Appendix B

Probability

This appendix summarizes the probabilistic notions that are most important in the book. Although many readers will not need to be reminded of the basic definitions, they might still refer to it to check some easy probabilistic statements whose proof we have included here to avoid disrupting the arguments in the main part of the book. For convergence in law, we will refer mostly to the book of Billingsley [10] and, for random series and similar topics, to that of Li and Queffélec [83].

B.1 The Riesz Representation Theorem

Let X be a *locally compact* topological space (such as $\mathbf{R}^d$ for $d \geqslant 1$). We recall that *Radon measures* on X are certain measures for which compact subsets of X have finite measure, and which satisfy some regularity property (the latter requirement being unnecessary if any open set in X is a countable union of compact sets, as is the case of **R** for instance).

The Riesz representation theorem interprets Radon measures in terms of the corresponding integration functional. It can be taken as a definition (and it is indeed the definition in Bourbaki's theory of integration [17]); for a proof in the usual context where measures are defined "set-theoretically," see, for example, [106, Th 2.14].

Theorem B.1.1 *Let* X *be a locally compact topological space and* $\mathrm{C}_c(\mathrm{X})$ *the space of compactly supported continuous functions on* X*. For any linear form* $\lambda\colon \mathrm{C}_c(\mathrm{X}) \to \mathbf{C}$ *such that* $\lambda(f) \geqslant 0$ *if* $f \geqslant 0$*, there exists a unique Radon measure* μ *on* X *such that*

$$\lambda(f) = \int_{\mathrm{X}} f \, d\mu$$

for all $f \in \mathrm{C}_c(\mathrm{X})$.

Let $k \geqslant 1$ be an integer. If $X = \mathbf{R}^k$ (or an open set in $\mathbf{R}^k$), then Radon measures can be identified by the integration of much more regular functions. For instance, we have the following (see, e.g., [28, Th. 3.18]):

Proposition B.1.2 *Let* $C_c^\infty(\mathbf{R}^k)$ *be the space of smooth compactly-supported functions on* $\mathbf{R}$*. For any linear form* $\lambda\colon C_c^\infty(\mathbf{R}^k) \to \mathbf{C}$ *such that* $\lambda(f) \geqslant 0$ *if* $f \geqslant 0$*, there exists a unique Radon measure* μ *on* $\mathbf{R}^k$ *such that*

$$\lambda(f) = \int_{\mathbf{R}^k} f\, d\mu$$

for all $f \in C_c^\infty(\mathbf{R}^k)$.

Remark B.1.3 When applying either form of the Riesz representation theorem, we may wish to identify whether the measure μ obtained from the linear form λ is a probability measure on X or not. This is the case if and only

$$\sup_{\substack{f \in C_c(X) \\ 0 \leqslant f \leqslant 1}} \lambda(f) = 1,$$

(see, e.g., [17, Ch. 4, § 1, n^{o}8]) where, in the setting of Proposition B.1.2, we may also restrict f to be smooth.

Moreover, if a positive linear form $\lambda\colon C_c(X) \to \mathbf{C}$ admits an extension to a linear form $\lambda\colon C_b(X) \to \mathbf{C}$, where $C_b(X)$ is the space of continuous and bounded functions on X, which is still positive (so $\lambda(f) \geqslant 0$ if $f \in C_b(X)$ is non-negative), then the measure μ associated to λ is a probability measure if and only if $\lambda(1) = 1$, where 1 on the left is the constant function. (This is natural enough, but it is not entirely obvious; the underlying reason is that the positivity implies that

$$|\lambda(f)| \leqslant \|f\|_\infty \lambda(1)$$

where $\|f\|_\infty$ is the supremum norm for a bounded continuous function, so that λ is a continuous linear form on the Banach space $C_b(X)$.)

B.2 Support of a Measure

Let M be a topological space. If M is either second countable (i.e., there is basis of open sets that is countable) or compact, then any Borel measure μ on M has a well-defined closed *support*, denoted $\mathrm{supp}(\mu)$, which is characterized by either of the following properties: (1) it is the complement of the largest open set U, with respect to inclusion, such that $\mu(U) = 0$; or (2) it is the set of those $x \in M$ such that, for any open neighborhood U of x, we have $\mu(U) > 0$.

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

If X is a random variable with values in M, we will say that the *support of* X is the support of the law of X, which is a probability measure on M.

We need the following elementary property of the support of a measure:

Lemma B.2.1 *Let* M *and* N *be topological spaces that are each either second countable or compact. Let* μ *be a probability measure on* M*, and let* $f : \mathrm{M} \longrightarrow \mathrm{N}$ *be a continuous map. The support of* $f_*(\mu)$ *is the closure of* $f(\mathrm{supp}(\mu))$.

We recall that given a probability measure μ on M and a continuous map $f : \mathrm{M} \to \mathrm{N}$, the *image measure* $f_*(\mu)$ is defined by

$$f_*(\mu)(\mathrm{A}) = \mu(f^{-1}(\mathrm{A}))$$

for a measurable set $\mathrm{A} \subset \mathrm{N}$, and it satisfies

$$\int_{\mathrm{N}} \varphi(x) d(f_*\mu)(x) = \int_{\mathrm{M}} \varphi(f(y)) d\mu(y)$$

for $\varphi \geqslant 0$ and measurable, or $\varphi \circ f$ integrable with respect to μ.

Proof First, if $y = f(x)$ for some $x \in \mathrm{supp}(\mu)$, and if U is an open neighborhood of y, then we can find an open neighborhood $\mathrm{V} \subset \mathrm{M}$ of x such that $f(\mathrm{V}) \subset \mathrm{U}$. Then $(f_*\mu)(\mathrm{U}) \geqslant \mu(\mathrm{V}) > 0$. This shows that y belongs to the support of $f_*\mu$. Since the support is closed, we deduce that $\overline{f(\mathrm{supp}(\mu))} \subset \mathrm{supp}(f_*\mu)$.

For the converse, let $y \in \mathrm{N}$ be in the support of $f_*\mu$. For any open neighborhood U of y, we have $\mu(f^{-1}(\mathrm{U})) = (f_*\mu)(\mathrm{U}) > 0$. This implies that $f^{-1}(\mathrm{U}) \cap \mathrm{supp}(\mu)$ is not empty, and since U is arbitrary, that y belongs to the closure of $f(\mathrm{supp}(\mu))$. □

Recall that a family $(\mathrm{X}_i)_{i \in \mathrm{I}}$ of random variables, each taking possibly values in a different metric space M_i, is *independent* if, for any finite subset $\mathrm{J} \subset \mathrm{I}$, the joint distribution of $(\mathrm{X}_j)_{j \in \mathrm{J}}$ is the measure on $\prod \mathrm{M}_j$ which is the product measure of the laws of the X_j.

Lemma B.2.2 *Let* $\mathrm{X} = (\mathrm{X}_i)_{i \in \mathrm{I}}$ *be a finite family of random variables with values in a topological space* M *that is compact or second countable. Viewed as a random variable taking values in* M^{I}*, we have*

$$\mathrm{supp}(\mathrm{X}) = \prod_{i \in \mathrm{I}} \mathrm{supp}(\mathrm{X}_i).$$

Proof If $x = (x_i) \in \mathrm{M}^{\mathrm{I}}$, then an open neighborhood U of x contains a product set $\prod \mathrm{U}_i$, where U_i is an open neighborhood of x_i in M. Then we have

$$\mathbf{P}(\mathrm{X} \in \mathrm{U}) \geqslant \mathbf{P}(\mathrm{X} \in \prod_i \mathrm{U}_i) = \prod_i \mathbf{P}(\mathrm{X}_i \in \mathrm{U}_i)$$

by independence. If $x_i \in \text{supp}(X_i)$ for each i, then this is > 0, and hence $x \in \text{supp}(X)$.

Conversely, if $x \in \text{supp}(X)$, then for any $j \in I$, and any open neighborhood U of x_j, the set

$$V = \{y = (y_i)_{i \in I} \in M^I \mid y_j \in U\} \subset M^I$$

is an open neighborhood of x. Hence we have $\mathbf{P}(X \in V) > 0$, and since $\mathbf{P}(X \in V) = \mathbf{P}(X_i \in U)$, it follows that x_j is in the support of X_j. □

B.3 Convergence in Law

Let M be a metric space. We view it as given with the Borel σ-algebra generated by open sets, and we denote by $C_b(M)$ the Banach space of bounded complex-valued continuous functions on M, with the norm

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

Given a sequence (μ_n) of probability measures on M, and a probability measure μ on M, one says that μ_n converges weakly to μ if and only if, for any bounded and continuous function $f : M \longrightarrow \mathbf{R}$, we have

$$\int_M f(x) d\mu_n(x) \longrightarrow \int_M f(x) d\mu(x). \tag{B.1}$$

If $(\Omega, \Sigma, \mathbf{P})$ is a probability space and $(X_n)_{n \geqslant 1}$ is a sequence of M-valued random variables, and if X is an M-valued random variable, then one says that (X_n) converges in law to X if and only if the measures $X_n(\mathbf{P})$ converge weakly to $X(\mathbf{P})$. If μ is a probability measure on M, then we will also say that X_n converges to μ if the measures $X_n(\mathbf{P})$ converge weakly to μ.

The probabilistic versions of (B.1) in those cases is that

$$\mathbf{E}(f(X_n)) \longrightarrow \mathbf{E}(f(X)) = \int_M f d\mu \tag{B.2}$$

for all functions $f \in C_b(M)$.

Remark B.3.1 If $M = \mathbf{R}$, convergence in law is often introduced in terms of the distribution function $F_X(x) = \mathbf{P}(X \leqslant x)$ of a real-valued random variable X. Precisely, it is classical (see, e.g., [9, Th. 25.8]) that a a sequence of real-valued random variables (X_N) converges in law to a random variable X

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

if and only if $F_{X_m}(x) \to F_X(x)$ for all $x \in \mathbf{R}$ such that F_X is continuous at x (which is true for all but at most countably many x, namely, all x such that $\mathbf{P}(X = x) = 0$).

The definition immediately implies the following very useful fact, which we state in probabilistic language (we will refer to it as the composition principle).

Proposition B.3.2 *Let* M *be a metric space. Let* (X_n) *be a sequence of* M-*valued random variables such that* X_n *converges in law to a random variable* X. *For any metric space* N *and any continuous function* $\varphi\colon M \to N$*, the* N-*valued random variables* $\varphi \circ X_n$ *converge in law to* $\varphi \circ X$.

Proof For any continuous and bounded function $f : N \longrightarrow \mathbf{C}$, the composite $f \circ \varphi$ is bounded and continuous on M, and therefore convergence in law implies that

$$\mathbf{E}(f(\varphi(X_n))) \longrightarrow \mathbf{E}(f(\varphi(X))).$$

By definition, this formula, valid for all f, means that $\varphi(X_n)$ converges in law to $\varphi(X)$. □

Checking the condition (B.2) for all $f \in C_b(M)$ may be difficult. A number of convenient criteria, and properties, of convergence in law are related to weakening this requirement to only *certain* "test functions" f, which may be more regular, or have special properties. We will discuss some of these in the next sections.

One often important consequence of convergence in law is a simple relation with the support of the limit of a sequence of random variables.

Lemma B.3.3 *Let* M *be a second countable or compact topological space. Let* (X_n) *be a sequence of* M-*valued random variables, defined on some probability spaces* Ω_n. *Assume that* (X_n) *converges in law to some random variable* X, *and let* $N \subset M$ *be the support of the law of* X.

(1) *For any* $x \in N$ *and for any open neighborhood* U *of* x, *we have*

$$\liminf_{n\to+\infty} \mathbf{P}(X_n \in U) > 0,$$

and in particular there exists some $n \geqslant 1$ *and some* $\omega \in \Omega_n$ *such that* $X_n(\omega) \in U$.

(2) *For any* $x \in M$ *not belonging to* N, *there exists an open neighborhood* U *of* x *such that*

$$\limsup_{n\to+\infty} \mathbf{P}(X_n \in U) = 0.$$

Proof For (1), a standard equivalent form of convergence in law is that, for any open set $U \subset M$, we

$$\liminf_{n\to+\infty} \mathbf{P}(X_n \in U) \geqslant \mathbf{P}(X \in U)$$

(see [10, Th. 2.1, (i) and (iv)]). If $x \in N$ and U is an open neighborhood of x, then by definition we have $\mathbf{P}(X \in U) > 0$, and therefore

$$\liminf_{n\to+\infty} \mathbf{P}(X_n \in U) > 0.$$

For (2), if $x \in M$ is not in N, there exists an open neighborhood V of x such that $\mathbf{P}(X \in V) = 0$. For some $\delta > 0$, this neighborhood contains the closed ball C of radius δ around f, and by [10, Th. 2.1., (i) and (iii)], we have

$$0 \leqslant \limsup_{p\to+\infty} \mathbf{P}(X_n \in C) \leqslant \mathbf{P}(X \in C) = 0,$$

hence the second assertion with U the open ball of radius δ. □

Another useful relation between support and convergence in law is the following:

Corollary B.3.4 *Let* M *be a second countable or compact topological space. Let* (X_n) *be a sequence of* M*-valued random variables, defined on some probability spaces* Ω_n *such that* (X_n) *converges in law to a random variable* X*. Let* g *be a continuous function on* M *such that* $g(X_n)$ *converges in probability to zero; that is, we have*

$$\lim_{n\to+\infty} \mathbf{P}_n(|g(X_n)| > \delta) = 0$$

for all $\delta > 0$*. The support of* X *is then contained in the zero set of* g*.*

Proof Let N be the support of X. Suppose that there exists $x \in N$ such that $|g(x)| = \delta > 0$. Since the set of all $y \in M$ such that $|g(y)| > \delta$ is an open neighborhood of x, we have

$$\liminf_{n\to+\infty} \mathbf{P}(|g(X_n)| > \delta) > 0$$

by the previous lemma; this contradicts the assumption, which implies that

$$\lim_{n\to+\infty} \mathbf{P}(|g(X_n)| > \delta) = 0.$$

□

Remark B.3.5 Another proof is that it is well known (and elementary) that convergence in probability implies convergence in law, so in the situation of the corollary, the sequence $(g(X_n))$ converges to 0 in law. Since it also converges

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

to $g(\mathrm{X})$ by composition, we have $\mathbf{P}(g(\mathrm{X}) \neq 0) = 0$, which precisely means that the support of X is contained in the zero set of g.

We also recall an important definition that is a property of weak-compactness for a family of probability measures (or random variables).

Definition B.3.6 (Tightness) Let M be a complete separable metric space. Let $(\mu_i)_{i\in \mathrm{I}}$ be a family of probability measures on M. One says that (μ_i) is *tight* if for any $\varepsilon > 0$, there exists a compact subset $\mathrm{K} \subset \mathrm{M}$ such that $\mu_i(\mathrm{K}) \geqslant 1 - \varepsilon$ for all $i \in \mathrm{I}$.

It is a non-obvious fact that a single probability measure on a complete separable metric space is tight (see [10, Th. 1.3]).

B.4 Perturbation and Convergence in Law

As we suggested in Section 1.2, we will often prove convergence in law of the sequences of random variables that interest us by showing that they are obtained by "perturbation" of other sequences that are more accessible. In this section, we explain how to handle some of these perturbations.

A very useful tool for this purpose is the following property, which is a first example of reducing the proof of convergence in law to more regular test functions than all bounded continuous functions.

Let M be a metric space, with distance d. Recall that a continuous function $f : \mathrm{M} \to \mathbf{C}$ is said to be a *Lipschitz function* if there exists a real number $\mathrm{C} \geqslant 0$ such that

$$|f(x) - f(y)| \leqslant \mathrm{C}d(x, y)$$

for all $(x, y) \in \mathrm{M} \times \mathrm{M}$. We then say that C is a *Lipschitz constant* for f (it is, of course, not unique).

Proposition B.4.1 *Let* M *be a complete separable metric space. Let* (X_n) *be a sequence of* M*-valued random variables, and* μ *a probability measure on* M. *Then* X_n *converges in law to* μ *if and only if we have*

$$\mathbf{E}(f(\mathrm{X}_n)) \to \int_{\mathrm{M}} f(x) d\mu(x)$$

for all bounded Lipschitz functions $f : \mathrm{M} \longrightarrow \mathbf{C}$.

In other words, it is enough to prove the convergence property (B.2) for Lipschitz test functions.

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

Proof A classical argument shows that convergence in law of (X_n) to μ is equivalent to

$$\mu(F) \geqslant \limsup_{n\to+\infty} \mathbf{P}(X_n \in F) \tag{B.3}$$

for all closed subsets F of M (see, e.g., [10, Th. 2.1, (iii)]).

However, the proof that convergence in law *implies* this property uses only Lipschitz test functions f (see, e.g., [10, (ii)⇒(iii), p. 16, and (1.1), p. 8], where it is only stated that the relevant functions f are uniformly continuous, but this is shown by checking that they are Lipschitz). Hence the assumption that (B.2) holds for Lipschitz functions implies (B.3) for all closed subsets F, and consequently it implies convergence in law. □

We can now deduce various corollaries concerning perturbation of sequences that converge in law.

