B Mathematical Appendices

In this appendix, the reader can find a number of basic definitions and results concerning some of the mathematical tools that are used throughout the book. Given their wide range, it is not possible to discuss these tools in a self-contained manner in this appendix. Nevertheless, we believe that gathering a coherent set of notions and notations could be useful to the reader.

Although most of the proofs can be found in the literature (we provide references for most of them), often in a much more general form, we have occasionally provided explicit elementary derivations tailored for the particular use made in the book. The results are not always stated in their most general form, in order to avoid introducing too many concepts and notations.

Since the elementary notions borrowed from topology are used only in the case of metric spaces and are always presented and developed from scratch, they are not exposed in a systematic way.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

B.1 Real analysis

B.1.1 Elementary Inequalities

Lemma B.1 (Comparing arithmetic and geometric means)**.** *For any collection $x_1, \dots, x_n$ of nonnegative real numbers,*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \Big\{\prod_{i=1}^{n} x_i\Big\}^{1/n}, \tag{B.1}$$

with equality if and only if $x_1 = x_2 = \cdots = x_n$.

Lemma B.2 (Hölder's inequality, finite form)**.** *For all $(x_1, \dots, x_n), (y_1, \dots, y_n) \in \mathbb{R}^n$ and all $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\sum_{k=1}^{n} |x_k y_k| \leq \Big(\sum_{k=1}^{n} |x_k|^p\Big)^{1/p} \Big(\sum_{k=1}^{n} |x_k|^q\Big)^{1/q}.$$

Lemma B.3 (Stirling's Formula)**.** *For all $n \geq 1$,*

$$e^{\frac{1}{12n+1}} \sqrt{2\pi n} n^n e^{-n} \leq n! \leq e^{\frac{1}{12n}} \sqrt{2\pi n} n^n e^{-n}. \tag{B.2}$$

A proof of this version of Stirling's Formula can be found in [285].

B.1.2 Double sequences

We say that a double sequence $(a_{m,n})_{m,n\geq 1}$ is **nondecreasing** if

$$m \leq m', n \leq n' \implies a_{m,n} \leq a_{m',n'},$$

and **nonincreasing** if $(-a_{m,n})_{m,n\geq 1}$ is nondecreasing. It is **bounded above** (resp. **below**) if there exists $C < \infty$ such that $a_{m,n} \leq C$ (resp. $a_{m,n} \geq -C$), for all $m, n \geq 1$.

Lemma B.4. *Let $(a_{m,n})_{m,n\geq 1}$ be a nondecreasing double sequence bounded above. Then,*

$$\lim_{m\to\infty} \lim_{n\to\infty} a_{m,n} = \lim_{n\to\infty} \lim_{m\to\infty} a_{m,n} = \lim_{m,n\to\infty} a_{m,n} = \sup\big\{a_{m,n} : m, n \geq 1\big\}. \tag{B.3}$$

Proof. $(a_{m,n})_{m,n\geq 1}$ being bounded, $s \stackrel{\text{def}}{=} \sup_{m,n} a_{m,n}$ is finite. Let $\epsilon > 0$, and take m_0, n_0 such that $a_{m_0,n_0} \geq s - \epsilon$. $(a_{m,n})$ being nondecreasing, we deduce that

$$s \geq a_{m,n} \geq s - \epsilon, \qquad \forall\, m \geq m_0, n \geq n_0.$$

Consequently, $\lim_{m,n\to\infty} a_{m,n} = s$. For all fixed $m \geq 1$, the sequence $(a_{m,n})_{n\geq 1}$ is nondecreasing and bounded, and thus converges to some limit s_m. For a fixed $\epsilon > 0$, let m_1, n_1 be such that

$$|a_{m,n} - s| \leq \tfrac{\epsilon}{2}, \qquad \forall\, m \geq m_1, n \geq n_1.$$

For fixed m, we can also find $n_2(m)$ such that

$$|a_{m,n} - s_m| \leq \tfrac{\epsilon}{2}, \qquad \forall\, n \geq n_2(m).$$

Consequently,

$$|s_m - s| \leq \epsilon, \qquad \forall\, m \geq m_1,$$

which implies that $\lim_{m\to\infty} s_m = s$. We have thus proved (B.3). □

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

B.1.3 Subadditive sequences

A sequence $(a_n)_{n\geq 1} \subset \mathbb{R}$ is called **subadditive** if

$$a_{n+m} \leq a_n + a_m \quad \forall m, n .$$

Lemma B.5. *If $(a_n)_{n\geq 1}$ is subadditive, then*

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_n \frac{a_n}{n} .$$

Proof. Let $\alpha \stackrel{\text{def}}{=} \inf_n \frac{a_n}{n}$, and fix $\alpha' > \alpha$. Let ℓ be such that $\frac{a_\ell}{\ell} \leq \alpha'$. For all n, there exists k and $0 \leq j < \ell$ such that $n = k\ell + j$. We can then use the definition of α, and k times the subadditivity of a_n to write

$$\alpha n \leq a_n = a_{k\ell+j} \leq k a_\ell + a_j .$$

Dividing by n,

$$\alpha \leq \liminf_{n\to\infty} \frac{a_n}{n} \leq \limsup_{n\to\infty} \frac{a_n}{n} \leq \frac{a_\ell}{\ell} \leq \alpha' .$$

The desired result follows by letting $\alpha' \downarrow \alpha$. □

On the lattice $\mathbb{Z}^d$, a similar property holds. Let us denote by $\mathscr{R}$ the set of all **parallelepipeds** of $\mathbb{Z}^d$, that is sets of the form $\Lambda = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d] \cap \mathbb{Z}^d$. A set function $a : \mathscr{R} \to \mathbb{R}$ is **subadditive** if $R_1, R_2 \in \mathscr{R}$, $R_1 \cup R_2 \in \mathscr{R}$ implies

$$a(R_1 \cup R_2) \leq a(R_1) + a(R_2) .$$

Let, as usual, $\mathsf{B}(n) = \{-n, \ldots, n\}^d$.

Lemma B.6. *Let $a : \mathscr{R} \to \mathbb{R}$ be subadditive and such that $a(\Lambda + i) = a(\Lambda)$ for all $\Lambda \in \mathscr{R}$ and all $i \in \mathbb{Z}^d$. Then*

$$\lim_{n\to\infty} \frac{a(\mathsf{B}(n))}{|\mathsf{B}(n)|} = \inf_{\Lambda\in\mathscr{R}} \frac{a(\Lambda)}{|\Lambda|} .$$

The proof is a d-dimensional adaptation of the one given above for sequences $(a_n)_{n\geq 1}$; we leave it as an exercise (a proof can be found in [134]).

