\mathcal{S} \times \mathcal{S}$ into $[0,1]$ that has the property that for all $x \in \mathcal{S}$,

$$\sum_{y \in \mathcal{S}} Q(x,y) = 1.$$

DEFINITION 9.3.7. *We say that the collection of random variables $(X_n)_{n \geq 0}$ is a Markov chain with kernel Q started from x_0 if $X_0 = x_0$ almost surely and if for all $n \geq 1$ and $x_1, \ldots, x_n$ in $\mathcal{S}$,*

$$P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = Q(x_0, x_1) \times \cdots \times Q(x_{n-1}, x_n).$$

The simple intuition is the following: When the process is at x at time n, then it chooses to jump to y at time $n+1$ with a (conditional probability) $Q(x,y)$. The simple random walk is just the case where $Q(x,y)$ is equal to $1/(2d)$ if y is a neighbor of x and to 0 otherwise.

We can note that this definition describes fully the law *of a Markov chain with kernel Q started from x_0. It is not difficult to show that such a Markov chain actually exists (one can build $(X_n)_{n \geq 0}$ inductively out of a sequence of independent uniform random variables in $[0,1]$).*

The previous notion of harmonic functions in D will be replaced by the class of functions H (defined on $\overline{D}$) that are in the kernel of $Q - I$, i.e., such that for all $x \in D \subset \mathcal{S}$,

$$\sum_{y \in \mathcal{S}} H(y) Q(x,y) = H(x).$$

It is the easy to check that for a Markov chain with kernel Q started from x, then for such a function H, the process $(H(X_{\min(n,T)})_{n \geq 0}$ will be a martingale (when T is the first time at which X_n is in ∂D). One can then (this will depend on further properties of the Markov chain) adapt the results that we derived for simple random walks to this more general case. Markov chains are one main object of interest in the Applied Stochastic Processes *course.*