The first result along these lines is quite standard, and the second is a bit more ad hoc but will be convenient in Chapter 6.

Corollary B.4.2 *Let* M *be a separable Banach space. Let* (X_n) *and* (Y_n) *be sequences of* M*-valued random variables. Assume that the sequence* (X_n) *converges in law to a random variable* X.

If the sequence (Y_n) *converges in probability to* 0*, or if* (Y_n) *converges to* 0 *in* L^p *for some fixed* $p \geqslant 1$*, with the possibility that* $p = +\infty$*, then the sequence* $(X_n + Y_n)_n$ *converges in law to* X *in* M.

Proof Let $f : M \longrightarrow \mathbf{C}$ be a bounded Lipschitz continuous function, and C a Lipschitz constant of f. For any n, we have

$$\begin{aligned} |\mathbf{E}(f(X_n + Y_n)) - \mathbf{E}(f(X))| \leqslant |\mathbf{E}(f(X_n + Y_n)) - \mathbf{E}(f(X_n))| \\ + |\mathbf{E}(f(X_n)) - \mathbf{E}(f(X))|. \end{aligned} \tag{B.4}$$

First assume that (Y_n) converges to 0 in L^p, and that $p < +\infty$. Then we obtain

$$\begin{aligned} |\mathbf{E}(f(X_n + Y_n)) - \mathbf{E}(f(X))| &\leqslant C\,\mathbf{E}(|Y_n|) + |\mathbf{E}(f(X_n)) - \mathbf{E}(f(X))| \\ &\leqslant C\,\mathbf{E}(|Y_n|^p)^{1/p} + |\mathbf{E}(f(X_n)) - \mathbf{E}(f(X))| \end{aligned}$$

which converges to 0, hence $(X_n + Y_n)$ converges in law to X. If $p = +\infty$, a similar argument, left to the reader, applies.

Suppose now that (Y_n) converges to 0 in probability. Let $\varepsilon > 0$. For n large enough, the second term in (B.4) is $\leqslant \varepsilon$ since X_n converges in law to X. For the first, we fix another parameter $\delta > 0$ and write

$$|\mathbf{E}(f(X_n + Y_n)) - \mathbf{E}(f(X_n))| \leqslant C\delta + 2\|f\|_\infty \mathbf{P}(|Y_n| > \delta)$$

by separating the integral depending on whether $|Y_n| \leqslant \delta$ or not. Take $\delta = C^{-1}\varepsilon$, so the first term here is $\leqslant \varepsilon$. Then since (Y_n) converges in probability to 0, we have

$$2\|f\|_\infty \mathbf{P}(|Y_n| > \delta) \leqslant \varepsilon$$

for all n large enough, and therefore

$$|\mathbf{E}(f(X_n + Y_n)) - \mathbf{E}(f(X))| \leqslant 3\varepsilon$$

for all n large enough. The result now follows from Proposition B.4.1. □

Here is the second variant, where we do not attempt to optimize the assumptions.

Corollary B.4.3 *Let $m \geqslant 1$ be an integer. Let (X_n) and (Y_n) be sequences of $\mathbf{R}^m$-valued random variables, let (α_n) be a sequence in $\mathbf{R}^m$ and (β_n) a sequence of real numbers. Assume*

(1) *the sequence (X_n) converges in law to a random variable X, and $\|X_n\|$ is bounded by a constant $N \geqslant 0$, independent of n;*

(2) *for all n, we have $\|Y_n\| \leqslant \beta_n$;*

(3) *we have $\alpha_n \to (1, \ldots, 1)$ and $\beta_n \to 0$ as $n \to +\infty$.*

Then the sequence $(\alpha_n \cdot X_n + Y_n)_n$ converges in law to X in $\mathbf{R}^m$, where here $\cdot$ denotes the componentwise product of vectors.[1]

Proof We begin as in the previous corollary. Let $f : \mathbf{R}^m \longrightarrow \mathbf{C}$ be a bounded Lipschitz continuous function, and C its Lipschitz constant. For any n, we now have

$$|\mathbf{E}(f(\alpha_n \cdot X_n + Y_n)) - \mathbf{E}(f(X))| \leqslant |\mathbf{E}(f(\alpha_n \cdot X_n + Y_n)) - \mathbf{E}(f(\alpha_n \cdot X_n))| + |\mathbf{E}(f(\alpha_n \cdot X_n)) - \mathbf{E}(f(X_n))| + |\mathbf{E}(f(X_n)) - \mathbf{E}(f(X))|.$$

The last term tends to 0 since X_n converges in law to X. The second is at most

$$C|\alpha_n - (1, \ldots, 1)\| \mathbf{E}(\|X_n\|) \leqslant CN\|\alpha_n - (1, \ldots, 1)\| \to 0,$$

and the first is at most

$$C\,\mathbf{E}(\|Y_n\|) \leqslant C\beta_n \to 0.$$

The result now follows from Proposition B.4.1. □

[1] I.e., we have $(a_1, \ldots, a_m) \cdot (b_1, \ldots, b_m) = (a_1 b_1, \ldots, a_m b_m)$.

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

The last instance of perturbations is slightly different. It amounts, in practice, to using some "auxiliary parameter m" to approximate a sequence of random variables; when the error in such an approximation is suitably small, and the approximations converge in law for each fixed m, we obtain convergence in law.

Proposition B.4.4 *Let* M *be a finite-dimensional Banach space. Let* $(X_n)_{n\geqslant 1}$ *and* $(X_{n,m})_{n\geqslant m\geqslant 1}$ *be* M*-valued random variables. Define* $E_{n,m} = X_n - X_{n,m}$. *Assume that*

(1) *for each* $m \geqslant 1$*, the random variables* $(X_{n,m})_{n\geqslant m}$ *converge in law to a random variable* Y_m*;*

(2) *we have*

$$\lim_{m\to+\infty} \limsup_{n\to+\infty} \mathbf{E}(\|E_{n,m}\|) = 0.$$

Then the sequences (X_n) *and* (Y_m) *converge in law as* $n \to +\infty$ *and have the same limit distribution.*

The second assumption means in practice that

$$\mathbf{E}(\|E_{n,m}\|) \leqslant f(n,m)$$

where $f(n,m) \to 0$ as m tends to $+\infty$, uniformly for $n \geqslant m$.

A statement of this kind can be found also, for instance, in [10, Th. 3.2], but the latter assumes that it is already known that (Y_m) converges in law.

Proof We begin by proving that (X_n) converges in law. Let $f : M \longrightarrow \mathbf{R}$ be a bounded Lipschitz continuous function, and C a Lipschitz constant for f. For any $n \geqslant 1$ and any $m \leqslant n$, we have

$$|\,\mathbf{E}(f(X_n)) - \mathbf{E}(f(X_{n,m}))| \leqslant C\,\mathbf{E}(\|E_{n,m}\|),$$

hence

$$\mathbf{E}(f(X_{n,m})) - C\,\mathbf{E}(\|E_{n,m}\|) \leqslant \mathbf{E}(f(X_n)) \leqslant \mathbf{E}(f(X_{n,m})) + C\,\mathbf{E}(\|E_{n,m}\|).$$

Fix first $m \geqslant 1$. By the first assumption, the expectations $\mathbf{E}(f(X_{n,m}))$ converge to $\mathbf{E}(f(Y_m))$ as $n \to +\infty$. Then these inequalities imply that we have

$$\limsup_{n\to+\infty} \mathbf{E}(f(X_n)) - \liminf_{n\to+\infty} \mathbf{E}(f(X_n)) \leqslant 2C \limsup_{n\to+\infty} \mathbf{E}(\|E_{n,m}\|)$$

(because any limit of a convergent subsequence of $\mathbf{E}(f(X_n))$ will lie in an interval of length at most the right-hand side). Letting m go to infinity, the second assumption allows us to conclude that

$$\limsup_{n\to+\infty} \mathbf{E}(f(\mathrm{X}_n)) - \liminf_{n\to+\infty} \mathbf{E}(f(\mathrm{X}_n)) = 0,$$

so that the sequence $(\mathbf{E}(f(\mathrm{X}_n)))_{n\geqslant 1}$ converges.

Now consider the map μ defined on bounded Lipschitz functions on M by

$$\mu(f) = \lim_{n\to+\infty} \mathbf{E}(f(\mathrm{X}_n)).$$

It is elementary that μ is linear and that it is positive (in the sense that if $f \geqslant 0$, we have $\mu(f) \geqslant 0$) and satisfies $\mu(1) = 1$. By the Riesz representation theorem (see Proposition B.1.2 and Remark B.1.3, noting that a finite-dimensional Banach space is locally compact), it follows that μ "is" a probability measure on M. It is then tautological that (X_n) converges in law to a random vector X with probability law μ by Proposition B.4.1.

It remains to prove that the sequence (Y_m) also converges in law with limit X. We again consider the Lipschitz function f, with Lipschitz constant C, and write

$$|\,\mathbf{E}(f(\mathrm{X}_n)) - \mathbf{E}(f(\mathrm{X}_{n,m}))| \leqslant \mathrm{C}\,\mathbf{E}(\|\mathrm{E}_{n,m}\|).$$

For a fixed m, we let $n \to +\infty$. Since we have proved that (X_n) converges to X, we deduce by the first assumption that

$$|\,\mathbf{E}(f(\mathrm{X})) - \mathbf{E}(f(\mathrm{Y}_m))| \leqslant \mathrm{C} \limsup_{n\to+\infty} \mathbf{E}(\|\mathrm{E}_{n,m}\|).$$

Since the right-hand side converges to 0 by the second assumption, we conclude that

$$\mathbf{E}(f(\mathrm{Y}_m)) \to \mathbf{E}(f(\mathrm{X}))$$

and, finally, that (Y_m) converges to X. □

Remark B.4.5 If one knows that Y_n converges in law, one also obtains convergence in law (by a straightforward adaptation of the previous argument) if Assumption (2) of the proposition is replaced by

$$\lim_{m\to+\infty} \limsup_{n\to+\infty} \mathbf{P}(\|\mathrm{E}_{n,m}\| > \delta) = 0 \tag{B.5}$$

for any $\delta > 0$; see again [10, Th. 3.2].

Remark B.4.6 Although we have stated all results in the case where the random variables are defined on the same probability space, the proofs do not rely on this fact, and the statements apply also if they are defined on spaces depending on n, with obvious adaptations of the assumptions. For instance, in the last statement, we can take X_n and $X_{n,m}$ to be defined on a space Ω_n (independent of m) and the second assumption means that

$$\lim_{m\to+\infty} \limsup_{n\to+\infty} \mathbf{E}_n(|E_{n,m}|) = 0.$$

B.5 Convergence in Law in a Finite-Dimensional Vector Space

We will use two important criteria for convergence in law for random variables with values in a finite-dimensional real vector space V, which both amount to testing (B.1) for a restricted set of functions. Another important criterion applies to variables with values in a compact topological group and is reviewed in Section B.6.

The first result is valid in all cases and is based on the Fourier transform. Given an integer $m \geqslant 1$ and a probability measure μ on $\mathbf{R}^m$, recall that the *characteristic function* (or *Fourier transform*) of μ is the function

$$\varphi_\mu \colon \mathbf{R}^m \longrightarrow \mathbf{C}$$

defined by

$$\varphi_\mu(t) = \int_{\mathbf{R}^m} e^{it\cdot x} d\mu(x),$$

where $t \cdot x = t_1x_1 + \cdots + t_mx_m$ is the standard inner product. This is a continuous bounded function on $\mathbf{R}^m$. For a random vector X with values in $\mathbf{R}^m$, we denote by φ_X the characteristic function of $X(\mathbf{P})$, namely,

$$\varphi_X(t) = \mathbf{E}(e^{it\cdot X}).$$

We state two (obviously equivalent) versions of P. Lévy's theorem for convenience:

Theorem B.5.1 (Lévy Criterion) *Let $m \geqslant 1$ be an integer.*

(1) *Let (μ_n) be a sequence of probability measures on $\mathbf{R}^m$, and let μ be a probability measure on $\mathbf{R}^m$. Then (μ_n) converges weakly to μ if and only if, for any $t \in \mathbf{R}^m$, we have*

$$\varphi_{\mu_n}(t) \longrightarrow \varphi_\mu(t)$$

as $n \to +\infty$.

(2) *Let $(\Omega, \Sigma, \mathbf{P})$ be a probability space. Let $(X_n)_{n\geqslant 1}$ be $\mathbf{R}^m$-valued random vectors on Ω, and let X be an $\mathbf{R}^m$-valued random vector. Then (X_n) converges in law to X if and only if, for all $t \in \mathbf{R}^m$, we have*

$$\mathbf{E}(e^{it\cdot \mathrm{X}_n}) \longrightarrow \mathbf{E}(e^{it\cdot \mathrm{X}}).$$

For a proof, see, for example, [9, Th. 26.3] in the case $m = 1$.

Remark B.5.2 In fact, the precise version of Lévy's Theorem does not require to know in advance the limit of the sequence: if a sequence (μ_n) of probability measures is such that, for all $t \in \mathbf{R}^m$, we have

$$\varphi_{\mu_n}(t) \longrightarrow \varphi(t)$$

for *some* function φ, *and* if φ is continuous at 0, then one can show that φ is the characteristic function of a probability measure μ (and hence that μ_n converges weakly to μ); see for instance [9, p. 350, cor. 1]. So, for instance, it is not necessary to know beforehand that $\varphi(t) = e^{-t^2/2}$ is the characteristic function of a probability measure in order to prove the Central Limit Theorem using Lévy's Criterion.

Lemma B.5.3 *Let $m \geqslant 1$ be an integer. Let $(\mathrm{X}_n)_{n\geqslant 1}$ be a sequence of random variables with values in $\mathbf{R}^m$ on some probability space. Let (β_n) be a sequence of positive real numbers such that $\beta_n \to 0$ as $n \to +\infty$. If (X_n) converges in law to an $\mathbf{R}^m$-valued random variable X, then for any sequence (Y_n) of $\mathbf{R}^m$-valued random variables such that $\|\mathrm{X}_n - \mathrm{Y}_n\|_\infty \leqslant \beta_n$ for all $n \geqslant 1$, the random variables Y_n converge to X.*

Proof We use Lévy's criterion. We fix $t \in \mathbf{R}^m$ and write

$$\mathbf{E}(e^{it\cdot \mathrm{Y}_n}) - \mathbf{E}(e^{it\cdot \mathrm{X}}) = \mathbf{E}(e^{it\cdot \mathrm{Y}_n} - e^{it\cdot \mathrm{X}_n}) + \mathbf{E}(e^{it\cdot \mathrm{X}_n} - e^{it\cdot \mathrm{X}}).$$

By Lévy's Theorem and our assumption on the convergence of the sequence (X_n), the second term on the right converges to 0 as $n \to +\infty$. For the first, we can simply apply the dominated convergence theorem to derive the same conclusion: we have

$$\|\mathrm{X}_n - \mathrm{Y}_n\|_\infty \leqslant \beta_n \to 0,$$

hence

$$e^{it\cdot \mathrm{Y}_n} - e^{it\cdot \mathrm{X}_n} = e^{it\cdot \mathrm{Y}_n}\big(1 - e^{it\cdot (\mathrm{X}_n - \mathrm{Y}_n)}\big) \to 0$$

(pointwise) as $n \to +\infty$. Moreover, we have

$$\left| e^{it\cdot \mathrm{Y}_n} - e^{it\cdot \mathrm{X}_n} \right| \leqslant 2$$

for all $n \geqslant 1$. Hence the dominated convergence theorem implies that the expectation $\mathbf{E}(e^{it\cdot \mathrm{Y}_n} - e^{it\cdot \mathrm{X}_n})$ converges to 0.

Lévy's Theorem applied once more allows us to conclude that (Y_n) converges in law to X, as claimed. □

The second convergence criterion is known as the *method of moments*. It is more restrictive than Lévy's criterion, but is sometimes analytically more flexible, especially because it is often more manageable when there is no independence assumptions.

Definition B.5.4 (Mild measure) Let μ be a probability measure on $\mathbf{R}^m$. We will say that μ is *mild*[2] if the *absolute moments*

$$\mathrm{M}^a_{\boldsymbol{k}}(\mu) = \int_{\mathbf{R}^m} |x_1|^{k_1} \cdots |x_m|^{k_m} d\mu(x_1, \ldots, x_m)$$

exist for all tuples of nonnegative integers $\boldsymbol{k} = (k_1, \ldots, k_m)$ and if there exists $\delta > 0$ such that the power series

$$\sum_{k_i \geqslant 0} \mathrm{M}^a_{\boldsymbol{k}}(\mu) \frac{z_1^{k_1} \cdots z_m^{k_m}}{k_1! \cdots k_m!}$$

converges in the region

$$\{(z_1, \ldots, z_m) \in \mathbf{C}^m \mid |z_i| \leqslant \delta\}.$$

If a measure μ is mild, then it follows in particular that the moments

$$\mathrm{M}_{\boldsymbol{k}}(\mu) = \int_{\mathbf{R}^m} x_1^{k_1} \cdots x_m^{k_m} d\mu(x_1, \ldots, x_m)$$

exist for all $\boldsymbol{k} = (k_1, \ldots, k_m)$ with k_i nonnegative integers.