B.1.4 Functions defined by series

Theorem B.7. *Let $I \subset \mathbb{R}$ be an open interval. For each $k \geq 1$, let $\phi_k : I \to \mathbb{R}$ be C^1. Assume that there exists a summable sequence $(\epsilon_k)_{k\geq 1} \subset \mathbb{R}_{\geq 0}$ such that $\sup_{x\in I} |\phi_k(x)| \leq \epsilon_k$, $\sup_{x\in I} |\phi_k'(x)| \leq \epsilon_k$. Then $f(x) \stackrel{\text{def}}{=} \sum_{k\geq 1} \phi_k(x)$ is well defined and C^1 on I. Moreover, $f'(x) = \sum_{k\geq 1} \phi_k'(x)$.*

Proof. Since $\sum_k \epsilon_k < \infty$, $\sum_k \phi_k(x)$ is an absolutely convergent series for all $x \in I$, defining a function $f : I \to \mathbb{R}$. Then, fix $x \in I$ and take some small $h > 0$:

$$\frac{f(x+h) - f(x)}{h} = \sum_k \frac{\phi_k(x+h) - \phi_k(x)}{h} .$$

By the mean-value theorem, there exists $\tilde{x} \in [x, x+h]$ such that $|\frac{\phi_k(x+h)-\phi_k(x)}{h}| = |\phi_k'(\tilde{x})| \leq \epsilon_k$. Using Exercise B.15, we can therefore interchange $h \downarrow 0$ with $\sum_k$. The same argument with $h \uparrow 0$ then gives $f'(x) = \sum_k \phi_k'(x)$. A similar argument guarantees that f is C^1. □

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

B.2 Convex functions

In this section, we gather a few elementary results about convex functions of one real variable. Rockafellar's book [287] is a standard reference on the subject; another nice and accessible reference is the book [286] by Roberts and Varberg.

We will use I to denote an (not necessarily bounded) open interval in $\mathbb{R}$, that is, $I = (a, b)$ with $-\infty \le a < b \le +\infty$.

Definition B.8. *A function $f : I \to \mathbb{R}$ is **convex** if*

$$f\big(\alpha x + (1-\alpha) y\big) \le \alpha f(x) + (1-\alpha) f(y), \qquad \forall x, y \in I, \forall \alpha \in [0,1]. \tag{B.4}$$

*When the inequality is strict for all $x \neq y$ and all $\alpha \in (0,1)$, f is **strictly convex**. If $-f$ is (strictly) convex, then f is said to be **(strictly) concave**.*

For the cases considered in the book, f always has a continuous extension to the boundary of I (when I is finite). Sometimes, we need to extend the domain of f from a finite I to the whole of $\mathbb{R}$; in such cases, one can do that by setting $f(x) \stackrel{\text{def}}{=} +\infty$ for all $x \notin I$. The definition of convexity can then be extended, allowing f to take infinite values in B.4.

The following exercise is elementary, but emphasizes a property of convex functions that will be used repeatedly in the sequel; it is illustrated on Figure B.1.

Exercise B.1. *Show that $f : I \to \mathbb{R}$ is convex if and only if, for any $x < y < z$ in I,*

$$f(y) \le \frac{z-y}{z-x} f(x) + \frac{y-x}{z-x} f(z). \tag{B.5}$$

From this, deduce that if f is finite, then, for any $x < y < z$ in I,

$$\frac{f(y)-f(x)}{y-x} \le \frac{f(z)-f(x)}{z-x} \le \frac{f(z)-f(y)}{z-y}. \tag{B.6}$$

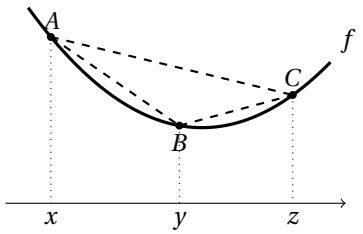


Figure B.1: The geometrical meaning of (B.6): for any triple of points on the graph of a convex function, slope$(AB) \le$ slope$(AC) \le$ slope(BC).

Exercise B.2. *Show that $f : I \to \mathbb{R}$ is convex if and only if, for all $\alpha_1, \dots, \alpha_n \in [0,1]$ such that $\alpha_1 + \cdots + \alpha_n = 1$ and all $x_1, \dots, x_n \in I$,*

$$f\Big(\sum_{k=1}^{n} \alpha_k x_k\Big) \le \sum_{k=1}^{n} \alpha_k f(x_k).$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

An important property is that limits of convex functions are convex:

Exercise B.3. *Show that if $(f_n)_{n\geq 1}$ is a sequence of convex functions from I to $\mathbb{R}$, then $x \mapsto \limsup_{n\to\infty} f_n(x)$ is convex. In particular, if $f(x) \stackrel{\text{def}}{=} \lim_{n\to\infty} f_n(x)$ exists (in $\mathbb{R}\cup\{+\infty\}$) for all $x \in I$, then it is also convex.*

B.2.1 Convexity vs. continuity

Proposition B.9. *Let $f : I \to \mathbb{R}$ be convex. Then f is locally Lipschitz: for all compact $K \subset I$, there exists $C_K < \infty$ such that $|f(x) - f(y)| \leq C_K|x - y|$ for all $x, y \in K$. In particular, f is continuous.*

Proof. Let $K \subset I$ be compact, and let $\epsilon > 0$ be small enough to ensure that $K_\epsilon \stackrel{\text{def}}{=} \{z : d(z, K) \leq \epsilon\} \subset I$. Let also $M \stackrel{\text{def}}{=} \sup_{z\in K_\epsilon} f(z)$, $m \stackrel{\text{def}}{=} \inf_{z\in K_\epsilon} f(z)$. Observe that both m and M are finite. (Otherwise, there would exist an interior point $x_* \in I$ and a sequence $x_n \to x_*$, $f(x_n) \uparrow +\infty$. Then, for all pair $z < x_* < z'$ one would get, for all sufficiently large n, $f(x_n) > \max\{f(z), f(z')\}$, a contradiction with the convexity of f.) Let $x, y \in K$, and set $z \stackrel{\text{def}}{=} y + \epsilon\frac{y-x}{|y-x|} \in K_\epsilon$. Then $y = (1-\lambda)x + \lambda z$ with $\lambda = \frac{|y-x|}{\epsilon+|y-x|}$, and therefore $f(y) \leq (1-\lambda)f(x) + \lambda f(z)$, which gives after rearrangement

$$f(y) - f(x) \leq \lambda(f(z) - f(x)) \leq \lambda(M - m) \leq \frac{M-m}{\epsilon}|y - x|. \qquad \square$$

Lemma B.10. *Let $(f_n)_{n\geq 1}$ be a sequence of convex functions on I converging pointwise to $f : I \to \mathbb{R}$. Then $f_n \to f$ uniformly on all compacts $K \subset I$.*

Proof. Fix some compact $K \subset I$, and let $a' < a < b < b'$ in I such that $[a, b] \supset K$.

It follows from (B.6) that, for all n and all distinct $x, y \in [a, b]$,

$$\frac{f_n(a) - f_n(a')}{a - a'} \leq \frac{f_n(y) - f_n(x)}{y - x} \leq \frac{f_n(b') - f_n(b)}{b' - b}.$$

By pointwise convergence, the leftmost and rightmost ratios converge to finite values as $n \to \infty$. Therefore, there exists C, independent of n, such that

$$|f_n(y) - f_n(x)| \leq C|y - x| \quad \forall x, y \in [a, b].$$

Letting $n \to \infty$ in the last display shows that the same is also true for the limiting function f.

Fix $\epsilon > 0$. Let $N \in \mathbb{N}$ and define $\delta = (b - a)/N$ and $x_k = a + k\delta$, $k = 0, \ldots, N$. Pointwise convergence implies that there exists n_0 such that, for all $n \geq n_0$,

$$|f_n(x_k) - f(x_k)| < \tfrac{1}{3}\epsilon, \quad \forall k \in \{0, \ldots, N\}.$$

Let $z \in [a, b]$ and let $k \in \{0, \ldots, N\}$ be such that $|x_k - z| < \delta$. Then, for all $n \geq n_0$,

$$|f_n(z) - f(z)| \leq \underbrace{|f_n(z) - f_n(x_k)|}_{\leq C\delta} + \underbrace{|f_n(x_k) - f(x_k)|}_{\leq \epsilon/3} + \underbrace{|f(x_k) - f(z)|}_{\leq C\delta} \leq \epsilon,$$

provided we choose N such that $C\delta \leq \epsilon/3$. $\square$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

A function $f : I \to \mathbb{R}$ is said to be **midpoint-convex** if

$$f\Big(\frac{x+y}{2}\Big) \le \frac{f(x)+f(y)}{2}, \qquad \forall\, x, y \in I. \tag{B.7}$$

Clearly, a convex function is also midpoint-convex.

Lemma B.11. *If f is midpoint-convex and continuous, then it is convex.*

Proof. Using the continuity of f, it suffices to show that (B.4) holds in the case where $\alpha \in \mathscr{D} \stackrel{\text{def}}{=} \bigcup_{m\ge 1} \mathscr{D}_m$, with $\mathscr{D}_m \stackrel{\text{def}}{=} \big\{\frac{k}{2^m} : 0 \le k < 2^m\big\}$. Observe first that (B.7) means that (B.4) holds for all $x, y \in I$ and for $\alpha \in \mathscr{D}_1$.

We can now proceed by induction. Assume that (B.4) holds for all $\alpha \in \mathscr{D}_m$. Let $z = \alpha x + (1-\alpha) y$, with $\alpha \in \mathscr{D}_{m+1} \setminus \mathscr{D}_m$; with no loss of generality, we can assume that $\alpha > 1/2$. Let $z' \stackrel{\text{def}}{=} 2z - x = \alpha' x + (1-\alpha') y$ where $\alpha' \stackrel{\text{def}}{=} 2\alpha - 1 \in \mathscr{D}_m$. Applying (B.7) and the induction assumption, we get

$$\begin{aligned} f(z) = f(\tfrac{1}{2}x + \tfrac{1}{2}z') &\le \tfrac{1}{2} f(x) + \tfrac{1}{2} f(z') \\ &\le \tfrac{1}{2} f(x) + \tfrac{1}{2}\big\{\alpha' f(x) + (1-\alpha') f(y)\big\} = \alpha f(x) + (1-\alpha) f(y)\,, \end{aligned}$$

so (B.4) also holds for all $\alpha \in \mathscr{D}_{m+1}$. □

B.2.2 Convexity vs. differentiability

The **one-sided derivatives** of a function f at a point x are defined by

$$\partial^+ f(x) = \frac{\partial f}{\partial x^+} \stackrel{\text{def}}{=} \lim_{z \downarrow x} \frac{f(z) - f(x)}{z - x}\,,$$
$$\partial^- f(x) = \frac{\partial f}{\partial x^-} \stackrel{\text{def}}{=} \lim_{z \uparrow x} \frac{f(z) - f(x)}{z - x}\,.$$

These quantities are always well defined for a convex function, and enjoy several useful properties:

Theorem B.12. *Let $f : I \to \mathbb{R}$ be convex. The following properties hold.*

1. *$\partial^+ f(x)$ and $\partial^- f(x)$ exist at all points $x \in I$.*
2. *$\partial^- f(x) \le \partial^+ f(x)$, for all $x \in I$.*
3. *$\partial^+ f(x) \le \partial^- f(y)$ for all $x < y$ in I.*
4. *$\partial^+ f$ and $\partial^- f$ are nondecreasing.*
5. *$\partial^+ f$ is right-continuous, $\partial^- f$ is left-continuous.*
6. *$\big\{x : \partial^+ f(x) \ne \partial^- f(x)\big\}$ is at most countable.*
7. *Let $(g_n)_{n\ge 1}$ be a sequence of convex functions from I to $\mathbb{R}$ converging pointwise to a function g. If g is differentiable at x, then $\lim_{n\to\infty} \partial^+ g_n(x) = \lim_{n\to\infty} \partial^- g_n(x) = g'(x)$.*

Note that Item 6 shows that a convex function $f : I \to \mathbb{R}$ is differentiable everywhere outside an at most countable set.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Proof. From (B.6), we see that

$$x \mapsto \frac{f(y)-f(x)}{y-x} \quad \text{and} \quad y \mapsto \frac{f(y)-f(x)}{y-x} \quad \text{are nondecreasing.} \tag{B.8}$$

1. In I, consider $x < y$ and a decreasing sequence $(z_k)_{k\geq 1}$ with $z_k > y$, for all k, and $z_k \downarrow y$. From (B.8), the sequence $\big(\big(f(z_k)-f(y)\big)/(z_k-y)\big)_{k\geq 1}$ is nonincreasing, and (B.6) implies that it is bounded below by $\big(f(y)-f(x)\big)/(y-x)$. It follows that the sequence converges, which establishes the existence of $\partial^+ f(y)$. A similar argument proves the existence of $\partial^- f(y)$.

2. Taking $x \uparrow y$ in the left-hand side, followed by $z \downarrow y$ in the right-hand side of (B.6) gives $\partial^- f(y) \leq \partial^+ f(y)$.

3. Let $x < y$ in I. It follows from (B.8) that

$$\partial^+ f(x) \leq \frac{f(y)-f(x)}{y-x} \leq \partial^- f(y)\,. \tag{B.9}$$

4. This is a consequence of the second and third points.

5. We prove the claim for $\partial^+ f$; the other one is treated in the same way. On the one hand, it follows from the monotonicity of $\partial^+ f$ that $\lim_{y\downarrow x}\partial^+ f(y)$ exists and $\lim_{y\downarrow x}\partial^+ f(y) \geq \partial^+ f(x)$. On the other hand, we know from Proposition B.9 that f is continuous. It thus follows from (B.9) that, for each $z > x$,

$$\frac{f(z)-f(x)}{z-x} = \lim_{y\downarrow x}\frac{f(z)-f(y)}{z-y} \geq \lim_{y\downarrow x}\partial^+ f(y)\,.$$

Letting $z \downarrow x$, we obtain that $\partial^+ f(x) \geq \lim_{y\downarrow x}\partial^+ f(y)$ and the claim follows.