If X is a random variable, we will say as usual that a random vector $\mathrm{X} = (\mathrm{X}_1, \ldots, \mathrm{X}_m)$ is *mild* if its law $\mathrm{X}(\mathbf{P})$ is mild. The moments and absolute moments are then

$$\mathrm{M}_{\boldsymbol{k}}(\mathrm{X}) = \mathbf{E}(\mathrm{X}_1^{k_1} \cdots \mathrm{X}_m^{k_m}), \quad \text{and} \quad \mathrm{M}^a_{\boldsymbol{k}}(\mathrm{X}) = \mathbf{E}(|\mathrm{X}_1|^{k_1} \cdots |\mathrm{X}_m|^{k_m}).$$

We again give two versions of the method of moments for weak convergence when the limit is mild:

Theorem B.5.5 (Method of moments) *Let $m \geqslant 1$ be an integer.*

(1) *Let (μ_n) be a sequence of probability measures on $\mathbf{R}^m$ such that all moments $\mathrm{M}_{\boldsymbol{k}}(\mu_n)$ exist, and let μ be a probability measure on $\mathbf{R}^m$.* Assume that μ is mild. *Then (μ_n) converges weakly to μ if for any m-tuple $\boldsymbol{k}$ of nonnegative integers, we have*

$$\mathrm{M}_{\boldsymbol{k}}(\mu_n) \longrightarrow \mathrm{M}_{\boldsymbol{k}}(\mu)$$

as $n \to +\infty$.

[2] There doesn't seem to be an especially standard name for this notion.

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

(2) *Let* $(\Omega, \Sigma, \mathbf{P})$ *be a probability space. Let* $(X_n)_{n\geqslant 1}$ *be* $\mathbf{R}^m$*-valued random vectors on* Ω *such that all moments* $M_{\boldsymbol{k}}(X_n)$ *exist, and let* Y *be an* $\mathbf{R}^m$*-valued random vector.* Assume that Y is mild. *Then* (X_n) *converges in law to* Y *if for any* m*-tuple* $\boldsymbol{k}$ *of non-negative integers, we have*

$$\mathbf{E}(X_{n,1}^{k_1}\cdots X_{n,m}^{k_m}) \longrightarrow \mathbf{E}(Y_1^{k_1}\cdots Y_n^{k_m}).$$

For a proof (in the case $m = 1$), see, for instance, [9, Th. 30.2 and Th. 30.1].

This only gives one implication in comparison with the Lévy Criterion. It is often useful to have a converse, and here is one such statement:

Theorem B.5.6 (Converse of the method of moments) *Let* $(\Omega, \Sigma, \mathbf{P})$ *be a probability space.*

Let $m \geqslant 1$ *be an integer and let* $(X_n)_{n\geqslant 1}$ *be a sequence of* $\mathbf{R}^m$*-valued random vectors on* Ω *such that all moments* $M_{\boldsymbol{k}}(X_n)$ *exist and such that there exist constants* $c_{\boldsymbol{k}} \geqslant 0$ *with*

$$\mathbf{E}(|X_{n,1}|^{k_1}\cdots|X_{n,m}|^{k_m}) \leqslant c_{\boldsymbol{k}} \tag{B.6}$$

for all $n \geqslant 1$*. Assume that* X_n *converges in law to a random vector* Y*. Then* Y *is mild and for any* m*-tuple* $\boldsymbol{k}$ *of nonnegative integers, we have*

$$\mathbf{E}(X_{n,1}^{k_1}\cdots X_{n,m}^{k_m}) \longrightarrow \mathbf{E}(Y_1^{k_1}\cdots Y_n^{k_m}).$$

Proof See [9, Th 25.12 and Cor.] for a proof (again for $m = 1$). □

Example B.5.7 This converse applies, in particular, if (X_n) is a sequence of real-valued random variables given by

$$X_n = \frac{B_1 + \cdots + B_n}{\sigma_n}$$

where the variables $(B_i)_{i\geqslant 1}$ are independent and satisfy

$$\mathbf{E}(B_i) = 0, \qquad |B_i| \leqslant 1, \qquad \sigma_n^2 = \sum_{i=1}^{n} \mathbf{V}(B_i) \to +\infty \text{ as } n \to +\infty.$$

Then the Central Limit Theorem (see Theorem B.7.2) implies that the sequence (X_n) converges in law to a standard Gaussian random variable Y. Moreover, X_n is bounded (by n/σ_n), so all its moments exist. We will check that this sequence satisfies the uniform integrability condition (B.6), from which we deduce the convergence of moments

$$\lim_{n\to+\infty} \mathbf{E}(X_n^k) = \mathbf{E}(Y^k)$$

for all integers $k \geqslant 0$ (the moments of Y are described explicitly in Proposition B.7.3).

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

For any $k \geqslant 0$, there exists a constant $C_k \geqslant 0$ such that

$$|x|^k \leqslant C_k(e^x + e^{-x})$$

for all $x \in \mathbf{R}$. It follows that if we can show that there exists $D \geqslant 0$ such that

$$\mathbf{E}(e^{X_n}) \leqslant D \quad \text{and} \quad \mathbf{E}(e^{-X_n}) \leqslant D \tag{B.7}$$

for all $n \geqslant 1$, then we obtain $\mathbf{E}(|X_n|^k) \leqslant 2C_k D$ for all n, which gives the desired conclusion. Note that, since X_n is bounded, these expectations make sense, and moreover, we may assume that we only consider n large enough so that $\sigma_n \geqslant 1$.

To prove (B.7), fix more generally $t \in [-1, 1]$. Since the (B_i) are independent random variables, we have

$$\mathbf{E}(e^{tX_n}) = \prod_{i=1}^{n} \mathbf{E}\left(\exp\left(\frac{tB_i}{\sigma_n}\right)\right).$$

Since we assumed that $\sigma_n \geqslant 1$ and $|B_i| \leqslant 1$, we have $|tB_i/\sigma_n| \leqslant 1$, hence

$$\exp\left(\frac{tB_i}{\sigma_n}\right) \leqslant 1 + \frac{tB_i}{\sigma_n} + \frac{t^2 B_i^2}{\sigma_n^2}$$

(because $e^x \leqslant 1 + x + x^2$ for $|x| \leqslant 1$, as can be checked using basic calculus). We then obtain further

$$\mathbf{E}(e^{tX_n}) \leqslant \prod_{i=1}^{n} \left(1 + \frac{t^2}{\sigma_n^2}\mathbf{E}(B_i^2)\right)$$

since $\mathbf{E}(B_i) = 0$. Using $1 + x \leqslant e^x$, this leads to

$$\mathbf{E}(e^{tX_n}) \leqslant \exp\left(\frac{t^2}{\sigma_n^2}\sum_{i=1}^{n}\mathbf{E}(B_i^2)\right) = \exp(t^2).$$

Applying this with $t = 1$ and $t = -1$, we get (B.7) with $D = e$, hence also (B.6), for all n large enough.

Remark B.5.8 In the case $m = 2$, one often deals with random variables that are naturally seen as complex-valued, instead of $\mathbf{R}^2$-valued. In that case, it is sometimes quite useful to use the complex moments

$$\tilde{M}_{k_1,k_2}(X) = \mathbf{E}(X^{k_1}\bar{X}^{k_2})$$

of a $\mathbf{C}$-valued random variable instead of $M_{k_1,k_2}(X)$. The corresponding statements are that X is mild if and only if the power series

$$\sum\sum_{k_1,k_2 \geqslant 0} \tilde{M}_{k_1,k_2}(X)\frac{z_1^{k_1} z_2^{k_2}}{k_1!\,k_2!}$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

converges in a region

$$\{(z_1, z_2) \in \mathbf{C} \mid |z_1| \leqslant \delta, \quad |z_2| \leqslant \delta\}$$

for some $\delta > 0$, and that if X is mild, then (X_n) converges weakly to X if

$$\tilde{M}_{k_1,k_2}(X_n) \longrightarrow \tilde{M}_{k_1,k_2}(X)$$

for all $k_1, k_2 \geqslant 0$. Example B.5.7 extends to the complex-valued case.

Example B.5.9 (1) Any bounded random vector is mild. Indeed, if $\|X\|_\infty \leqslant B$, say, then we get

$$M^a_{\boldsymbol{k}}(X) \leqslant B^{k_1+\cdots+k_m},$$

and therefore

$$\sum\sum_{k_i \geqslant 0} M^a_{\boldsymbol{k}}(\mu) \frac{|z_1|^{k_1} \cdots |z_m|^{k_m}}{k_1! \cdots k_m!} \leqslant e^{B|z_1|+\cdots+B|z_m|},$$

so that the power series converges, in that case, for all $\boldsymbol{z} \in \mathbf{C}^m$.

(2) Any Gaussian random vector is mild (see Section B.7).

(3) If X is mild, and Y is another random vector with $|Y_i| \leqslant |X_i|$ (almost surely) for all i, then Y is also mild.

B.6 The Weyl Criterion

One important special case of convergence in law is known as *equidistribution* in the context of topological groups in particular. We only consider compact groups here for simplicity. Let G be such a group. Then there exists on G a unique Borel probability measure μ_G which is invariant under left (and right) translations: for any integrable function $f : G \longrightarrow \mathbf{C}$ and for any fixed $g \in G$, we have

$$\int_G f(gx)d\mu_G(x) = \int_G f(xg)d\mu_G(x) = \int_G f(x)d\mu_G(x).$$

This measure is called the (probability) Haar measure on G (see, e.g., [17, VII, §1, n. 2, th. 1 and prop. 2]).

If a G-valued random variable X is distributed according to μ_G, one says that X is *uniformly distributed* on G.

Example B.6.1 (1) Let $G = \mathbf{S}^1$ be the multiplicative group of complex numbers of modulus 1. This group is isomorphic to $\mathbf{R}/\mathbf{Z}$ by the isomorphism $\theta \mapsto e(\theta)$. The measure μ_G is then identified with the Lebesgue measure $d\theta$ on $\mathbf{R}/\mathbf{Z}$. In other words, for any integrable function $f\colon \mathbf{S}^1 \to \mathbf{C}$, we have

$$\int_{\mathbf{S}^1} f(z)d\mu_{\mathrm{G}}(z) = \int_{\mathbf{R}/\mathbf{Z}} f(e(\theta))d\theta = \int_0^1 f(e(\theta))d\theta.$$

(2) If $(\mathrm{G}_i)_{i\in\mathrm{I}}$ is any family of compact groups, each with a probability Haar measure μ_i, then the (possibly infinite) tensor product

$$\bigotimes_{i\in\mathrm{I}} \mu_i$$

is the probability Haar measure μ on the product G of the groups G_i. Probabilistically, one would interpret this as saying that μ is the law of a family (X_i) of *independent* random variables, where each X_i is uniformly distributed on G_i.

(3) Let G be the nonabelian compact group $\mathrm{SU}_2(\mathbf{C})$, that is,

$$\mathrm{G} = \left\{ \begin{pmatrix} \alpha & \bar{\beta} \\ -\beta & \bar{\alpha} \end{pmatrix} \mid \alpha,\ \beta \in \mathbf{C},\ |\alpha|^2 + |\beta|^2 = 1 \right\}.$$

Writing $\alpha = a + ib$, $\beta = c + id$, we can identify G, as a topological space, with the unit 3-sphere

$$\{(a,b,c,d) \in \mathbf{R}^4 \mid a^2 + b^2 + c^2 + d^2 = 1\}$$

in $\mathbf{R}^4$. Then the left multiplication by some element on G is the restriction of a rotation of $\mathbf{R}^4$. Hence the surface (Lebesgue) measure μ_0 on the 3-sphere is a Borel-invariant measure on G. By uniqueness, we see that the probability Haar measure on G is

$$\mu = \frac{1}{2\pi^2}\mu_0$$

(since the surface area of the 3-sphere is $2\pi^2$).

Consider now the trace $\mathrm{Tr}\colon \mathrm{G} \longrightarrow \mathbf{R}$, which is given by $(a,b,c,d) \mapsto 2a$ in the sphere coordinates. One can show that the direct image $\mathrm{Tr}_*(\mu)$ is the so-called *Sato–Tate* measure

$$\mu_{\mathrm{ST}} = \frac{1}{\pi}\sqrt{1 - \frac{x^2}{4}}dx,$$

supported on $[-2,2]$ (for probabilists, this is also a semicircle law); equivalently, if we write the trace as

$$\mathrm{Tr}(g) = 2\cos(\theta)$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

for a unique $\theta \in [0, \pi]$, then this measure is identified with the measure

$$\frac{2}{\pi} \sin^2 \theta d\theta$$

on $[0, \pi]$ (for a proof, see, e.g., [18, Ch. 9, p. 58, example]). One obtains from either description of μ_{ST} the expectation and variance

$$\int_{\mathbf{R}} t d\mu_{\mathrm{ST}} = 0 \quad \text{and} \quad \int_{\mathbf{R}} t^2 d\mu_{\mathrm{ST}} = 1. \tag{B.8}$$

(4) If G is a finite group, then the probability Haar measure is just the normalized counting measure: for any function f on G, the integral of f is

$$\frac{1}{|\mathrm{G}|} \sum_{x \in \mathrm{G}} f(x).$$

For a topological group G, a *unitary character* χ of G is a continuous homomorphism

$$\chi : \mathrm{G} \longrightarrow \mathbf{S}^1.$$

The *trivial character* is the character $g \mapsto 1$ of G. The set of all characters of G is denoted $\widehat{\mathrm{G}}$. It has a structure of abelian group by multiplication of functions. If G is locally compact, then $\widehat{\mathrm{G}}$ is a locally compact topological group with the topology of uniform convergence on compact sets (for this theory, see, e.g., [19]).

In general, $\widehat{\mathrm{G}}$ may be reduced to the trivial character (this is the case if $\mathrm{G} = \mathrm{SL}_2(\mathbf{R})$, for instance). Assume now that G is locally compact and abelian. Then it is a fundamental fact (known as *Pontryagin duality*; see, e.g., [70, §7.3] for a survey or [19, II, §1, n. 5, th. 2] for the details) that there are “many” characters, in a suitable sense. If G is compact, then a simple version of this assertion is that $\widehat{\mathrm{G}}$ is an orthonormal basis of the space $\mathrm{L}^2(\mathrm{G}, \mu)$, where μ is the probability Haar measure on G.

For an integrable function $f \in \mathrm{L}^1(\mathrm{G}, \mu)$, its *Fourier transform* is the function $\widehat{f} : \widehat{\mathrm{G}} \longrightarrow \mathbf{C}$ defined by

$$\widehat{f}(\chi) = \int_{\mathrm{G}} f(x) \overline{\chi(x)} d\mu(x)$$

for all $\chi \in \widehat{\mathrm{G}}$. For a compact commutative group G, and $f \in \mathrm{L}^2(\mathrm{G}, \mu)$, we have

$$f = \sum_{\chi \in \widehat{\mathrm{G}}} \widehat{f}(\chi) \chi,$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

as a series converging in $L^2(G, \mu)$. It follows easily that a function $f \in L^1(G, \mu)$ is almost everywhere constant if and only if $\widehat{f}(\chi) = 0$ for all $\chi \neq 1$.

The following relation is immediate from the invariance of Haar measure: for f integrable and any fixed $y \in G$, if we let $g(x) = f(xy)$, then g is well defined as an integrable function, and

$$\widehat{g}(\chi) = \int_G f(xy)\overline{\chi(x)}d\mu(x) = \chi(y)\int_G f(x)\overline{\chi(x)}d\mu(x) = \chi(y)\widehat{f}(y).$$

Example B.6.2 (1) The characters of $\mathbf{S}^1$ are given by

$$z \mapsto z^m$$

for $m \in \mathbf{Z}$. Equivalently, the characters of $\mathbf{R}/\mathbf{Z}$ are given by $x \mapsto e(hx)$, where $e(z) = \exp(2i\pi z)$. More generally, the characters of $(\mathbf{R}/\mathbf{Z})^n$ are of the form

$$x = (x_1, \ldots, x_n) \mapsto e(h_1 x_1 + \cdots + h_n x_n) = e(h \cdot x)$$

for some (unique) $h \in \mathbf{Z}^n$ (see, e.g., [19, p. 236, cor. 3]).