6. Since I can be written as the union of countably many closed intervals and since a countable union of countable sets is countable, it is enough to prove the statement for an arbitrary closed interval $[a,b]$ contained in I. Let $\epsilon > 0$ such that $[a-\epsilon, b+\epsilon] \subset I$. Since f is continuous, $M \stackrel{\text{def}}{=} \sup_{x\in[a-\epsilon,b+\epsilon]} |f(x)| < \infty$. It thus follows from (B.9) that

$$\partial^+ f(b) \leq \frac{f(b+\epsilon)-f(b)}{\epsilon} \leq \frac{2M}{\epsilon}$$

and

$$\partial^- f(a) \geq \frac{f(a)-f(a-\epsilon)}{\epsilon} \geq -\frac{2M}{\epsilon}\,.$$

By what we saw above, $\partial^- f(a) \leq \partial^\pm f(x) \leq \partial^+ f(b)$ for all $x \in [a,b]$, we deduce that $\sup_{x\in[a,b]} |\partial^\pm f(x)| \leq 2M/\epsilon$. For $r \in \mathbb{N}$, let

$$\mathscr{A}_r = \big\{x \in [a,b] : \partial^+ f(x) - \partial^- f(x) \geq \tfrac{1}{r}\big\}\,.$$

Since

$$\big\{x \in [a,b] : \partial^+ f(x) > \partial^- f(x)\big\} = \bigcup_{r\geq 1}\mathscr{A}_r,$$

it suffices to prove that each $\mathscr{A}_r$ is finite. Consider n distinct points $x_1 < x_2 < \ldots < x_n$ from $\mathscr{A}_r$. Then,

$$\partial^+ f(x_n) - \partial^- f(x_1) = \sum_{k=1}^{n}\big(\partial^+ f(x_k) - \partial^- f(x_k)\big) \geq \frac{n}{r}\,,$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

which implies $n \le r\big(\partial^+ f(x_n) - \partial^- f(x_1)\big) \le 4Mr/\epsilon$; $\mathscr{A}_r$ is therefore finite.

7.

Using again (B.9), for any $h > 0$,

$$\limsup_{n\to\infty} \partial^+ g_n(x) \le \limsup_{n\to\infty} \frac{g_n(x+h) - g_n(x)}{h} = \frac{g(x+h) - g(x)}{h}.$$

Letting $h \downarrow 0$ gives $\partial^+ g(x) \ge \limsup_{n\to\infty} \partial^+ g_n(x)$. A similar argument yields $\partial^- g(x) \le \liminf_{n\to\infty} \partial^- g_n(x)$. Therefore,

$$\partial^- g(x) \le \liminf_{n\to\infty} \partial^- g_n(x) \le \limsup_{n\to\infty} \partial^+ g_n(x) \le \partial^+ g(x),$$

and the differentiability of g at x indeed implies that

$$g'(x) = \lim_{n\to\infty} \partial^- g_n(x) = \lim_{n\to\infty} \partial^+ g_n(x). \qquad \square$$

We say that $f : I \mapsto \mathbb{R}$ has a **supporting line of slope m at x_0** if

$$f(x) \ge m(x - x_0) + f(x_0), \quad \forall x \in I. \tag{B.10}$$

Theorem B.13. *A function $f : I \to \mathbb{R}$ is convex if and only if f has a supporting line at each point $x \in I$. Moreover, in that case, there is a supporting line at x of slope m for all $m \in [\partial^- f(x), \partial^+ f(x)]$.*

Proof. Suppose first that f has a supporting line at each point of I. Let $x < y$ be two points of I, $\alpha \in [0,1]$ and $z = \alpha x + (1-\alpha) y$. By assumption, there exists m such that $f(u) \ge f(z) + m(u - z)$ for all $u \in I$. Applying this at x and y, we deduce that

$$\alpha f(x) + (1-\alpha) f(y) \ge f(z) + m\big(\underbrace{\alpha(x-z) + (1-\alpha)(y-z)}_{=0}\big),$$

which implies that $\alpha f(x) + (1-\alpha) f(y) \ge f(\alpha x + (1-\alpha) y)$ as desired.

Assume now that f is convex and let $x_0 \in I$. Let $m \in [\partial^- f(x_0), \partial^+ f(x_0)]$. By (B.9), $\frac{f(x) - f(x_0)}{x - x_0} \ge \partial^+ f(x_0) \ge m$ for all $x > x_0$, and $\frac{f(x) - f(x_0)}{x - x_0} \le \partial^- f(x_0) \le m$ for all $x < x_0$, which implies $f(x) \ge m(x - x_0) + f(x_0)$ for all x. $\square$

We also remind the reader of a well-known property that relates convexity to the positivity of the second derivative of a twice-differentiable function:

Exercise B.4. *Let f be twice-differentiable at each point of I. Show that f is convex if and only if $f''(x) \ge 0$ for all $x \in I$.*

Note that a sequence of strictly convex functions $(f_n)_{n\ge1}$ converging pointwise can have a limit that is not strictly convex; consider, for example, $f_n(x) = |x|^{1+1/n}$. A twice-differentiable function f for which one can find $c > 0$ such that $f''(x) > c$ for all x is said to be **strongly convex**. Note that a function can be strictly convex and fail to be strongly convex, for example $x \mapsto x^4$.

Exercise B.5. *Let $(f_n)_{n\ge1}$ be a sequence of twice-differentiable strongly convex functions such that $f = \lim_n f_n(x)$ exists and is finite everywhere. Show that f is strictly convex.*

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

B.2.3 The Legendre transform

Definition B.14. *Let $f : I \to \mathbb{R} \cup \{+\infty\}$. The **Legendre–Fenchel Transform** (or simply **Legendre transform** [1]) of f is defined by*

$$f^*(y) \stackrel{\text{def}}{=} \sup_{x \in I} \{yx - f(x)\}, \quad y \in \mathbb{R}. \tag{B.11}$$

We will always suppose, from now on, that there exists at least one point at which f is finite, which guarantees that $f^*(y) > -\infty$ for all $y \in \mathbb{R}$.

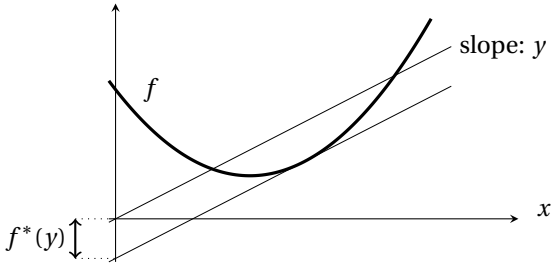


Figure B.2: Visualizing the Legendre transform: for a given $y \in \mathbb{R}$, $f^*(y)$ is the largest difference between the straight line $x \mapsto yx$ and the graph of f.

Exercise B.6. *Show that any Legendre transform is convex.*

Exercise B.7. *Compute the Legendre transform f_i^* of each of the following functions:*

$$f_1(x) = \tfrac{1}{2}x^2, \qquad f_2(x) = x^4, \qquad f_3(x) = \begin{cases} 0 & \text{if } x \in (-1,1), \\ +\infty & \text{if } x \notin (-1,1). \end{cases}$$

Compute also $f_i^{**} \stackrel{\text{def}}{=} (f_i^*)^*$ in each case. What do you observe?