(2) If $(G_i)_{i\in I}$ is any family of compact groups, each with the probability Haar measure μ_i, then the characters of the product G of the G_i are given in a unique way as follows: take a finite subset S of I, and for any $i \in I$, pick a nontrivial character χ_i of G_i, then define

$$\chi(x) = \prod_{i\in S} \chi_i(x_i)$$

for any $x = (x_i)_{i\in I}$ in G. Here the trivial character corresponds to $S = \emptyset$. See, for example, [70, Example 5.6.10] for a proof.

In particular, if I is a finite set, this computation shows that the group of characters of G is isomorphic to the product of the groups of characters of the G_i and that the isomorphism is such that a family (χ_i) of characters of the groups G_i is mapped to the character

$$(x_i) \mapsto \prod_{i\in I} \chi_i(x_i).$$

(3) If G is a finite abelian group, then the group $\widehat{G}$ of characters of G is also finite, and it is isomorphic to G. This can be seen from the structure theorem for finite abelian groups, which shows that any finite abelian group is a direct product of some finite cyclic groups (see, e.g., [103, Th. B-3.13]) combined with the previous example and the explicit computation of the dual group of a finite cyclic group $\mathbf{Z}/q\mathbf{Z}$ for $q \geqslant 1$: an isomorphism from $\mathbf{Z}/q\mathbf{Z}$ to $\widehat{\mathbf{Z}/q\mathbf{Z}}$ is given by sending $a \pmod q$ to the character

$$x \mapsto e\left(\frac{ax}{q}\right),$$

which is well defined because replacing a and x by other integers congruent modulo q does not change the value of $e(ax/q)$.

In this case, one can also prove elementarily that the characters form an orthonormal basis of the finite-dimensional vector space $C(G)$ of complex-valued functions on G, which in this case has the inner product

$$\langle f, g\rangle = \frac{1}{|G|}\sum_{x\in G} f(x)\overline{g(x)}.$$

Indeed, one can also reduce to the case of cyclic groups by checking that there is a unique isomorphism

$$C(G_1)\otimes C(G_2) \to C(G_1\times G_2)$$

such that a pure tensor $f_1 \otimes f_2$ is mapped to the function $(x_1, x_2) \mapsto f_1(x_1)f_2(x_2)$. In particular, the characters of $G_1 \times G_2$ (which belong to $C(G_1 \times G_2)$) correspond under this isomorphism to the pure tensors $\chi_1 \otimes \chi_2$.

In addition, under this isomorphism, we have

$$\langle f_1\otimes f_2, g_1\otimes g_2\rangle = \langle f_1, g_1\rangle\, \langle f_2, g_2\rangle$$

for any functions f_i and g_i on G_i. This implies that if the characters of G_1 and those of G_2 form orthornomal bases of their respective spaces of functions, then so do the characters of $G_1 \times G_2$.

And in the case of $G = \mathbf{Z}/q\mathbf{Z}$, we can simply compute using the explicit description of the characters $\chi_a : x \mapsto e(ax/q)$ for $a \in \mathbf{Z}/q\mathbf{Z}$ that

$$\langle \chi_a, \chi_b\rangle = \frac{1}{q}\sum_{0\leqslant x\leqslant q-1} e\left(\frac{ax}{q}\right) e\left(-\frac{bx}{q}\right),$$

which is equal to 1 if $a = b$, and otherwise is

$$\frac{1-e(q(a-b)/q)}{1-e((a-b)/q)}$$

by summing a finite geometric sum, and is therefore zero, as we wanted.

The Weyl Criterion is a criterion for a sequence of G-valued random variables to converge in law to a uniformly distributed random variable on G. We state it for compact abelian groups only:

Theorem B.6.3 (Weyl's Criterion) *Let* G *be a compact abelian group. A sequence* (X_n) *of* G*-valued random variables converges in law to a uniformly distributed random variable on* G *if and only if, for any nontrivial character* χ *of* G*, we have*

$$\lim_{n\to+\infty} \mathbf{E}(\chi(X_n)) \longrightarrow 0.$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

Remark B.6.4 (1) Note that the orthogonality of characters implies that

$$\int_{\mathrm{G}} \chi(x) d\mu_{\mathrm{G}}(x) = \langle \chi, 1 \rangle = 0$$

for any nontrivial character χ of G. Hence the Weyl Criterion has the same flavor of Lévy's Criterion (note that, for any $t \in \mathbf{R}^m$, the function $x \mapsto e^{ix\cdot t}$ is a character of $\mathbf{R}^m$).

(2) If G is compact, but not necessarily abelian, there is a version of the Weyl Criterion using as "test functions" the traces of irreducible finite-dimensional representations of G (see [70, §5.5] for an account).

The best-known example of application of the Weyl Criterion is to prove the following equidistribution theorem of Kronecker:

Theorem B.6.5 (Kronecker) *Let $d \geqslant 1$ be an integer. Let z be an element of $\mathbf{R}^d$ and let* T (resp. $\widetilde{\mathrm{T}}$) *be the closure of the subgroup of* $(\mathbf{R}/\mathbf{Z})^d$ *generated by the class of z* (resp. generated by the classes of the elements yz for $y \in \mathbf{R}$).

(1) *As* $\mathrm{N} \to +\infty$, *the probability measures on* $(\mathbf{R}/\mathbf{Z})^d$ *defined by*

$$\frac{1}{\mathrm{N}} \sum_{0 \leqslant n < \mathrm{N}} \delta_{nz}$$

converge in law to the probability Haar measure on T.

(2) *Let* λ *denote the Lebesgue measure on* $\mathbf{R}$. *As* $\mathrm{X} \to +\infty$, *the probability measures* μ_{X} *on* $(\mathbf{R}/\mathbf{Z})^d$ *defined by*

$$\mu_{\mathrm{X}}(\mathrm{A}) = \frac{1}{\mathrm{X}} \lambda(\{x \in [0,\mathrm{X}] \mid xz \in \mathrm{A}\}),$$

for a measurable subset A *of* $(\mathbf{R}/\mathbf{Z})^d$, *converge in law to the probability Haar measure on* $\widetilde{\mathrm{T}}$.

Proof We only prove the "continuous" version in (2), since the first one is easier (and better known). First note that each probability measure μ_{X} has support contained in $\widetilde{\mathrm{T}}$ by definition, so it can be viewed as a measure on $\widetilde{\mathrm{T}}$.

From the theory of compact abelian groups, we know that any character χ of $\widetilde{\mathrm{T}}$ can be extended to a character of $(\mathbf{R}/\mathbf{Z})^d$ (see, e.g., [19, p. 226, th. 4], applied to the exact sequence $1 \to \widetilde{\mathrm{T}} \to (\mathbf{R}/\mathbf{Z})^d$), which is therefore of the form

$$v \mapsto e(n \cdot v)$$

for some $n \in \mathbf{Z}^d$ (by Example B.6.2, (1)). We then have

$$\int_{(\mathbf{R}/\mathbf{Z})^d} \chi(v) d\mu_{\mathrm{X}}(v) = \frac{1}{\mathrm{X}} \int_0^{\mathrm{X}} e((x\,n) \cdot z) dx.$$

Suppose that χ is a nontrivial character of $\widetilde{\mathrm{T}}$; since the classes of yz for $y \in \mathbf{R}$ generate a dense subgroup of $\widetilde{\mathrm{T}}$, we have then $n \cdot z \neq 0$. Hence

$$\int_{(\mathbf{R}/\mathbf{Z})^d} \chi(v) d\mu_{\mathrm{X}}(v) = \frac{1}{\mathrm{X}} \frac{e((\mathrm{X}\,n) \cdot z) - 1}{2i\pi n \cdot z} \to 0$$

as $\mathrm{X} \to +\infty$. We conclude by an application of the Weyl Criterion. □

Example B.6.6 In order to apply Theorem B.6.5 in practice, we need to identify the subgroup T (or $\widetilde{\mathrm{T}}$). The following special cases are quite often sufficient (writing $z = (z_1, \ldots, z_d) \in \mathbf{R}^d$):

1. we have $\mathrm{T} = (\mathbf{R}/\mathbf{Z})^d$ if and only if $(1, z_1, \ldots, z_d)$ are $\mathbf{Q}$-linearly independent;
2. we have $\widetilde{\mathrm{T}} = (\mathbf{R}/\mathbf{Z})^d$ if and only if $(z_1, \ldots, z_d)$ are $\mathbf{Q}$-linearly independent.

For instance, if $d = 1$, then the first condition means that $z = z_1$ is irrational, and the second means that z is nonzero.

We check (1), leaving (2) as an exercise. If $(1, z_1, \ldots, z_d)$ are *not* $\mathbf{Q}$-linearly independent, then multiplying a nontrivial linear dependency relation with a suitable nonzero integer, we obtain a relation

$$m_0 + \sum_{i=1}^{d} m_i z_i = 0,$$

where $m_i \in \mathbf{Z}$ and not all m_i are zero, in fact, not all m_i with $i \geqslant 1$ are zero (since this would also imply that $m_0 = 0$). Then the class of nz modulo $\mathbf{Z}^d$ is, for all $n \in \mathbf{Z}$, an element of the proper closed subgroup

$$\{x = (x_1, \ldots, x_d) \in (\mathbf{R}/\mathbf{Z})^d \mid m_1 x_1 + \cdots + m_d x_d = 0\},$$

which implies that T is also contained in that subgroup, hence is not all of $(\mathbf{R}/\mathbf{Z})^d$.

Conversely, a simple argument is to check that if $(1, z_1, \ldots, z_d)$ are $\mathbf{Q}$-linearly independent, then a direct application of the Weyl Criterion proves that the probability measures

$$\frac{1}{\mathrm{N}} \sum_{0 \leqslant n < \mathrm{N}} \delta_{nz}$$

converge in law to the probability Haar measure on $(\mathbf{R}/\mathbf{Z})^d$ (because nontrivial characters of this group correspond to $(m_i) \in \mathbf{Z}^d$, and the integral against the measure above is

$$\frac{1}{N}\sum_{n=1}^{N} e((m_1 z_1 + \cdots + m_d z_d)n),$$

where the real number $m_1 z_1 + \cdots + m_d z_d$ is not an integer by the linear independence, so that the sum tends to 0 by summing a finite geometric series).

B.7 Gaussian Random Variables

By definition, a random vector X with values in $\mathbf{R}^m$ is called a *(centered) Gaussian vector* if there exists a nonnegative quadratic form Q on $\mathbf{R}^m$ such that the characteristic function φ_X of X is of the form

$$\varphi_X(t) = e^{-Q(t)/2}$$

for $t \in \mathbf{R}^m$. The quadratic form can be recovered from X by the relation

$$Q(t_1, \ldots, t_m) = \sum_{1 \leqslant i,j \leqslant m} a_{i,j} t_i t_j,$$

with $a_{i,j} = \mathbf{E}(X_i X_j)$, and the (symmetric) matrix $(a_{i,j})_{1\leqslant i,j\leqslant m}$ is called the *correlation matrix* of X. The components X_i of X are independent if and only if $a_{i,j} = 0$ if $i \neq j$, that is, if and only if the components of X are orthogonal.

If X is a Gaussian random vector, then X is mild, and in fact

$$\sum_{\boldsymbol{k}} M_{\boldsymbol{m}}(X) \frac{t_1^{k_1} \cdots t_m^{k_m}}{k_1! \cdots k_m!} = \mathbf{E}(e^{t\cdot X}) = e^{Q(t)/2}$$

for $t \in \mathbf{R}^m$, so that the power series converges on all of $\mathbf{C}^m$. The Laplace transform $\psi_X(z) = \mathbf{E}(e^{z\cdot X})$ is also defined for all $z \in \mathbf{C}^m$, and in fact

$$\mathbf{E}(e^{z\cdot X}) = e^{Q(z)/2}. \tag{B.9}$$

For $m = 1$, this means that a random variable is a centered Gaussian if and only if there exists $\sigma \geqslant 0$ such that

$$\varphi_X(t) = e^{-\sigma^2 t/2}, \tag{B.10}$$

and in fact we have

$$\mathbf{E}(X^2) = \mathbf{V}(X) = \sigma^2.$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

If $\sigma = 1$, then we say that X is a *standard Gaussian random variable* (also sometimes called a standard normal random variable). We then have

$$\mathbf{P}(a < \mathrm{X} < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx$$

for all real numbers $a < b$.

Exercise B.7.1 We recall a standard proof of the fact that the measure on $\mathbf{R}$ given by

$$\mu = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

is indeed a Gaussian probability measure with variance 1.

(1) Define

$$\varphi(t) = \varphi_\mu(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} e^{itx - x^2/2} dx$$

for $t \in \mathbf{R}$. Prove that φ is of class C^1 on $\mathbf{R}$ and satisfies $\varphi'(t) = -t\varphi(t)$ for all $t \in \mathbf{R}$ and $\varphi(0) = 1$.

(2) Deduce that $\varphi(t) = e^{-t^2/2}$ for all $t \in \mathbf{R}$. [**Hint**: This is an elementary argument with ordinary differential equations, but because the order is 1, one can define $g(t) = e^{t^2/2}\varphi(t)$ and check by differentiation that $g'(t) = 0$ for all $t \in \mathbf{R}$.]

We will use the following simple version of the Central Limit Theorem:

Theorem B.7.2 *Let* $\mathrm{B} \geqslant 0$ *be a fixed real number. Let* (X_n) *be a sequence of independent real-valued random variables with* $|\mathrm{X}_n| \leqslant \mathrm{B}$ *for all n. Let*

$$\alpha_n = \mathbf{E}(\mathrm{X}_n) \quad \text{and} \quad \beta_n = \mathbf{V}(\mathrm{X}_n^2).$$

Let $\sigma_\mathrm{N} \geqslant 0$ *be defined by*

$$\sigma_\mathrm{N}^2 = \beta_1 + \cdots + \beta_\mathrm{N}$$

for $\mathrm{N} \geqslant 1$. *If* $\sigma_\mathrm{N} \to +\infty$ *as* $n \to +\infty$, *then the random variables*

$$\mathrm{Y}_\mathrm{N} = \frac{(\mathrm{X}_1 - \alpha_1) + \cdots + (\mathrm{X}_\mathrm{N} - \alpha_\mathrm{N})}{\sigma_\mathrm{N}}$$

converge in law to a standard Gaussian random variable.

Proof Although this is a very simple case of the general Central Limit Theorem for sums of independent random variables (indeed, even of Lyapunov's well-known version), we give a proof using Lévy's criterion for convenience.

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

First of all, we may assume that $\alpha_n = 0$ for all n by replacing X_n by $X_n - \alpha_n$ (up to replacing B by 2B, since $|\alpha_n| \leqslant B$).

By independence of the variables (X_n), the characteristic function φ_N of Y_N is given by

$$\varphi_N(t) = \mathbf{E}(e^{itY_N}) = \prod_{1\leqslant n\leqslant N} \mathbf{E}(e^{itX_n/\sigma_N}).$$

Fix $t \in \mathbf{R}$. Since tX_n/σ_N is bounded (because t is fixed), we have a Taylor expansion around 0 of the form

$$e^{itX_n/\sigma_N} = 1 + \frac{itX_n}{\sigma_N} - \frac{t^2X_n^2}{2\sigma_N^2} + O\left(\frac{|t|^3|X_n|^3}{\sigma_N^3}\right),$$

for $1 \leqslant n \leqslant N$.

Consequently, we obtain

$$\varphi_{X_n}\left(\frac{t}{\sigma_N}\right) = \mathbf{E}(e^{itX_n/\sigma_N}) = 1 - \frac{1}{2}\left(\frac{t}{\sigma_N}\right)^2 \mathbf{E}(X_n^2) + O\left(\left(\frac{|t|}{\sigma_N}\right)^3 \mathbf{E}(|X_n|^3)\right).$$

Observe that with our assumption, we have

$$\mathbf{E}(|X_n|^3) \leqslant B\,\mathbf{E}(X_n^2) = B\beta_n.$$

Moreover, for N large enough (depending on t, but t is fixed), the modulus of

$$-\frac{1}{2}\left(\frac{t}{\sigma_N}\right)^2 \mathbf{E}(X_n^2) + O\left(\left(\frac{|t|}{\sigma_N}\right)^3 \mathbf{E}(|X_n|^3)\right)$$

is less than 1, so that we can use Proposition A.2.2 and deduce that

$$\begin{aligned}
\varphi_N(t) &= \exp\left(\sum_{n=1}^{N} \log \mathbf{E}(e^{itX_n/\sigma_N})\right)\\
&= \exp\left(-\frac{t^2}{2\sigma_N}\sum_{n=1}^{N}\beta_n + O\left(\frac{B|t|^3}{\sigma_N^3}\sum_{n=1}^{N}\beta_n\right)\right)\\
&= \exp\left(-\frac{t^2}{2} + O\left(\frac{B|t|^3}{\sigma_N}\right)\right) \longrightarrow \exp(-t^2/2)
\end{aligned}$$

as $N \to +\infty$; we conclude, then, by Lévy's Criterion and (B.10). ☐

If one uses directly the method of moments to get convergence in law to a Gaussian random variable, it is useful to know the values of their moments. We only state the one-dimensional and the simplest complex case:

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

Proposition B.7.3 (1) *Let* X *be a real-valued Gaussian random variable with expectation* 0 *and variance* σ^2*. For* $k \geqslant 0$*, we have*

$$\mathbf{E}(\mathrm{X}^k) = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ \sigma^k \frac{k!}{2^{k/2}(k/2)!} = \sigma^k \cdot (1 \cdot 3 \cdot \cdots \cdot (k-1)) & \text{if } k \text{ is even.} \end{cases}$$

(1) *Let* X *be a complex-valued Gaussian random variable with covariance matrix*

$$\begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix}$$

for some $\sigma > 0$*. For* $k \geqslant 0$ *and* $l \geqslant 0$*, we have*

$$\mathbf{E}(\mathrm{X}^k \bar{\mathrm{X}}^l) = \begin{cases} 0 & \text{if } k \neq l, \\ \sigma^k 2^k k! & \text{if } k = l. \end{cases}$$

Exercise B.7.4 Prove this proposition.