As can be seen by solving the previous exercise, f^{**} is not always equal to f. Nevertheless,

Exercise B.8. *Show that, for all $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, $f^{**} \le f$.*

Let us see two more examples in which the geometrical effect of applying two successive Legendre transforms is made transparent:

Exercise B.9. *If $f(x) = \big||x| - 1\big|$, show that*

$$f^{**}(x) = \begin{cases} -x - 1 & \text{if } x < -1, \\ 0 & \text{if } |x| \le 1, \\ +x - 1 & \text{if } x > +1. \end{cases}$$

[1] Actually, the latter form is usually reserved for a particular case; nevertheless, we use the term *Legendre transform* everywhere in this book.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Exercise B.10. *Let $f(x) = x^4 - x^2$. Using the geometrical picture of Figure B.2, study qualitatively f^* and f^{**}.*

With the above examples in mind, we now impose restrictions on f to guarantee that $f^{**} = f$.

We call f **lower semi-continuous at x** if, for any sequence $x_n \to x$,

$$\liminf_{n\to\infty} f(x_n) \geq f(x)\,.$$

Exercise B.11. *Show that any Legendre transform is lower semi-continuous.*

Lemma B.15. *Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be convex and lower semi-continuous. For all x_0, if $\alpha \in \mathbb{R}$ is such that $\alpha < f(x_0)$, then there exists an **affine function** $h(x) = ax + b$ such that $h \leq f$ and $h(x_0) \geq \alpha$.*

Proof. For simplicity, we assume that $f(x_0) < +\infty$ (the case $f(x_0) = +\infty$ is treated similarly). If either $\partial^+ f(x_0)$, or $\partial^- f(x_0)$, is finite, then Theorem B.13 implies the result. If $\partial^- f(x_0) = +\infty$, convexity implies that $f(x) < f(x_0)$ for all $x \in (x_0 - \delta, x_0)$ (with $\delta > 0$ sufficiently small), and $f(x) = +\infty$ for all $x > x_0$. Let

$$a \stackrel{\text{def}}{=} \inf\big\{m \geq 0 : m(x - x_0) + \alpha \leq f(x),\, \forall\, x \in I\big\}\,.$$

We claim that $a < \infty$. Indeed, assume that $a = \infty$. Then, there would exist a sequence $x_n < x_0$, $x_n \uparrow x_0$, with $f(x_n) \leq \alpha < f(x_0)$, giving $\liminf_n f(x_n) < f(x_0)$, which would contradict the lower semi-continuity of f. When $a < \infty$, the affine function $h(x) = a(x - x_0) + \alpha$ satisfies the requirements. The remaining cases are treated similarly. □

Exercise B.12. *Show that if f has a supporting line of slope m at x_0, then f^* has a supporting line of slope x_0 at m.*

The **epigraph** of an arbitrary function $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$\mathrm{epi}\,(f) \stackrel{\text{def}}{=} \big\{(x, y) \in \mathbb{R}^2 : y \geq f(x)\big\}\,.$$

Exercise B.13. *Let $f : I \to \mathbb{R} \cup \{+\infty\}$.*

1. *Show that f is convex if and only if $\mathrm{epi}\,(f)$ is convex*[2].
2. *Show that f is lower semi-continuous if and only if $\mathrm{epi}\,(f)$ is closed.*

Definition B.16. *The **convex envelope** (or **convex hull**) of f, denoted $\mathrm{CE}\,f$, is defined as the unique convex function g whose epigraph is*

$$C \stackrel{\text{def}}{=} \bigcap \big\{F \subset \mathbb{R}^2 : F \text{ closed, convex, } F \supset \mathrm{epi}\,(f)\big\}\,. \tag{B.12}$$

That is,

$$\mathrm{CE}\,f(x) \stackrel{\text{def}}{=} \inf\big\{y : (x, y) \in C\big\}\,.$$

[2] $A \subset \mathbb{R}^2$ is **convex** if $z_1, z_2 \in A$, $\lambda \in [0,1]$ implies $\lambda z_1 + (1-\lambda) z_2 \in A$.

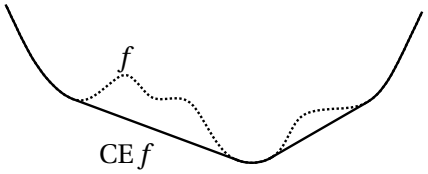
Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Clearly, if f is convex and lower semi-continuous, then $\mathrm{CE}\, f = f$.

Observe that, since C is closed, $(x, \mathrm{CE}\, f(x)) \in C$ for all x. Moreover, C is convex, which implies that $x \mapsto \mathrm{CE}\, f(x)$ is convex. In fact, $\mathrm{epi}\,(\mathrm{CE}\, f) = C$. Since C is closed, this implies (Exercise B.13) that $\mathrm{CE}\, f$ is lower semi-continuous.

In words, as will be seen in the next exercise, $\mathrm{CE}\, f$ is the largest convex function g such that $g \le f$:



Exercise B.14. *If g is convex, lower semi-continuous and $g \le f$, then $g \le \mathrm{CE}\, f$.*

Theorem B.17. *If $f : \mathbb{R} \to \mathbb{R}$ is lower semi-continuous,*

$$f^{**} = \mathrm{CE}\, f\,.$$

Proof. We have already seen that $f^{**} \le f$. Since f^{**} is convex and lower semi-continuous, this implies $f^{**} \le \mathrm{CE}\, f$ (Exercise B.14). To establish the reverse inequality at a point x_0, $f^{**}(x_0) \ge \mathrm{CE}\, f(x_0)$, we must show that, for all $\alpha \in \mathbb{R}$ satisfying $\alpha < \mathrm{CE}\, f(x_0)$,

$$\text{there exists } y \in \mathbb{R} \text{ such that } x_0 y - f^*(y) \ge \alpha\,. \tag{B.13}$$

Since $\mathrm{CE}\, f$ is also lower semi-continuous, there exists, by Lemma B.15, an affine function h such that (i) $h \le f$ and (ii) $\alpha \le h(x_0) \le f(x_0)$. If $h(x) = ax + b$, (i) means that $ax + b \le f(x)$ for all x, which gives $f^*(a) \le -b$. Then, (ii) implies that $\alpha \le ax_0 + b$. Combining these bounds gives $ax_0 - f^*(a) \ge \alpha$, which implies (B.13). □

Corollary B.18. *If $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous, then*

$$(\mathrm{CE}\, f)^* = f^*\,.$$

Proof. By Theorem B.17, $\mathrm{CE}\, f = f^{**}$, and so $(\mathrm{CE}\, f)^* = f^{***}$. Since f^* is lower semi-continuous, we have again by Theorem B.17 that $f^{***} = (f^*)^{**} = \mathrm{CE}\, f^*$. But f^* is convex, which implies that $\mathrm{CE}\, f^* = f^*$. □

In particular, we proved:

Theorem B.19. *If $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is lower-semicontinuous and convex,*

$$f^{**} = f\,.$$

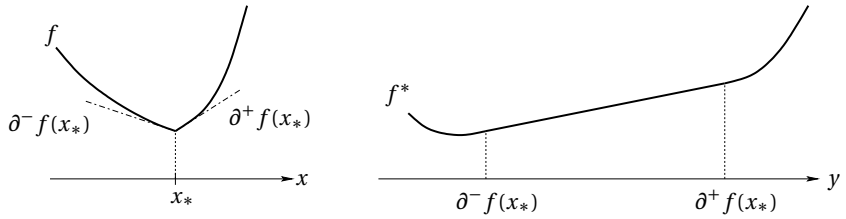
B.2.4 Legendre transform of non-differentiable functions

We have seen that the right and left derivatives of a convex function f at a point x_*, $\partial^+ f(x_*)$ and $\partial^- f(x_*)$, are well defined (Theorem B.12). If f is not differentiable at x_*, then $\partial^- f(x_*) < \partial^+ f(x_*)$, so f can have more than one supporting line at x_*,

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

which has an important consequence on the qualitative behavior of the Legendre Transform.



Theorem B.20. *Let f be a convex function. Then:*

1. *If f is not differentiable at x_*, then f^* is affine on the interval $[\partial^- f(x_*), \partial^+ f(x_*)]$.*
2. *If f is affine on some interval $[a,b]$, with a slope m, then f^* is non-differentiable at m and $\partial^+ f^*(m) \geq b > a \geq \partial^- f^*(m)$.*

Note that if a continuous function f is not convex on an interval contining x, then $\mathrm{CE}\, f$ must be affine on that interval. In that case, the above theorem, combined with Corollary B.18, shows that f^* cannot be differentiable.

Proof. By Theorem B.13, for each value $m \in [\partial^- f(x_*), \partial^+ f(x_*)]$, the line $x \mapsto m(x - x_*) + f(x_*)$ is a supporting line for f at x_*. By Exercise B.12, this implies that f^* admits, at each $m \in [\partial^- f(x_*), \partial^+ f(x_*)]$, a supporting line with the same slope x_*. Since f^* is convex, all these supporting lines actually coincide, which implies that f^* is affine on the interval.

If $\mathrm{CE}\, f$ is affine on $[a,b]$, with slope m, one has in particular that $f^*(m) = (\mathrm{CE}\, f)^*(m) = ma - f(a) = mb - f(b)$. Then, for all $\epsilon > 0$,

$$f^*(m+\epsilon) - f^*(m) \geq \big\{(m+\epsilon)b - f(b)\big\} - f^*(m) = \epsilon b,$$

and therefore $\partial^+ f^*(m) \geq b$. Similarly, $\partial^- f^*(m) \leq a$. □

B.3 Complex analysis

Let $D \subset \mathbb{C}$ be a domain (that is, open and connected). Remember that a function $f : D \to \mathbb{C}$ is **holomorphic** if

$$f'(z) \stackrel{\text{def}}{=} \lim_{w \to z} \frac{f(w) - f(z)}{w - z}$$

exists and is finite at each $z \in D$. It is well known that holomorphic functions have derivatives of all orders, and that f is holomorphic if and only if it is **analytic**, that is, if and only if it can be represented at each point $z_0 \in D$ by a convergent Taylor series:

$$f(z) = \sum_{n \geq 0} a_n (z - z_0)^n,$$

where $a_n = \frac{1}{n!} f^{(n)}(z_0)$ and z belongs to a small disk around z_0. Therefore, holomorphic and analytic should be considered as synonyms in this section.

We start with the following fundamental result of complex analysis.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Theorem B.21 (Cauchy's integral theorem)**.** *Let $D \subset \mathbb{C}$ be open and simply connected, and let f be holomorphic on D. Then*

$$\oint_\gamma f(\xi)\mathrm{d}\xi = 0,$$

for all closed paths $\gamma \subset D$.

Proof. See, for example, [336, Theorem 4.14]. □

Corollary B.22. *Let $D \subset \mathbb{C}$ be open and simply connected, and let f be holomorphic on D. Then, there exists a function F, holomorphic on D, such that $F' = f$.*

Proof. We fix some point $z_0 \in D$. Since D is open and connected, any other $z \in D$ can be joined from z_0 by a continuous path $\gamma_z \subset D$. Let

$$F(z) \stackrel{\text{def}}{=} \int_{\gamma_z} f(\xi)\mathrm{d}\xi.$$

By Theorem B.21, this definition does not depend on the choice of the path γ_z. Choose $r > 0$ small enough to ensure that the disc $B(z,r) \stackrel{\text{def}}{=} \{w \in \mathbb{C} : |w - z| \le r\} \subset D$. Then,

$$F(w) = F(z) + \int_{[z,w]} f(\xi)\mathrm{d}\xi, \qquad \forall\, w \in B(z,r),$$

where $[z,w]$ is the straight line segment connecting z to w. But f being holomorphic implies in particular that $f(\xi) = f(z) + O(|\xi - z|)$ for all $\xi \in B(z,r)$, and so

$$\lim_{w\to z} \frac{F(w) - F(z)}{w - z} = \lim_{w\to z} \frac{1}{w - z} \int_{[z,w]} f(\xi)\mathrm{d}\xi = f(z)\,.$$

This implies that F is holomorphic and that $F' = f$. □

Theorem B.23. *Let f be a holomorphic function on a simply connected open set $D \subset \mathbb{C}$, which has no zeroes on D. Then, there exists a function g analytic on D, called a **branch of the logarithm of f on D**, such that $f = e^g$.*

Proof. Our assumptions imply that f'/f is holomorphic on D. Corollary B.22 thus implies the existence of a function F, holomorphic on D, such that $F' = f'/f$. In particular,

$$(fe^{-F})' = f'e^{-F} - fF'e^{-F} \equiv 0.$$

Therefore, there exists $c \in \mathbb{C}$ such that $fe^{-F} = e^c$, or equivalently $f = e^{F+c}$. □

Remark B.24. 1. Let g be a branch of the logarithm of f on D. Then $\mathfrak{Re}\, g = \log|f|$. Indeed,

$$|f| = |e^g| = |e^{\mathfrak{Re}\, g} e^{\mathrm{i}\,\mathfrak{Im}\, g}| = e^{\mathfrak{Re}\, g}.$$

2. Let g_1 and g_2 be two branches of the logarithm of f on D. Since, for each $z \in D$,

$$e^{g_2(z) - g_1(z)} = \frac{e^{g_2(z)}}{e^{g_1(z)}} = \frac{f(z)}{f(z)} = 1\,,$$

we conclude that $g_2(z) = g_1(z) + 2\mathrm{i}k(z)\pi$ for some $k(z) \in \mathbb{Z}$. However, $z \mapsto k(z) = (g_2(z) - g_1(z))/2\mathrm{i}\pi$ is continuous and integer-valued; it is therefore constant on D. This implies that $g_2 = g_1 + 2\mathrm{i}k\pi$ for some $k \in \mathbb{Z}$.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

3. Assume that the domain D in Theorem B.23 is such that $D \cap \mathbb{R}$ is connected. Suppose also that $f(z) \in \mathbb{R}_{>0}$ for $z \in D \cap \mathbb{R}$. Then there is a branch g of the logarithm of f on D such that $g(z) \in \mathbb{R}$ for all $z \in D \cap \mathbb{R}$; in particular, g coincides with the usual logarithm of f (seen as a real function) on $D \cap \mathbb{R}$. Indeed, it suffices to observe that the function F in the proof can be constructed by starting from a point $z_0 \in D \cap \mathbb{R}$ at which one can fix $g(z_0) = \log|f(z_0)|$ and use the fact that $F(z) = \int_{z_0}^{z} f'(x)/f(x)\,\mathrm{d}x$ for $z \in D \cap \mathbb{R}$. ⋄

It is well known that the limit of a sequence of analytic functions need not be analytic. Let us see how additional conditions can be imposed to guarantee that the limiting function is also analytic.