B.8 Sub-Gaussian Random Variables

Gaussian random variables have many remarkable properties. It is a striking fact that a number of these, especially with respect to integrability properties, are shared by a much more general class of random variables.

Definition B.8.1 (Sub-Gaussian random variable) Let $\sigma > 0$ be a real number. A real-valued random variable X is σ^2-sub-Gaussian if we have

$$\mathbf{E}(e^{t\mathrm{X}}) \leqslant e^{\sigma^2 t^2/2}$$

for all $t \in \mathbf{R}$. A complex-valued random variable X is σ^2-sub-Gaussian if $\mathrm{X} = \mathrm{Y} + i\mathrm{Z}$ with Y and Z real-valued σ^2-sub-Gaussian random variables.

If X is a real σ^2-sub-Gaussian random variable, then we obtain immediately good Gaussian-type upper bounds for the tail of the distribution: for any $b > 0$, using first a general auxiliary parameter $t > 0$, we have

$$\mathbf{P}(\mathrm{X} > b) \leqslant \mathbf{P}(e^{t\mathrm{X}} > e^{bt}) \leqslant \frac{\mathbf{E}(e^{t\mathrm{X}})}{e^{bt}} \leqslant e^{\sigma^2 t^2/2 - bt},$$

and selecting $t = -\frac{1}{2}b^2/\sigma^2$, we get

$$\mathbf{P}(\mathrm{X} > b) \leqslant e^{-\frac{1}{2}b^2/\sigma^2}.$$

The right-hand side is a standard upper bound for the probability $\mathbf{P}(\mathrm{N} > b)$ for a centered Gaussian random variable N with variance σ^2, so this inequality justifies the name "sub-Gaussian."

A Gaussian random variable is sub-Gaussian by (B.9). But there are many more examples, in particular the random variables described in the next proposition.

Proposition B.8.2 (1) *Let* X *be a complex-valued random variable and* $m > 0$ *a real number such that* $\mathbf{E}(\mathrm{X}) = 0$ *and* $|\mathrm{X}| \leqslant m$. *Then* X *is* m^2*-sub-Gaussian.*

(2) *Let* X_1 *and* X_2 *be independent random variables such that* X_i *is* σ_i^2*-sub-Gaussian. Then* $\mathrm{X}_1 + \mathrm{X}_2$ *is* $(\sigma_1^2 + \sigma_2^2)$*-sub-Gaussian.*

Proof (1) We may assume that X is real-valued, and by considering $m^{-1}\mathrm{X}$ instead of X, we may assume that $|\mathrm{X}| \leqslant 1$, and of course that X is not almost surely 0. In particular, the Laplace transform $\varphi(t) = \mathbf{E}(e^{t\mathrm{X}})$ is well defined, and $\varphi(t) > 0$ for all $t \in \mathbf{R}$. Moreover, it is easy to check that φ is smooth on $\mathbf{R}$ with

$$\varphi'(t) = \mathbf{E}(\mathrm{X}e^{t\mathrm{X}}) \quad \text{and} \quad \varphi''(t) = \mathbf{E}(\mathrm{X}^2 e^{t\mathrm{X}})$$

(by differentiating under the integral sign) and in particular

$$\varphi(0) = 1 \quad \text{and} \quad \varphi'(0) = \mathbf{E}(\mathrm{X}) = 0.$$

We now define $f(t) = \log(\varphi(t)) - \frac{1}{2}t^2$. The function f is also smooth and satisfies $f(0) = f'(0) = 0$. Moreover, we compute that

$$f''(t) = \frac{\varphi''(t)\varphi(t) - \varphi'(t)^2 - \varphi(t)^2}{\varphi(t)^2}.$$

The formula for φ'' and the condition $|\mathrm{X}| \leqslant 1$ imply that $0 \leqslant \varphi''(t) \leqslant \varphi(t)$ for all $t \in \mathbf{R}$. Therefore

$$\varphi''(t)\varphi(t) - \varphi'(t)^2 - \varphi(t)^2 \leqslant -\varphi'(t)^2 \leqslant 0,$$

and hence $f''(t) \leqslant 0$ for all $t \in \mathbf{R}$. This means that the derivative of f is decreasing, so that $f'(t) \leqslant 0$ for $t \geqslant 0$, and $f'(t) \geqslant 0$ for $t \leqslant 0$. Thus f is nondecreasing when $t \leqslant 0$ and nonincreasing when $t \geqslant 0$. In particular, we have $f(t) \leqslant f(0) = 0$ for all $t \in \mathbf{R}$, which translates exactly to the condition $\mathbf{E}(e^{t\mathrm{X}}) \leqslant e^{t^2/2}$ defining a sub-Gaussian random variable.

(2) Since X_1 and X_2 are independent and sub-Gaussian, we have

$$\mathbf{E}(e^{t(\mathrm{X}_1+\mathrm{X}_2)}) = \mathbf{E}(e^{t\mathrm{X}_1})\,\mathbf{E}(e^{t\mathrm{X}_2}) \leqslant \exp\left(\tfrac{1}{2}(\sigma_1^2 + \sigma_2^2)t^2\right)$$

for any $t \in \mathbf{R}$. ☐

Proposition B.8.3 *Let $\sigma > 0$ be a real number, and let X be a σ^2-sub-Gaussian random variable, either real or complex-valued. For any integer $k \geqslant 0$, there exists $c_k \geqslant 0$ such that*

$$\mathbf{E}(|X|^k) \leqslant c_k \sigma^k.$$

Proof The random variable $Y = \sigma^{-1}X$ is 1-sub-Gaussian. As in the proof of Theorem B.5.6 (2), we observe that there exists $c_k \geqslant 0$ such that

$$|Y|^k \leqslant c_k(e^{X_k} + e^{-X_k}),$$

and therefore

$$\sigma^{-k}\,\mathbf{E}(|X|^k) = \mathbf{E}(|Y|^k) \leqslant c_k(e^{1/2} + e^{-1/2}),$$

which gives the result. □

Remark B.8.4 A more precise argument leads to specific values of c_k. For instance, if X is real-valued, one can show that the inequality holds with $c_k = k2^{k/2}\Gamma(k/2)$.

B.9 Poisson Random Variables

Let $\lambda > 0$ be a real number. A random variable X is said to have a Poisson distribution with parameter $\lambda \in [0, +\infty[$ if and only if it is integral-valued and if, for any integer $k \geqslant 0$, we have

$$\mathbf{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

One checks immediately that

$$\mathbf{E}(X) = \lambda \quad \text{and} \quad \mathbf{V}(X) = \lambda$$

and that the characteristic function of X is

$$\varphi_X(t) = e^{-\lambda}\sum_{k\geqslant 0} e^{ikt}\frac{\lambda^k}{k!} = \exp(\lambda(e^{it} - 1)). \tag{B.11}$$

Proposition B.9.1 *Let (λ_n) be a sequence of real numbers such that $\lambda_n \to +\infty$ as $n \to +\infty$. Then*

$$\frac{X_n - \lambda_n}{\sqrt{\lambda_n}}$$

converges in law to a standard Gaussian random variable.

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

Proof Use the Lévy Criterion: the characteristic function φ_n of X_n is given by

$$\varphi_n(t) = \mathbf{E}(e^{it(X_n-\lambda_n)/\sqrt{\lambda_n}}) = \exp\Big(-it\sqrt{\lambda_n} + \lambda_n(e^{it/\sqrt{\lambda_n}} - 1)\Big)$$

for $t \in \mathbf{R}$, by (B.11). Since

$$-\frac{it}{\sqrt{\lambda_n}} + \lambda_n(e^{it/\sqrt{\lambda_n}} - 1) = it\sqrt{\lambda_n} + \lambda_n\left(\frac{it}{\sqrt{\lambda_n}} - \frac{t^2}{2\lambda_n} + \mathrm{O}\left(\frac{|t|^3}{\lambda_n^{3/2}}\right)\right)$$
$$= -\frac{t^2}{2} + \mathrm{O}\left(\frac{|t|^3}{\lambda_n^{1/2}}\right),$$

we obtain $\varphi_n(t) \to \exp(-t^2/2)$, which is the characteristic function of a standard Gaussian random variable. □

B.10 Random Series

We will need some fairly elementary results on certain random series, especially concerning almost sure convergence. We first have a well-known sufficient criterion of Kolmogorov for convergence in the case of independent summands:

Theorem B.10.1 (Kolmogorov) *Let $(X_n)_{n\geqslant 1}$ be a sequence of independent complex-valued random variables such that both series*

$$\sum_{n\geqslant 1} \mathbf{E}(X_n), \tag{B.12}$$

$$\sum_{n\geqslant 1} \mathbf{V}(X_n) \tag{B.13}$$

converge. Then the series

$$\sum_{n\geqslant 1} X_n$$

converges almost surely, and hence also in law. Moreover, its sum X is square integrable and has expectation $\sum \mathbf{E}(X_n)$.

Proof By replacing X_n with $X_n - \mathbf{E}(X_n)$, we reduce to the case where $\mathbf{E}(X_n) = 0$ for all n. Assuming that this is the case, we will prove that the series converges almost surely by checking that the sequence of partial sums

$$S_N = \sum_{1\leqslant n\leqslant N} X_n$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

is almost surely a Cauchy sequence. For this purpose, denote

$$Y_{N,M} = \sup_{1 \leqslant k \leqslant M} |S_{N+k} - S_N|$$

for N, M $\geqslant 1$. For fixed N, $Y_{N,M}$ is an increasing sequence of random variables; we denote by $Y_N = \sup_{k \geqslant 1} |S_{N+k} - S_N|$ its limit. Because of the estimate

$$|S_{N+k} - S_{N+l}| \leqslant |S_{N+k} - S_N| + |S_{N+l} - S_N| \leqslant 2Y_N$$

for $N \geqslant 1$ and $k, l \geqslant 1$, we have

$$\begin{aligned}\{(S_N)_{N \geqslant 1} \text{ is not Cauchy}\} &= \bigcup_{k \geqslant 1} \bigcap_{N \geqslant 1} \bigcup_{k \geqslant 1} \bigcup_{l \geqslant 1} \{|S_{N+k} - S_{N+l}| > 2^{-k}\} \\ &\subset \bigcup_{k \geqslant 1} \bigcap_{N \geqslant 1} \{Y_N > 2^{-k-1}\}.\end{aligned}$$

It is therefore sufficient to prove that

$$\mathbf{P}\left(\bigcap_{N \geqslant 1} \{Y_N > 2^{-k-1}\}\right) = 0$$

for each $k \geqslant 1$, or what amounts to the same thing, to prove that for any $\varepsilon > 0$, we have

$$\lim_{N \to +\infty} \mathbf{P}(Y_N > \varepsilon) = 0$$

(which means that Y_N converges to 0 in probability).

We begin by estimating $\mathbf{P}(Y_{N,M} > \varepsilon)$. *If* $Y_{N,M}$ was defined as $S_{N+M} - S_N$ (without the sup over $k \leqslant M$), this would be easy using the Markov inequality. To handle it, we use Kolmogorov's Maximal Inequality (see Lemma B.10.3): since the $(X_n)_{N+1 \leqslant n \leqslant N+M}$ are independent, this shows that for any $\varepsilon > 0$, we have

$$\mathbf{P}(Y_{N,M} > \varepsilon) = \mathbf{P}\left(\sup_{k \leqslant M} \left| \sum_{1 \leqslant n \leqslant k} X_{N+n} \right| > \varepsilon\right) \leqslant \frac{1}{\varepsilon^2} \sum_{n=N+1}^{N+M} \mathbf{V}(X_n).$$

Letting $M \to +\infty$, we obtain

$$\mathbf{P}(Y_N > \varepsilon) \leqslant \frac{1}{\varepsilon^2} \sum_{n \geqslant N+1} \mathbf{V}(X_n).$$

From the assumption on the convergence of the series of variance, this tends to 0 as $N \to +\infty$, which shows that the partial sums converge almost surely as claimed.

Now let $X = \sum X_n$ be the sum of the series, defined almost surely. For $N \geqslant 1$ and $M \geqslant 1$, we have

$$\|S_{M+N} - S_N\|_{L^2}^2 = \mathbf{E}\left(\left|\sum_{n=N+1}^{N+M} X_n\right|^2\right) = \sum_{n=N+1}^{N+M} \mathbf{E}(|X_n|^2) = \sum_{n=N+1}^{N+M} \mathbf{V}(X_n).$$

The assumption that (B.13) converges therefore implies that $(S_N)_{N\geqslant 1}$ is a Cauchy sequence in L^2, hence converges. Its limit necessarily coincides (almost surely) with the sum X, which shows that X is square-integrable. It follows that it is integrable and that its expectation can be computed as the sum of $\mathbf{E}(X_n)$. □

Remark B.10.2 (1) This result is a special case of Kolmogorov's Three Series Theorem, which gives a necessary and sufficient condition for almost sure convergence of a series of independent complex random variables (X_n); namely, it is enough that for some $c > 0$, and necessary that for all $c > 0$, the series

$$\sum_n \mathbf{P}(|X_n| > c), \qquad \sum_n \mathbf{E}(X_n^c), \qquad \sum_n \mathbf{V}(X_n^c)$$

converge, where $X_n^c = X_n$ if $|X_n| \leqslant c$ and $X_n^c = 0$ otherwise (see, e.g., [9, Th. 22.8] for the full proof, or try to reduce it to the previous case).

(2) It is worth mentioning two further results for context: (1) the event "the series converges" is an asymptotic event, in the sense that it doesn't depend on any finite number of the random variables – Kolmogorov's Zero–One Law then shows that this event can only have probability 0 or 1 – and (2) a theorem of P. Lévy shows that, again for independent summands, the almost sure convergence is equivalent to convergence in law or to convergence in probability. For proofs and discussion of these facts, see, for instance, [83, §0.III].

Here is Kolmogorov's maximal inequality:

Lemma B.10.3 *Let* $M \geqslant 1$ *be an integer,* $Y_1, \ldots, Y_M$ *independent complex random variables in* L^2 *with* $\mathbf{E}(Y_n) = 0$ *for all* n*. Then for any* $\varepsilon > 0$*, we have*

$$\mathbf{P}\left(\sup_{1\leqslant k\leqslant M} |Y_1 + \cdots + Y_k| > \varepsilon\right) \leqslant \frac{1}{\varepsilon^2}\sum_{n=1}^{M} \mathbf{V}(Y_n).$$

Proof Let

$$S_n = Y_1 + \cdots + Y_n$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

for $1 \leqslant n \leqslant \mathrm{M}$. We define a random variable T with values in $[0, +\infty]$ by $\mathrm{T} = \infty$ if $|\mathrm{S}_n| \leqslant \varepsilon$ for all $n \leqslant \mathrm{M}$, and otherwise

$$\mathrm{T} = \inf\{n \leqslant \mathrm{M} \mid |\mathrm{S}_n| > \varepsilon\}.$$

We then have

$$\Big\{ \sup_{1 \leqslant k \leqslant \mathrm{M}} |\mathrm{Y}_1 + \cdots + \mathrm{Y}_k| > \varepsilon \Big\} = \bigcup_{1 \leqslant n \leqslant \mathrm{M}} \{\mathrm{T} = n\},$$

and the union is disjoint. In particular, we get

$$\mathbf{P}\Big(\sup_{1 \leqslant k \leqslant \mathrm{M}} |\mathrm{S}_k| > \varepsilon \Big) = \sum_{n=1}^{\mathrm{M}} \mathbf{P}(\mathrm{T} = n).$$

We now note that $|\mathrm{S}_n|^2 \geqslant \varepsilon^2$ on the event $\{\mathrm{T} = n\}$, so that we can also write

$$\mathbf{P}\Big(\sup_{1 \leqslant k \leqslant \mathrm{M}} |\mathrm{S}_k| > \varepsilon \Big) \leqslant \frac{1}{\varepsilon^2} \sum_{n=1}^{\mathrm{M}} \mathbf{E}(|\mathrm{S}_n|^2 \mathbf{1}_{\{\mathrm{T}=n\}}). \tag{B.14}$$

We claim next that

$$\mathbf{E}(|\mathrm{S}_n|^2 \mathbf{1}_{\{\mathrm{T}=n\}}) \leqslant \mathbf{E}(|\mathrm{S}_\mathrm{M}|^2 \mathbf{1}_{\{\mathrm{T}=n\}}) \tag{B.15}$$

for all $n \leqslant \mathrm{M}$.