Remember that a family $\mathscr{A}$ of functions on $\mathbb{C}$ is **locally uniformly bounded** on a set $D \subset \mathbb{C}$ if, for each $z \in D$, there exists a real number M and a neighborhood $\mathscr{U}$ of z such that $|f(w)| \le M$ for all $w \in \mathscr{U}$ and all $f \in \mathscr{A}$.

Theorem B.25 (Vitali Convergence Theorem)**.** *Let D be an open, connected subset of $\mathbb{C}$ and $(f_n)_{n\ge1}$ be a sequence of analytic functions on D, which are locally uniformly bounded and converge on a set having a cluster point in D. Then the sequence $(f_n)_{n\ge1}$ converges locally uniformly on D to an analytic function.*

Proof. See [71, p. 154]. □

Theorem B.26 (Hurwitz Theorem)**.** *Let D be an open subset of $\mathbb{C}$ and $(f_n)_{n\ge1}$ be a sequence of analytic functions, which converge, locally uniformly, on D to an analytic function f. If $f_n(z) \neq 0$, for all $z \in D$ and for all n, then either f vanishes identically, or f is never zero on D.*

Proof. See [71, Corollary 2.6]. □

The following theorem is the complex counterpart to Theorem B.7. (Notice that, in the complex case, no control is needed on the series of the derivatives.)

Theorem B.27 (Weierstrass' Theorem on uniformly convergent series of analytic functions)**.** *Let $D \subset \mathbb{C}$ be a domain. For each k, let $f_k : D \to \mathbb{C}$ be an analytic function. If the series*

$$f(z) \stackrel{\text{def}}{=} \sum_k f_k(z)$$

is uniformly convergent on every compact subset $K \subset D$, then it defines an analytic function on D. Moreover, for each $n \in \mathbb{N}$, $\sum_k f_k^{(n)}$ converges uniformly on every compact $K \subset D$ and

$$f^{(n)}(z) = \sum_k f_k^{(n)}(z), \quad \forall z \in D.$$

Proof. See [228, Volume 1, Theorem 15.6]. □

Let $U, V \subset \mathbb{C}$. A continuous function $F : U \times V \to \mathbb{C}$ is said to be **analytic** on $U \times V$ if $F(\cdot, z)$ is analytic on U for any fixed $z \in V$ and $F(z, \cdot)$ is analytic on V for any fixed $z \in U$.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Theorem B.28 (Implicit function theorem)**.** *Let $(\omega, z) \mapsto F(\omega, z)$ be an analytic function on an open domain $U \times V \subset \mathbb{C}^2$. Let $(\omega_0, z_0) \in U \times V$ be such that $F(\omega_0, z_0) = 0$ and $\frac{\partial F}{\partial z}(\omega_0, z_0) \neq 0$. Then there exists an open subset $U_0 \subset U$ containing ω_0 and an analytic map $\varphi : U_0 \to V$ such that*

$$F(\omega, \varphi(\omega)) = 0 \quad \text{for all } \omega \in U_0 .$$

Proof. See [228, Volume 2, Theorem 3.11]. □

B.4 Metric spaces

All topological notions used in the book (in particular those of Chapter 6) concern topologies induced by a *metric*. Let χ be an arbitrary set. A map $d : \chi \times \chi \to \mathbb{R}_{\geq 0}$ is a **metric (on χ)** (or **distance**) if it satisfies: (i) $d(x, y) = 0$ if and only if $x = y$, (ii) $d(x, y) = d(y, x)$ for all $x, y \in \chi$, (iii) $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in \chi$. The pair (χ, d) is then called a **metric space**.

The **open ball centered at $x \in \chi$ of radius $\epsilon > 0$** is $B_\epsilon(x) \stackrel{\text{def}}{=} \{y \in \chi : d(y, x) < \epsilon\}$. A set $A \subset \chi$ is **open** if, for each $x \in A$, there exists $\epsilon > 0$ such that $B_\epsilon(r) \subset A$. A set A is **closed** if $A^c \stackrel{\text{def}}{=} \chi \setminus A$ is open. Arbitrary unions and finite intersections of open sets are open. A sequence $(x_n)_{n\geq 1} \subset \chi$ **converges** to $x_* \in \chi$ (denoted $x_n \to x_*$) if, for all $\epsilon > 0$, there exists n_0 such that $x_n \in B_\epsilon(x_*)$ for all $n \geq n_0$. A set $F \subset \chi$ is closed if and only if $(x_n)_{n\geq 1} \subset F$, $x_n \to x_*$ implies $x_* \in F$. A set $D \subset \chi$ is **dense** if, for all $x \in \chi$ and all $\epsilon > 0$, $D \cap B_\epsilon(x) \neq \varnothing$. χ is **separable** if there exists a countable dense subset $D \subset \chi$.

On $\chi = \mathbb{R}^n$, one usually uses the *Euclidean metric* inherited from the Euclidean norm: $d(x, y) \stackrel{\text{def}}{=} \|x - y\|_2$; on $\chi = \mathbb{C}$, one uses the modulus: $d(w, z) \stackrel{\text{def}}{=} |w - z|$.

A function $f : \chi \to \chi'$ is **continuous** if $f(x_n) \to f(x_*)$ whenever $x_n \to x_*$. Equivalently, f is continuous if and only if $f^{-1}(A') \subset \chi$ is open for each open set $A' \subset \chi'$.

A metric space (χ, d) is **sequentially compact** (or simply **compact**) if there exists, for each sequence $(x_n)_{n\geq 1} \subset \chi$, a subsequence $(x_{n_k})_{k\geq 1}$ and some $x_* \in \chi$ such that $x_{n_k} \to x_*$ when $k \to \infty$. A compact metric space is always separable.

An introduction to metric spaces can be found in [284, Chapter 1].

B.5 Measure Theory

This section and the two following ones contain several definitions and results concerning measure theory and integration. Many detailed books exist on the subject, among which the one by Bogachev [30].

B.5.1 Measures and probability measures

Throughout this section, Ω denotes an arbitrary set and $\mathscr{P}(\Omega)$ the collection of all subsets of Ω. The complement of a set $A \subset \Omega$ will be denoted $A^c \stackrel{\text{def}}{=} \Omega \setminus A$.