Indeed, if we write $\mathrm{S}_\mathrm{M} = \mathrm{S}_n + \mathrm{R}_n$, the independence assumption shows that R_n is independent of $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$ and in particular is independent of the indicator function of the event $\{\mathrm{T} = n\}$, which only depends on $\mathrm{X}_1, \ldots, \mathrm{X}_n$. Moreover, we have $\mathbf{E}(\mathrm{R}_n) = 0$. Now, taking the modulus square in the definition and multiplying by this indicator function, we get

$$|\mathrm{S}_\mathrm{M}|^2 \mathbf{1}_{\{\mathrm{T}=n\}} = |\mathrm{S}_n|^2 \mathbf{1}_{\{\mathrm{T}=n\}} + \mathrm{S}_n \overline{\mathrm{R}}_n \mathbf{1}_{\{\mathrm{T}=n\}} + \overline{\mathrm{S}}_n \mathrm{R}_n \mathbf{1}_{\{\mathrm{T}=n\}} + |\mathrm{R}_n|^2 \mathbf{1}_{\{\mathrm{T}=n\}}.$$

Taking then the expectation, and using the positivity of the last term, this gives

$$\mathbf{E}(|\mathrm{S}_\mathrm{M}|^2 \mathbf{1}_{\{\mathrm{T}=n\}}) \geqslant \mathbf{E}(|\mathrm{S}_n|^2 \mathbf{1}_{\{\mathrm{T}=n\}}) + \mathbf{E}(\mathrm{S}_n \overline{\mathrm{R}}_n \mathbf{1}_{\{\mathrm{T}=n\}}) + \mathbf{E}(\overline{\mathrm{S}}_n \mathrm{R}_n \mathbf{1}_{\{\mathrm{T}=n\}}).$$

But, by independence, we have

$$\mathbf{E}(\mathrm{S}_n \overline{\mathrm{R}_n} \mathbf{1}_{\{\mathrm{T}=n\}}) = \mathbf{E}(\mathrm{S}_n \mathbf{1}_{\{\mathrm{T}=n\}})\, \mathbf{E}(\overline{\mathrm{R}_n}) = 0,$$

and similarly, $\mathbf{E}(\overline{\mathrm{S}_n} \mathrm{R}_n \mathbf{1}_{\{\mathrm{T}=n\}}) = 0$. Thus we get the bound (B.15).

Using this in (B.14), this gives

$$\mathbf{P}\Big(\sup_{1 \leqslant k \leqslant \mathrm{M}} |\mathrm{S}_k| > \varepsilon \Big) \leqslant \frac{1}{\varepsilon^2} \sum_{n=1}^{\mathrm{M}} \mathbf{E}(|\mathrm{S}_\mathrm{M}|^2 \mathbf{1}_{\{\mathrm{T}=n\}}) \leqslant \frac{1}{\varepsilon^2} \mathbf{E}(|\mathrm{S}_\mathrm{M}|^2)$$

by positivity once again. □

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

Exercise B.10.4 Deduce from Kolmogorov's Theorem the nontrivial direction of the Borel–Cantelli Lemma: if $(A_n)_{n\geqslant 1}$ is a sequence of independent events such that

$$\sum_{n\geqslant 1} \mathbf{P}(A_n) = +\infty,$$

then an element of the underlying probability space belongs almost surely to infinitely many of the sets A_n.

The second result we need is more subtle. It concerns similar series, but *without* the independence assumption, which is replaced by an orthogonality condition.

Theorem B.10.5 (Menshov–Rademacher) *Let* (X_n) *be a sequence of complex-valued random variables such that* $\mathbf{E}(X_n) = 0$ *and*

$$\mathbf{E}(X_n\overline{X_m}) = \begin{cases} 0 & \text{if } n \neq m, \\ 1 & \text{if } n = m. \end{cases}$$

Let (a_n) *be any sequence of complex numbers such that*

$$\sum_{n\geqslant 1} |a_n|^2(\log n)^2 < +\infty.$$

Then the series

$$\sum_{n\geqslant 1} a_n X_n$$

converges almost surely, and hence also in law.

Remark B.10.6 Consider the probability space $\Omega = \mathbf{R}/\mathbf{Z}$ with the Lebesgue measure and the random variables $X_n(t) = e(nt)$ for $n \in \mathbf{Z}$. One easily sees (adapting to double-sided sequences and symmetric partial sums) that Theorem B.10.5 implies that the series

$$\sum_{n\in\mathbf{Z}} a_n e(nt)$$

converges almost everywhere (with respect to Lebesgue measure), provided

$$\sum_{n\in\mathbf{Z}} |a_n|^2(\log |n|)^2 < +\infty.$$

This may be proved more directly (see, e.g., [121, III, th. 4.4]) using properties of Fourier series, but it is far from obvious. Note that, in this case, a very famous theorem of Carleson shows that the condition may be replaced with

$\sum |a_n|^2 < +\infty$. On the other hand, Menshov proved that Theorem B.10.5 *cannot* in general be relaxed in this way: for general orthonormal sequences, the term $(\log n)^2$ cannot be replaced by any positive function $f(n)$ such that $f(n) = o((\log n)^2)$, even for $\mathbf{R}/\mathbf{Z}$.

We begin with a lemma that will play an auxiliary role similar to Kolmogorov's inequality.

Lemma B.10.7 *Let* $(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{N})$ *be orthonormal random variables,* $(a_1, \ldots, a_\mathrm{N})$ *be complex numbers, and* $\mathrm{S}_k = a_1\mathrm{X}_1 + \cdots + a_k\mathrm{X}_k$ *for* $1 \leqslant k \leqslant \mathrm{N}$. *We have*

$$\mathbf{E}\Big(\max_{1 \leqslant k \leqslant \mathrm{N}} |\mathrm{S}_k|^2\Big) \ll (\log \mathrm{N})^2 \sum_{n=1}^{\mathrm{N}} |a_n|^2,$$

where the implied constant is absolute.

Proof The basic ingredient is a simple combinatorial property, which we present a bit abstractly. We claim that there exists a family $\mathcal{J}$ of discrete intervals

$$\mathrm{I} = \{n_\mathrm{I}, \ldots, m_\mathrm{I} - 1\}, \qquad m_\mathrm{I} - n_\mathrm{I} \geqslant 1,$$

for $\mathrm{I} \in \mathcal{J}$, with the following two properties:

(1) any interval $1 \leqslant n \leqslant \mathrm{M}$ with $\mathrm{M} \leqslant \mathrm{N}$ is the disjoint union of $\ll \log \mathrm{N}$ intervals $\mathrm{I} \in \mathcal{J}$;

(2) an integer n with $1 \leqslant n \leqslant \mathrm{N}$ belongs to $\ll \log \mathrm{N}$ intervals in $\mathcal{J}$;

and in both cases the implied constant is independent of N.

To see this, let $n \geqslant 1$ be such that $2^{n-1} \leqslant \mathrm{N} \leqslant 2^n$ (so that $n \ll \log \mathrm{N}$), and consider for instance the family $\mathcal{J}$ of dyadic intervals

$$\mathrm{I}_{i,j} = \{n \mid 1 \leqslant n \leqslant \mathrm{N} \text{ and } i2^j \leqslant n < (i+1)2^j\}$$

for $0 \leqslant j \leqslant n$ and $1 \leqslant i \leqslant 2^{n-j}$. (The proofs of both properties in this case are left to the reader.)

Now, having fixed such a collection of intervals, we denote by T the smallest integer between 1 and N such that

$$\max_{1 \leqslant k \leqslant \mathrm{N}} |\mathrm{S}_k| = |\mathrm{S}_\mathrm{T}|.$$

By our first property of the intervals $\mathcal{J}$, we can write

$$\mathrm{S}_\mathrm{T} = \sum_{\mathrm{I}} \tilde{\mathrm{S}}_\mathrm{I},$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

where I runs over a set of $\ll \log N$ disjoint intervals in $\mathcal{J}$ and

$$\tilde{S}_I = \sum_{n \in I} a_n X_n$$

is the corresponding partial sum. By the Cauchy–Schwarz inequality, and the first property again, we get

$$|S_T|^2 \ll (\log N) \sum_{I} |\tilde{S}_I|^2 \ll (\log N) \sum_{I \in \mathcal{J}} |\tilde{S}_I|^2.$$

Taking the expectation and using orthonormality, we derive

$$\begin{aligned} \mathbf{E}\Big(\max_{1 \leqslant k \leqslant N} |S_k|^2\Big) = \mathbf{E}(|S_T|^2) &\ll (\log N) \sum_{I \in \mathcal{J}} \mathbf{E}(|\tilde{S}_I|^2) \\ &= (\log N) \sum_{I \in \mathcal{J}} \sum_{n \in I} |a_n|^2 \ll (\log N)^2 \sum_{1 \leqslant n \leqslant N} |a_n|^2 \end{aligned}$$

by the second property of the intervals $\mathcal{J}$. □

Proof of Theorem B.10.5 If the factor $(\log N)^2$ in Lemma B.10.7 was replaced by $(\log n)^2$ inside the sum, we would proceed just like the deduction of Theorem B.10.1 from Lemma B.10.3. Since this is not the case, a slightly different argument is needed.

We define

$$S_n = a_1 X_1 + \cdots + a_n X_n$$

for $n \geqslant 1$. For $j \geqslant 0$, we also define the dyadic sum

$$\tilde{S}_j = \sum_{2^j \leqslant n < 2^{j+1}} a_n X_n = S_{2^{j+1}-1} - S_{2^j}.$$

We first note that the series

$$T = \sum_{j \geqslant 0} (j+1)^2 |\tilde{S}_j|^2$$

converges almost surely. Indeed, since it is a series of nonnegative terms, it suffices to show that $\mathbf{E}(T) < +\infty$. But we have

$$\begin{aligned} \mathbf{E}(T) &= \sum_{j \geqslant 0} (j+1)^2 \, \mathbf{E}(|\tilde{S}_j|^2) \\ &= \sum_{j \geqslant 0} (j+1)^2 \sum_{2^j \leqslant n < 2^{j+1}} |a_n|^2 \ll \sum_{n \geqslant 1} |a_n|^2 (\log 2n)^2 < +\infty \end{aligned}$$

by orthonormality and by the assumption of the theorem.

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

Next, we observe that for $j \geqslant 0$ and $k \geqslant 0$, we have

$$|S_{2^{j+k}} - S_{2^j}| \leqslant \sum_{i=j}^{j+k-1} |\tilde{S}_i| \leqslant \Big(\sum_{j \leqslant i < j+k} \frac{1}{(i+1)^2} \Big)^{1/2} |T|^{1/2} \ll \Big(\frac{|T|}{j+1} \Big)^{1/2}$$

by the Cauchy–Schwarz inequality. We conclude that the sequence (S_{2^j}) is almost surely a Cauchy sequence and hence converges almost surely to a random variable S.

Finally, to prove that (S_n) converges almost surely to S, we observe that for any $n \geqslant 1$, and $j \geqslant 0$ such that $2^j \leqslant n < 2^{j+1}$, we have

$$|S_n - S_{2^j}| \leqslant M_j = \max_{2^j < k \leqslant 2^{j+1}} \Big| \sum_{m=2^j}^{k} a_n X_n \Big|. \tag{B.16}$$

Lemma B.10.7 implies that

$$\mathbf{E}\Big(\sum_{j \geqslant 0} M_j^2 \Big) = \sum_{j \geqslant 0} \mathbf{E}(M_j^2) \ll \sum_{n \geqslant 1} (\log 2n)^2 |a_n|^2 < +\infty,$$

which means in particular that M_j tends to 0 as $j \to +\infty$ almost surely. From (B.16) and the convergence of $(S_{2^j})_j$ to S, we deduce that (S_n) converges almost surely to S. This finishes the proof. □

We will also use information on the support of the distribution of a random series with independent summands.

Proposition B.10.8 *Let* B *be a separable Banach space. Let* $(X_n)_{n \geqslant 1}$ *be a sequence of* independent B-*valued random variables such that the series* $X = \sum X_n$ *converges almost surely.*[3] *The support of the law of* X *is the closure of the set of all convergent series of the form* $\sum x_n$, *where* x_n *belongs to the support of the law of* X_n *for all* $n \geqslant 1$.

Proof For $N \geqslant 1$, we write

$$S_N = \sum_{n=1}^{N} X_n, \qquad R_N = X - S_N.$$

The variables S_N and R_N are independent.

First, we observe that Lemmas B.2.1 and B.2.2 imply that the support of S_N is the closure of the set of elements $x_1 + \cdots + x_N$ with $x_n \in \operatorname{supp}(X_n)$ for $1 \leqslant n \leqslant N$ (apply Lemma B.2.1 to the law of $(X_1, \ldots, X_N)$ on B^N, which has

[3] Recall that by the result of P. Lévy mentioned in Remark B.10.2, this is in fact equivalent to convergence in law.

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

support the product of the $\mathrm{supp}(X_n)$ by Lemma B.2.2, and to the addition map $B^N \to B$).

We first prove that all convergent series $\sum x_n$ with $x_n \in \mathrm{supp}(X_n)$ belong to the support of X, hence the closure of this set is contained in the support of X, as claimed. Thus let $x = \sum x_n$ be of this type. Let $\varepsilon > 0$ be fixed.

For all N large enough, we have

$$\Big\| \sum_{n>N} x_n \Big\| < \varepsilon,$$

and it follows that $x_1 + \cdots + x_N$, which belongs to the support of S_N as first remarked also belongs to the open ball U_ε of radius ε around x. Hence

$$\mathbf{P}(S_N \in U_\varepsilon) > 0$$

for all N large enough (U_ε is an open neighborhood of some element in the support of S_N).

Now the almost sure convergence implies (by the dominated convergence theorem, for instance) that $\mathbf{P}(\|R_N\| > \varepsilon) \to 0$ as $N \to +\infty$. Therefore, taking N suitably large, we get

$$\begin{aligned}\mathbf{P}(\|X - x\| < 2\varepsilon) &\geqslant \mathbf{P}(\|S_N - x\| < \varepsilon \text{ and } \|R_N\| < \varepsilon)\\ &= \mathbf{P}(\|S_N - x\| < \varepsilon)\,\mathbf{P}(\|R_N\| < \varepsilon) > 0\end{aligned}$$

(by independence). Since ε is arbitrary, this shows that $x \in \mathrm{supp}(X)$, as claimed.

Conversely, let $x \in \mathrm{supp}(X)$. For any $\varepsilon > 0$, we have

$$\mathbf{P}\Big(\Big\| \sum_{n\geqslant 1} X_n - x \Big\| < \varepsilon\Big) > 0.$$

Since, for any $n_0 \geqslant 1$, we have

$$\mathbf{P}\Big(\Big\| \sum_{n\geqslant 1} X_n - x \Big\| < \varepsilon \text{ and } X_{n_0} \notin \mathrm{supp}(X_{n_0})\Big) = 0,$$

this means in fact that

$$\mathbf{P}\Big(\Big\| \sum_{n\geqslant 1} X_n - x \Big\| < \varepsilon \text{ and } X_n \in \mathrm{supp}(X_n) \text{ for all } n\Big) > 0.$$

In particular, we can find $x_n \in \mathrm{supp}(X_n)$ such that the series $\sum x_n$ converges and

$$\Big\| \sum_{n\geqslant 1} x_n - x \Big\| < \varepsilon,$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

and hence x belongs to the closure of the set of convergent series $\sum x_n$ with x_n in the support of X_n for all n. ☐

B.11 Some Probability in Banach Spaces

We consider in this section some facts about probability in a (complex) Banach space V. Most are relatively elementary. For simplicity, we will always assume that V is separable (so that, in particular, Borel measures on V have a well-defined support).

The first result concerns series

$$\sum_n X_n$$

where (X_n) is a sequence of *symmetric* random variables, which means that for any $N \geqslant 1$, and for any choice $(\varepsilon_1, \ldots, \varepsilon_N)$ of signs $\varepsilon_n \in \{-1, 1\}$ for $1 \leqslant n \leqslant N$, the random vectors

$$(X_1, \ldots, X_N) \text{ and } (\varepsilon_1 X_1, \ldots, \varepsilon_N X_N)$$

have the same distribution.

Symmetric random variables have remarkable properties. For instance, we have:

Proposition B.11.1 (Lévy) *Let* V *be a separable Banach space with norm* $\|\cdot\|$*, and* (X_n) *a sequence of* V*-valued random variables. Assume that the sequence* (X_n) *is symmetric. Let*

$$S_N = X_1 + \cdots + X_N$$

for $N \geqslant 1$. *For* $N \geqslant 1$ *and* $\varepsilon > 0$, *we have*

$$\mathbf{P}(\max_{1\leqslant n\leqslant N} \|S_N\| > \varepsilon) \leqslant 2\,\mathbf{P}(\|S_N\| > \varepsilon).$$

This result is known as *Lévy's reflection principle* and can be compared with Kolmogorov's maximal inequality (Lemma B.10.3).

Proof (1) Similarly to the proof of Lemma B.10.3, we define a random variable T by $T = \infty$ if $\|S_n\| \leqslant \varepsilon$ for all $n \leqslant N$, and otherwise

$$T = \inf\{n \leqslant N \mid \|S_n\| > \varepsilon\}.$$

Assume $T = k$ and consider the random variables

$$X'_n = X_n \text{ for } 1 \leqslant n \leqslant k, \quad X'_n = -X_n \text{ for } k+1 \leqslant n \leqslant N.$$

By assumption, the sequence $(X'_n)_{1\leqslant n\leqslant N}$ has the same distribution as $(X_n)_{1\leqslant n\leqslant N}$. Let S'_n denote the partial sums of the sequence (X'_n), and T' the analogue of T for the sequence (X'_n). The event $\{T'=k\}$ is the same as $\{T=k\}$ since $X'_n = X_n$ for $n \leqslant k$. On the other hand, we have

$$S'_N = X_1 + \cdots + X_k - X_{k+1} - \cdots - X_N = 2S_k - S_N.$$

Therefore

$$\mathbf{P}(\|S_N\|>\varepsilon \text{ and } T=k)=\mathbf{P}(\|S'_N\|>\varepsilon \text{ and } T'=k)=\mathbf{P}(\|2S_k-S_N\|>\varepsilon \text{ and } T=k).$$

By the triangle inequality we have

$$\{T=k\} \subset \{\|S_N\| > \varepsilon \text{ and } T=k\} \cup \{\|2S_k - S_N\| > \varepsilon \text{ and } T=k\}.$$

We deduce

$$\begin{aligned}
\mathbf{P}(\max_{1\leqslant n\leqslant N} \|S_n\| > \varepsilon) &= \sum_{k=1}^{N} \mathbf{P}(T=k) \\
&\leqslant \sum_{k=1}^{N} \mathbf{P}(\|S_N\| > \varepsilon \text{ and } T=k) \\
&\quad + \sum_{k=1}^{N} \mathbf{P}(\|2S_n - S_K\| > \varepsilon \text{ and } T=k) \\
&= 2\,\mathbf{P}(\|S_N\| > \varepsilon).
\end{aligned}$$

□

We now consider the special case where the Banach space V is $C([0,1])$, the space of complex-valued continuous functions on $[0,1]$ with the norm

$$\|f\|_\infty = \sup_{t\in[0,1]} |f(t)|.$$

For a $C([0,1])$-valued random variable X and any fixed $t \in [0,1]$, we will denote by $X(t)$ the complex-valued random variable that is the value of the random function X at t, that is, $X(t) = e_t \circ X$, where $e_t : C([0,1]) \longrightarrow \mathbf{C}$ is the evaluation at t.

Definition B.11.2 (Convergence of finite distributions) Let (X_n) be a sequence of $C([0,1])$-valued random variables, and let X be a $C([0,1])$-valued random variable. One says that (X_n) converges to X *in the sense of finite distributions* if and only if, for all integers $k \geqslant 1$, and for all

$$0 \leqslant t_1 < \cdots < t_k \leqslant 1,$$

the random vectors $(X_n(t_1), \ldots, X_n(t_k))$ converge in law to $(X(t_1), \ldots, X(t_k))$, in the sense of convergence in law in $\mathbf{C}^k$.

One sufficient condition for convergence of finite distributions is the following:

Lemma B.11.3 *Let* (X_n) *be a sequence of* $C([0,1])$*-valued random variables, and let* X *be a* $C([0,1])$*-valued random variable, all defined on the same probability space. Assume that, for any* $t \in [0,1]$*, the random variables* $(X_n(t))$ *converge in* L^1 *to* $X(t)$*. Then* (X_n) *converges to* X *in the sense of finite distributions.*

Proof Fix $k \geqslant 1$ and

$$0 \leqslant t_1 < \cdots < t_k \leqslant 1.$$

Let φ be a Lipschitz function on $\mathbf{C}^k$ (given the distance associated to the norm

$$\|(z_1, \ldots, z_k)\| = \sum_i |z_i|,$$

for instance) with Lipschitz constant $C \geqslant 0$. Then we have

$$\Big| \mathbf{E}(\varphi(X_n(t_1), \ldots, X_n(t_k))) - \mathbf{E}(\varphi(X(t_1), \ldots, X(t_k)))\Big| \leqslant C \sum_{i=1}^{k} \mathbf{E}(|X_n(t_i) - X(t_i)|),$$

which tends to 0 as $n \to +\infty$ by our assumption. Hence Proposition B.4.1 shows that $(X_n(t_1), \ldots, X_n(t_k))$ converges in law to $(X(t_1), \ldots, X(t_k))$. This proves the lemma. □

Convergence in finite distributions is a necessary condition for convergence in law of (X_n) to X, but it is not sufficient: a simple example (see [10, Example 2.5]) consists in taking the random variable X_n to be the *constant* random variable equal to the function f_n that is piecewise linear on $[0, 1/n]$, $[1/n, 1/(2n)]$ and $[1/(2n), 1]$, and such that $0 \mapsto 0$, $1/n \mapsto 1$, $1/(2n) \mapsto 0$ and $1 \mapsto 0$. Then it is elementary that X_n converges to the constant zero random variable in the sense of finite distributions, but that X_n does not converge in law to 0 (because f_n does not converge uniformly to 0).

Nevertheless, under the additional condition of *tightness* of the sequence of random variables (see Definition B.3.6), the convergence of finite distributions implies convergence in law.

Theorem B.11.4 *Let* (X_n) *be a sequence of* $C([0,1])$*-valued random variables, and let* X *be a* $C([0,1])$*-valued random variable. Suppose that* (X_n) *converges to* X *in the sense of finite distributions. Then* (X_n) *converges in law* *to* X *in the sense of* C([0, 1])*-valued random variables if and only if* (X_n) *is tight.*

For a proof, see, for example, [10, Th. 7.1]. The key ingredient is Prokhorov's Theorem (see [10, Th. 5.1]), which states that a tight family of random variables is relatively compact in the space $\mathcal{P}$ of probability measures on C([0, 1]), given the topology of convergence in law. To see how this implies the result, we note that convergence in the sense of finite distributions of a sequence implies at least that it has *at most* one limit in $\mathcal{P}$ (because probability measures on C([0, 1]) are uniquely determined by the family of their finite distributions; see [10, Ex. 1.3]). Suppose now that there exists a continuous bounded function f on C([0, 1]) such that

$$\mathbf{E}(f(X_n))$$

does not converge to $\mathbf{E}(f(X))$. Then there exists $\delta > 0$ and some subsequence (X_{n_k}) that satisfies $|\mathbf{E}(f(X_{n_k}) - f(X))| \geqslant \delta$ for all k. This subsequence also converges to X in the sense of finite distributions and by relative compactness admits a further subsequence that converges in law; but the limit of that further subsequence must then be X, which contradicts the inequality above.

Remark B.11.5 For certain purposes, it is important to observe that this proof of convergence in law is indirect and does not give quantitative estimates.

We will also use a variant of this result involving Fourier series. A minor issue is that we wish to consider functions f on [0, 1] that are not necessarily periodic, in the sense that $f(0)$ might differ from $f(1)$. However, we will have $f(0) = 0$. To account for this, we use the identity function in addition to the periodic exponentials to represent continuous functions with $f(0) = 0$.

We denote by $C_0([0, 1])$ the subspace of C([0, 1]) of functions vanishing at 0. We denote by e_0 the function $e_0(t) = t$, and for $h \neq 0$, we put $e_h(t) = e(ht)$. We denote further by $C_0(\mathbf{Z})$ the Banach space of complex-valued functions on $\mathbf{Z}$ converging to 0 at infinity with the sup norm. We define a continuous linear map FT: $C([0, 1]) \to C_0(\mathbf{Z})$ by mapping f to the sequence $(\widetilde{f}(h))_{h \in \mathbf{Z}}$ of its Fourier coefficients, where $\widetilde{f}(0) = f(1)$ and for $h \neq 0$ we have

$$\widetilde{f}(h) = \int_0^1 (f(t) - tf(1))e(-ht)dt = \int_0^1 (f - f(1)e_0)e_{-h}.$$

We want to relate convergence in law in $C_0([0, 1])$ with convergence, in law or in the sense of finite distributions, of these "Fourier coefficients" in $C_0(\mathbf{Z})$. Here convergence of finite distributions of a sequence (X_n) of $C_0(\mathbf{Z})$-valued random variables to X means that for any $\mathrm{H} \geqslant 1$, the vectors $(\mathrm{X}_{n,h})_{|h|\leqslant \mathrm{H}}$ converge in law to $(\mathrm{X}_h)_{|h|\leqslant \mathrm{H}}$, in the sense of convergence in law in $\mathbf{C}^{2\mathrm{H}+1}$.

First, since FT is continuous, Proposition B.3.2 gives immediately

Lemma B.11.6 *If* $(\mathrm{X}_n)_n$ *is a sequence of* $\mathrm{C}_0([0,1])$*-valued random variables that converges in law to a random variable* X*, then* $\mathrm{FT}(\mathrm{X}_n)$ *converges in law to* FT(X).

Next, we check that the Fourier coefficients determine the law of a $\mathrm{C}_0([0,1])$-valued random variable (this is the analogue of [10, Ex. 1.3]).

Lemma B.11.7 *If* X *and* Y *are* $\mathrm{C}_0([0,1])$*-valued random variables and if* FT(X) *and* FT(Y) *have the same finite distributions, then* X *and* Y *have the same law.*

Proof For $f \in \mathrm{C}_0([0,1])$, the function $g(t) = f(t) - tf(1)$ extends to a 1-periodic continuous function on $\mathbf{R}$. By Féjer's Theorem on the uniform convergence of Cesàro means of Fourier series of continuous periodic functions (see, e.g, [121, III, Th. 3.4]), we have

$$f(t) - tf(1) = \lim_{\mathrm{H}\to+\infty} \sum_{|h|\leqslant \mathrm{H}} \left(1 - \frac{|h|}{\mathrm{H}}\right) \widetilde{f}(h)e(ht)$$

uniformly for $t \in [0,1]$. Evaluating at $t = 0$, where the left-hand side vanishes, we deduce that

$$f = \lim_{\mathrm{H}\to+\infty} \mathrm{C}_\mathrm{H}(f),$$

where

$$\mathrm{C}_\mathrm{H}(f) = f(1)e_0 + \sum_{\substack{|h|\leqslant \mathrm{H}\\ h\neq 0}} \left(1 - \frac{|h|}{\mathrm{H}}\right) \widetilde{f}(h)(e_h - 1).$$

Note that $\mathrm{C}_\mathrm{H}(f) \in \mathrm{C}_0([0,1])$.

We now claim that $\mathrm{C}_\mathrm{H}(\mathrm{X})$ converges to X as $\mathrm{C}_0([0,1])$-valued random variables. Indeed, let φ be a bounded continuous function on $\mathrm{C}_0([0,1])$, say, with $|\varphi| \leqslant \mathrm{M}$. By the above, we have $\varphi(\mathrm{C}_\mathrm{H}(\mathrm{X})) \to \varphi(\mathrm{X})$ as $\mathrm{H} \to +\infty$ pointwise on $\mathrm{C}_0([0,1])$. Since $|\varphi(\mathrm{C}_\mathrm{H}(\mathrm{X}))| \leqslant \mathrm{M}$, which is integrable on the underlying probability space, Lebesgue's dominated convergence theorem implies that $\mathbf{E}(\varphi(\mathrm{C}_\mathrm{H}(\mathrm{X}))) \to \mathbf{E}(\varphi(\mathrm{X}))$. This proves the claim.

In view of the definition of $\mathrm{C}_\mathrm{H}(f)$, which only involves finitely many Fourier coefficients, the equality of finite distributions of FT(X) and FT(Y) implies by composition that for any $\mathrm{H} \geqslant 1$, the $\mathrm{C}_0([0,1])$-valued random variables $\mathrm{C}_\mathrm{H}(\mathrm{X})$ and $\mathrm{C}_\mathrm{H}(\mathrm{Y})$ have the same law. Since we have seen that $\mathrm{C}_\mathrm{H}(\mathrm{X})$

converges in law to X and that $C_H(Y)$ converges in law to Y, it follows that X and Y have the same law. □

Now comes the convergence criterion:

Proposition B.11.8 *Let* (X_n) *be a sequence of* $C_0([0,1])$*-valued random variables, and let* X *be a* $C_0([0,1])$*-valued random variable. Suppose that* $FT(X_n)$ *converges to* $FT(X)$ *in the sense of finite distributions. Then* (X_n) *converges in law to* X *in the sense of* $C_0([0,1])$*-valued random variables if and only if* (X_n) *is tight.*

Proof It is an elementary general fact that if (X_n) converges in law to X, then the family (X_n) is tight. We prove the converse assertion. It suffices to prove that any subsequence of (X_n) has a further subsequence that converges in law to X (see [10, Th. 2.6]). Because (X_n) is tight, so are any of its subsequences. By Prokhorov's Theorem ([10, Th. 5.1]), such a subsequence therefore contains a further subsequence, say, $(X_{n_k})_{k\geqslant 1}$, that converges in law to some probability measure Y. By Lemma B.11.6, the sequence of Fourier coefficients $FT(X_{n_k})$ converges in law to FT(Y). On the other hand, this sequence converges to FT(X) in the sense of finite distributions, by assumption. Hence FT(X) and FT(Y) have the same finite distributions, which implies that X and Y have the same law by Lemma B.11.7. □

Remark B.11.9 The example that was already mentioned before Theorem B.11.4 (namely, [10, Ex. 2.5]) also shows that the convergence of $FT(X_n)$ to FT(X) in the sense of finite distributions is not sufficient to conclude that (X_n) converges in law to X. Indeed, the sequence (X_n) in that example does not converge in law in C([0, 1]), but for $n \geqslant 1$, the (constant) random variable X_n satisfies $X_n(1) = 0$, and by direct computation, the Fourier coefficients (are constant and) satisfy also $|\widetilde{X}_n(h)| \leqslant n^{-1}$ for all $h \neq 0$, which implies that $FT(X_n)$ converges in law to the constant random variable equal to $0 \in C_0(\mathbf{Z})$.

In applications, we need some criteria to detect tightness. One such criterion is due to Kolmogorov:

Proposition B.11.10 (Kolmogorov's tightness criterion) *Let* (X_n) *be a sequence of* $C([0,1])$*-valued random variables. If there exist real numbers* $\alpha > 0$, $\delta > 0$, *and* $C \geqslant 0$ *such that, for any real numbers* $0 \leqslant s < t \leqslant 1$ *and any* $n \geqslant 1$, *we have*

$$\mathbf{E}(|X_n(t) - X_n(s)|^{\alpha}) \leqslant C|t - s|^{1+\delta}, \qquad \text{(B.17)}$$

then (X_n) *is tight.*

See, for instance, [81, Th. I.7] for a proof. The statement does not hold if the exponent $1+\delta$ is replaced by 1.

In fact, for some applications (as in [79]), one needs a variant where the single bound (B.17) is replaced by different ones depending on the size of $|t-s|$ relative to n. Such a result does not seem to follow formally from Proposition B.11.10, because the left-hand side in the inequality is not monotonic in terms of α (in contrast with the right-hand side, which is monotonic since $|t-s| \leqslant 1$). We state a result of this type and sketch the proof for completeness.

Proposition B.11.11 (Kolmogorov's tightness criterion, 2) *Let* (X_n) *be a sequence of* $\mathrm{C}([0,1])$*-valued random variables. Suppose that there exist positive real numbers*

$$\alpha_1, \alpha_2, \alpha_3, \quad \beta_2 < \beta_1, \quad \delta, \quad \mathrm{C}$$

such that for any real numbers $0 \leqslant s < t \leqslant 1$ *and any* $n \geqslant 1$*, we have*

$$\mathbf{E}(|\mathrm{X}_n(t) - \mathrm{X}_n(s)|^{\alpha_1}) \leqslant \mathrm{C}|t-s|^{1+\delta} \quad \text{if } 0 \leqslant |t-s| \leqslant n^{-\beta_1}, \tag{B.18}$$

$$\mathbf{E}(|\mathrm{X}_n(t) - \mathrm{X}_n(s)|^{\alpha_2}) \leqslant \mathrm{C}|t-s|^{1+\delta} \quad \text{if } n^{-\beta_1} \leqslant |t-s| \leqslant n^{-\beta_2}, \tag{B.19}$$

$$\mathbf{E}(|\mathrm{X}_n(t) - \mathrm{X}_n(s)|^{\alpha_3}) \leqslant \mathrm{C}|t-s|^{1+\delta} \quad \text{if } n^{-\beta_2} \leqslant |t-s| \leqslant 1. \tag{B.20}$$

Then (X_n) *is tight.*

Sketch of proof For $n \geqslant 1$, let $\mathrm{D}_n \subset [0,1]$ be the set of dyadic rational numbers with denominator 2^n. For $\delta > 0$, let

$$\omega(\mathrm{X}_n, \delta) = \sup\{|\mathrm{X}_n(t) - \mathrm{X}_n(s)| \mid s,t \in [0,1],\ |t-s| \leqslant \delta\}$$

denote the modulus of continuity of X_n, and for $n \geqslant 1$ and $k \geqslant 0$, let

$$\xi_{n,k} = \sup\{|\mathrm{X}_n(t) - \mathrm{X}_n(s)| \mid s,t \in \mathrm{D}_k,\ |s-t| = 2^{-k}\}.$$

We observe that for any $\alpha > 0$, we have

$$\mathbf{E}(\xi_{n,k}^{\alpha}) \leqslant \sum_{\substack{s,t \in \mathrm{D}_k \\ |t-s| = 2^{-k}}} \mathbf{E}(|\mathrm{X}_n(t) - \mathrm{X}_n(s)|^{\alpha}).$$

As in [60, p. 269], the key step is to prove that

$$\lim_{m \to +\infty} \limsup_{n \to +\infty} \mathbf{P}(\omega(\mathrm{X}_n, 2^{-m}) > \eta) = 0$$

for any $\eta > 0$ (the conclusion is then derived from this fact combined with the Ascoli–Arzela Theorem characterizing compact subsets of $\mathrm{C}([0,1])$). It is convenient here to use the notation $\min(a,b) = a \wedge b$. For fixed m and n, we then write

$$\begin{aligned}\mathbf{P}(\omega(\mathrm{X}_n, 2^{-m}) > \eta) \leqslant \mathbf{P}(&\sup_{|t-s|\leqslant 2^{-m}\wedge n^{-\beta_1}} |\mathrm{X}_n(t) - \mathrm{X}_n(s)| > \eta)\\ &+ \mathbf{P}(\sup_{2^{-m}\wedge n^{-\beta_1} < |t-s| \leqslant 2^{-m}\wedge n^{-\beta_2}} |\mathrm{X}_n(t) - \mathrm{X}_n(s)| > \eta)\\ &+ \mathbf{P}(\sup_{2^{-m}\wedge n^{-\beta_2} < |t-s| \leqslant 2^{-m}} |\mathrm{X}_n(t) - \mathrm{X}_n(s)| > \eta)\end{aligned}$$

(because, for a continuous function f on $[0,1]$, if $|f(t) - f(s)| > \eta$ for some (s,t) such that $|t-s| \leqslant 2^{-m}$, then there exist some dyadic rational numbers (s', t'), necessarily with denominator 2^n with $n \geqslant m$, such that $|f(t') - f(s')| > \eta$). Using (B.18), the first term is

$$\leqslant \sum_{\substack{k\geqslant m\\ 2^{-k}\leqslant n^{-\beta_1}}} \frac{\mathbf{E}(\xi_{n,k}^{\alpha_1})}{\eta^{\alpha_1}} \leqslant \frac{\mathrm{C}}{\eta^{\alpha_1}} \sum_{k\geqslant m} 2^k 2^{-k(1+\delta)} \leqslant \frac{\mathrm{C}}{\eta^{\alpha_1}} \frac{1}{1-2^{-\delta}},$$

and similarly, using (B.19), the second (resp. using (B.20), the third) is

$$\leqslant \sum_{\substack{k\geqslant m\\ \beta_2\log_2(n)\leqslant k\leqslant \beta_1\log_2(n)}} \frac{\mathbf{E}(\xi_{n,k}^{\alpha_2})}{\eta^{\alpha_2}} \leqslant \frac{\mathrm{C}}{\eta^{\alpha_2}} \sum_{k\geqslant m} 2^k 2^{-k(1+\delta)} \leqslant \frac{\mathrm{C}}{\eta^{\alpha_2}} \frac{1}{1-2^{-\delta}}$$

(resp. $\leqslant \mathrm{C}\eta^{-\alpha_3}(1-2^{-\delta})^{-1}$). The result follows. □

We will also use the following inequality of Talagrand, which gives a type of sub-Gaussian behavior of sums of random variable in Banach spaces, extending standard properties of real or complex-valued random variables.

Theorem B.11.12 (Talagrand) *Let* V *be a separable real Banach space and* V' *its dual. Let* $(\mathrm{X}_n)_{n\geqslant 1}$ *be a sequence of independent real-valued random variables with* $|\mathrm{X}_n| \leqslant 1$ *almost surely, and let* $(v_n)_{n\geqslant 1}$ *be a sequence of elements of* V. *Assume that the series* $\sum \mathrm{X}_n v_n$ *converges almost surely in* V. *Let* $m \geqslant 0$ *be a median of*

$$\Big\| \sum_{n\geqslant 1} \mathrm{X}_n v_n \Big\|.$$

Let $\sigma \geqslant 0$ *be the real number such that*

$$\sigma^2 = \sup_{\substack{\lambda\in \mathrm{V}'\\ \|\lambda\|\leqslant 1}} \sum_{n\geqslant 1} |\lambda(v_n)|^2.$$

For any real number $t > 0$*, we have*

$$\mathbf{P}\Big(\Big\|\sum_{n\geqslant 1} X_n v_n\Big\| \geqslant t\sigma + m\Big) \leqslant 4\exp\Big(-\frac{t^2}{16}\Big).$$

We recall that a median m of a real-valued random variable X is any real number such that

$$\mathbf{P}(X \geqslant m) \geqslant \frac{1}{2} \quad \text{and} \quad \mathbf{P}(X \leqslant m) \geqslant \frac{1}{2}.$$

A median always exists. If X is integrable, then Chebychev's inequality

$$\mathbf{P}(X \geqslant t) \leqslant \frac{\mathbf{E}(|X|)}{t} \tag{B.21}$$

shows that $m \leqslant 2\,\mathbf{E}(|X|)$.

Proof This follows easily from [114, Th. 13.2], which concerns finite sums, by passing to the limit. □

The application of this inequality will be the following, which is (in the case $V = \mathbf{R}$) partly a simple variant of a result of Montgomery-Smith [88].

Proposition B.11.13 *Let* V *be a separable real or complex Banach space and* V' *its dual. Let* $(X_n)_{n\geqslant 1}$ *be a sequence of independent random variables with* $|X_n| \leqslant 1$ *almost surely, which are either real- or complex-valued depending on the base field. Let* $(v_n)_{n\geqslant 1}$ *be a sequence of elements of* V*. Assume that the series* $\sum X_n v_n$ *converges almost surely in* V*, and let* X *be its sum.*

(1) *Assume that*

$$\sum_{n\leqslant N} \|v_n\| \ll \log(N), \qquad \sum_{n>N} \|v_n\|^2 \ll \frac{1}{N} \tag{B.22}$$

for all $N \geqslant 1$*. There exists a constant* $c > 0$ *such that for any* $A > 0$*, we have*

$$\mathbf{P}(\|X\| > A) \leqslant c\exp(-\exp(c^{-1}A)).$$

(2) *Assume that* V *is a real Banach space, that* (X_n) *is symmetric, identically distributed, and real-valued, and that there exists* $\lambda \in V'$ *of norm* 1 *such that*

$$\sum_{n\leqslant N} |\lambda(v_n)| \gg \log(N) \tag{B.23}$$

for $N \geqslant 1$*. Then there exists a constant* $c' > 0$ *such that for any* $A > 0$*, we have*

$$c^{-1}\exp(-\exp(cA)) \leqslant \mathbf{P}\left(|\lambda(X)| > A\right) \leqslant \mathbf{P}\left(\|X\| > A\right).$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

Proof We begin with (1), and we first check that we may assume that V is a real Banach space and that the random variables X_n are real-valued. To see this, if V is a complex Banach space, we view it as a real Banach space $V_{\mathbf{R}}$ (by restricting scalar multiplication), and we write $X_n = Y_n + iZ_n$ where Y_n and Z_n are real-valued random variables. Then $X = Y + iZ$ where

$$Y = \sum_{n \geqslant 1} Y_n v_n \quad \text{and} \quad Z = \sum_{n \geqslant 1} Z_n v_n$$

are both almost surely convergent series in $V_{\mathbf{R}}$ with independent real coefficients of absolute value $\leqslant 1$. We then have

$$\mathbf{P}(\|X\| > A) \leqslant \mathbf{P}\left(\|Y\| > \tfrac{1}{2}A \text{ or } \|Z\| > \tfrac{1}{2}A\right) \leqslant \mathbf{P}\left(\|Y\| > \tfrac{1}{2}A\right) + \mathbf{P}\left(\|Z\| > \tfrac{1}{2}A\right)$$

for any $A > 0$, by the triangle inequality. Since the assumptions (B.22) hold independently of whether V is viewed as a real or complex Banach space, we deduce that if (1) holds in the real case, then it also does for complex coefficients.

We now assume that V is a real Banach space. The idea is that if V was simply equal to **R**, then the series X would be a sub-Gaussian random variable, and standard estimates would give a sub-Gaussian upper bound for $\mathbf{P}(|X| > A)$, of the type $\exp(-cA^2)$. Such a bound would be essentially sharp for a *Gaussian* series. But although this is already quite strong, it is far from the truth here; intuitively, this is because, in the Gaussian case, the lower bound for the probability arises from the small but non-zero probability that a single summand (distributed like a Gaussian) might be very large. This cannot happen for the series X, because each X_n is absolutely bounded.

The actual proof "interpolates" between the sub-Gaussian behavior (given by Talagrand's inequality, when the Banach space is infinite-dimensional) and the boundedness of the coefficients (X_n) of the first few steps. This principle goes back (at least) to Montgomery-Smith [88] and has relations with the theory of interpolation of Banach spaces.

Fix an auxiliary parameter $s \geqslant 1$. We write $X = X^\sharp + X^\flat$, where

$$X^\sharp = \sum_{1 \leqslant n \leqslant s^2} X_n v_n \quad \text{and} \quad X^\flat = \sum_{n > s^2} X_n v_n.$$

Let m be a median of the real random variable $\|X^\flat\|$. Then, for any $\alpha > 0$ and $\beta > 0$, we have

$$\mathbf{P}(\|X\| \geqslant \alpha + \beta + m) \leqslant \mathbf{P}(\|X^\sharp\| \geqslant \alpha) + \mathbf{P}(\|X^\flat\| \geqslant m + \beta),$$

by the triangle inequality. We pick

$$\alpha = 8 \sum_{1 \leqslant n \leqslant s^2} \|v_n\|$$

so that by the assumption $|X_n| \leqslant 1$, we have

$$\mathbf{P}(\|X^\sharp\| \geqslant \alpha) = 0.$$

Then we take $\beta = s\sigma$, where $\sigma \geqslant 0$ is such that

$$\sigma^2 = \sup_{\|\lambda\| \leqslant 1} \sum_{n > s^2} |\lambda(v_n)|^2,$$

where λ runs over the elements of norm $\leqslant 1$ of the dual space V'. By Talagrand's Inequality (Theorem B.11.12), we have

$$\mathbf{P}(\|X^\flat\| \geqslant m + \beta) \leqslant 4 \exp\left(-\frac{s^2}{8}\right).$$

Hence, for all $s \geqslant 1$, we have

$$\mathbf{P}(\|X\| \geqslant \alpha + \beta + m) \leqslant 4 \exp\left(-\frac{s^2}{8}\right).$$

We now select s as large as possible so that $m + \alpha + \beta \leqslant A$. By Chebychev's inequality (B.21), we have

$$m \leqslant 2\,\mathbf{E}(\|X^\flat\|) \leqslant 2 \sum_{1 \leqslant h \leqslant s^2} \|v_n\|$$

so that

$$m + \alpha \ll \sum_{1 \leqslant n \leqslant s^2} \|v_n\| \ll \log(2s) \tag{B.24}$$

for any $s \geqslant 1$ by (B.22). Moreover, for any linear form λ with $\|\lambda\| \leqslant 1$, we have

$$\sum_{n > s^2} |\lambda(v_n)|^2 \ll \sum_{n > s^2} \|v_n\|^2 \ll \frac{1}{s^2}$$

so that $\sigma \ll s^{-1}$ and $\beta = s\sigma \ll 1$. It follows that

$$m + \alpha + \beta \leqslant c \log(cs) \tag{B.25}$$

for some constant $c \geqslant 1$ and all $s \geqslant 1$. We finally select s so that $c \log(cs) = A$, that is,

$$s = \frac{1}{c} \exp\left(\frac{A}{c}\right)$$

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press

(assuming, as we may, that A is large enough so that $s \geqslant 1$) and deduce that

$$\mathbf{P}(\|X\| \geqslant A) \leqslant 4\exp\Big(-\frac{s^2}{8}\Big) = 4\exp\Big(-\frac{1}{8c^2}\exp\Big(\frac{A}{c}\Big)\Big).$$

This gives the desired upper bound.

We now prove (2). Replacing the vectors v_n by the real numbers $\lambda(v_n)$ (recall that (2) is a statement for real Banach spaces and random variables), we may assume that $V = \mathbf{R}$. Let $\alpha > 0$ be such that

$$\sum_{n \leqslant N} |\lambda(v_n)| \geqslant \alpha \log(N)$$

for $N \geqslant 1$, and let β_n be a median of $|X_n|$. We then derive

$$\mathbf{P}(|X| > A) \geqslant \mathbf{P}\Big(X_n \geqslant \beta_n \text{ for } 1 \leqslant n \leqslant e^{(\alpha\beta)^{-1}A} \text{ and } \sum_{n > e^{A/(\alpha\beta)}} v_n X_n \geqslant 0\Big).$$

Since the random variables (X_n) are independent, this leads to

$$\mathbf{P}(|X| > A) \geqslant \Big(\frac{1}{4}\Big)^{\lfloor \exp(A/(\alpha\beta)) \rfloor} \mathbf{P}\Big(\sum_{n > e^{A/(\alpha\beta)}} v_n X_n \geqslant 0\Big).$$

Furthermore, since each X_n is symmetric, so is the sum

$$\sum_{n > e^{A/(\alpha\beta)}} v_n X_n,$$

which means that it has probability $\geqslant 1/2$ to be $\geqslant 0$. Therefore we have

$$\mathbf{P}(|X| > A) \geqslant \frac{1}{8} e^{-(\log 4)\exp(A/(\alpha\beta))}.$$

This is of the right form asymptotically, and thus the proof is completed. □

Remark B.11.14 (1) The typical example where the proposition applies is when $\|v_n\|$ is comparable to $1/n$.

(2) Many variations along these lines are possible. For instance, in Chapter 3, we encounter the situation where the vector v_n is zero unless $n = p$ is a prime p, in which case

$$\|v_p\| = \frac{1}{p^\sigma}$$

for some real number σ such that $1/2 < \sigma < 1$. In that case, we have

$$\sum_{n \leqslant N} \|v_n\| \gg \frac{N^{1-\sigma}}{\log N}, \qquad \sum_{n > N} \|v_n\|^2 \ll \frac{1}{N^{2\sigma-1}} \frac{1}{\log N}$$

for $N \geqslant 2$ (by the Prime Number Theorem) instead of (B.22), and the adaptation of the arguments in the proof of the proposition lead to

$$\mathbf{P}(\|X\| > A) \leqslant c \exp\left(-cA^{1/(1-\sigma)}(\log A)^{1/(2(1-\sigma))}\right).$$

(Indeed, check that (B.25) gives here

$$m + \alpha + \beta \ll \frac{s^{2(1-\sigma)}}{\sqrt{\log s}},$$

and we take

$$s = A^{1/(2(1-\sigma))}(\log A)^{1/(4(1-\sigma))}$$

in the final application of Talagrand's inequality.)

On the other hand, in Chapter 5, we have a case where (up to re-indexing), the assumptions (B.22) and (B.23) are replaced by

$$\sum_{n \leqslant N} \|v_n\| \gg (\log N)^2 \quad \text{and} \quad \sum_{n > N} \|v_n\|^2 \ll \frac{\log N}{N}.$$

Then we obtain by the same argument the estimates

$$\mathbf{P}(\|X\| > A) \leqslant c \exp(-\exp(c^{-1}A^{1/2})),$$
$$c^{-1} \exp(-\exp(cA^{1/2})) \leqslant \mathbf{P}(|\lambda(X)| > A) \leqslant \mathbf{P}(\|X\| > A)$$

for some real number $c > 0$.

https://doi.org/10.1017/9781108888226.010 Published online by Cambridge University Press