Definition B.29. *A collection $\mathscr{A} \subset \mathscr{P}(\Omega)$ is an **algebra** if (i) $\varnothing \in \mathscr{A}$, (ii) $A \in \mathscr{A}$ implies $A^c \in \mathscr{A}$, and (iii) $A, B \in \mathscr{A}$ implies $A \cup B \in \mathscr{A}$.*

Definition B.30. *A collection $\mathscr{F} \subset \mathscr{P}(\Omega)$ is a **σ-algebra** if (i) $\varnothing \in \mathscr{A}$, (ii) $A \in \mathscr{A}$ implies $A^c \in \mathscr{F}$, and (iii) $(A_n)_{n\geq 1} \subset \mathscr{F}$ implies $\bigcup_{n\geq 1} A_n \in \mathscr{F}$.*

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Given an arbitrary collection $\mathscr{S} \subset \mathscr{P}(\Omega)$ of subsets of Ω, there exists a smallest σ-algebra containing $\mathscr{S}$, called the ***σ*-algebra generated by** $\mathscr{S}$, denoted $\sigma(\mathscr{S})$ and given by

$$\sigma(\mathscr{S}) \stackrel{\text{def}}{=} \bigcap \{\mathscr{F} : \mathscr{F} \text{ a } \sigma\text{-algebra containing } \mathscr{S}\}\,.$$

(Note that the intersection of an arbitrary collection of σ-algebras is a σ-algebra.)

Example B.31. If (χ, d) is a metric space whose collection of open sets is denoted by $\mathscr{O}$, then $\mathscr{B} \stackrel{\text{def}}{=} \sigma(\mathscr{O})$ is called the ***σ*-algebra of Borel sets on** χ. When (χ, d) is the Euclidean space $\mathbb{R}^n$ (equipped with the Euclidean metric), this σ-algebra is denoted $\mathscr{B}(\mathbb{R}^n)$. ◇

A pair $(\Omega, \mathscr{F})$, where $\mathscr{F}$ is a σ-algebra of subsets of Ω, is called a **measurable space** and the sets $A \in \mathscr{F}$ are called **measurable**.

Definition B.32. *On a measurable space $(\Omega, \mathscr{F})$, a set function $\mu : \mathscr{F} \to [0, +\infty]$ is called a **measure** if the following holds:*

1. $\mu(\varnothing) = 0$.
2. *(**σ-additivity**) If $(A_n)_{n\geq 1} \subset \mathscr{F}$ is a sequence of pairwise disjoint sets, then $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$.*

The measure μ is **finite** if $\mu(\Omega) < \infty$; μ is a **probability measure** if $\mu(\Omega) = 1$. If there exists a sequence $(A_n)_{n\geq 1} \subset \mathscr{F}$ such that $\bigcup_{n\geq 1} A_n = \Omega$ and $\mu(A_n) < \infty$ for each n, then μ is ***σ*-finite**.

Let us remind the reader of two straightforward consequences of the above definition. First, by the σ-additivity of item 2 above,

$$\mu\Big(\bigcup_n A_n\Big) \leq \sum_n \mu(A_n)\,,$$

for any sequence $A_n \in \mathscr{F}$. In particular, if $\mu(A_n) = 0$ for all n, then

$$\mu\Big(\bigcup_n A_n\Big) = 0\,.$$

A property A, defined for each element $\omega \in \Omega$, occurs ***μ*-almost everywhere** (or **for *μ*-almost all *ω***) if there exists $B \in \mathscr{F}$ such that $\{\omega \in \Omega : A \text{ does not hold for } \omega\} \subset B$ and $\mu(B) = 0$. When μ is a probability measure, one usually says ***μ*-almost surely**.

Measures are usually constructed by defining a *finitely additive* set function on an algebra $\mathscr{A}$ and by extending it to the σ-algebra generated by $\mathscr{A}$.

Let $\mathscr{A}$ be an algebra. A set function $\mu_0 : \mathscr{A} \to [0, +\infty]$ is said to be **finitely additive** if $\mu_0(A \cup B) = \mu_0(A) + \mu_0(B)$ for all pairs of disjoint measurable sets; μ_0 is a **measure** if $\mu_0(\varnothing) = 0$ and if $\mu_0(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mu_0(A_n)$ holds for all sequences $(A_n)_{n\geq 1} \subset \mathscr{A}$ of pairwise disjoint sets for which $\bigcup_{n\geq 1} A_n \in \mathscr{A}$.

Theorem B.33 (Carathéodory's Extension Theorem)**.** *Let $\mu_0 : \mathscr{A} \to [0, +\infty]$ be a σ-finite measure on an algebra $\mathscr{A}$ and let $\mathscr{F} \stackrel{\text{def}}{=} \sigma(\mathscr{A})$. Then there exists a unique measure $\mu : \mathscr{F} \to [0, +\infty]$, called the **extension of μ_0**, which coincides with μ_0 on $\mathscr{A}$: $\mu(A) = \mu_0(A)$ for all $A \in \mathscr{A}$.*

The σ-algebra $\mathscr{F} = \sigma(\mathscr{A})$ is in general a much larger collection of sets than $\mathscr{A}$; nevertheless, each set $B \in \mathscr{F}$ can be approximated arbitrary well by sets in $A \in \mathscr{A}$ in the sense of measure theory:

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Lemma B.34. *Let μ be a probability measure on $(\Omega, \mathscr{F})$, where $\mathscr{F}$ is generated by an algebra $\mathscr{A}$: $\mathscr{F} = \sigma(\mathscr{A})$. Then, for all $B \in \mathscr{F}$ and all $\epsilon > 0$, there exists $A \in \mathscr{A}$ such that $\mu(B \triangle A) \leq \epsilon$.*

Proof. Let $\mathscr{G} \stackrel{\text{def}}{=} \{B \in \mathscr{F} : \forall \epsilon > 0, \exists A \in \mathscr{A} \text{ s.t. } \mu(B \triangle A) \leq \epsilon\}$. Since, obviously, $\mathscr{G} \supset \mathscr{A}$, it suffices to show that $\mathscr{G}$ is a σ-algebra. Since $\mu(B \triangle A) = \mu(B^c \triangle A^c)$, we see that $\mathscr{G}$ is stable under taking complements. Let $(B_n)_{n\geq 1} \subset \mathscr{G}$ and set $B = \bigcup_{n\geq 1} B_n$. Fix $\epsilon > 0$. For each n, let $A_n \in \mathscr{A}$ be such that $\mu(B_n \triangle A_n) \leq \epsilon/2^n$. Then, let $A = \bigcup_{n=1}^N A_n \in \mathscr{A}$. If N is large enough,

$$\mu(B \triangle A) \leq \sum_{n\geq 1} \mu(B_n \triangle A_n) \leq \epsilon.$$

Therefore, $B \in \mathscr{G}$. This shows that $\mathscr{G}$ is a σ-algebra. □

In measure theory, it is often useful to determine whether some property is verified by each measurable set of a σ-algebra $\mathscr{F}$. If $\mathscr{F}$ is generated by an algebra, then this can be done by checking conditions which are easier to verify than testing *each* $B \in \mathscr{F}$.

A collection $\mathscr{M} \subset \mathscr{P}(\Omega)$ is a **monotone class** if (i) $\Omega \in \mathscr{M}$, (ii) for any sequence $(A_n)_{n\geq 1} \subset \mathscr{M}$ such that $A_n \uparrow A$, one has $A \in \mathscr{M}$, and (iii) for any sequence $(A_n)_{n\geq 1} \subset \mathscr{M}$ such that $A_n \downarrow A$, one has $A \in \mathscr{M}$. As before, there always exists a smallest monotone class generated by a collection $\mathscr{S}$, denoted $\mathscr{M}(\mathscr{S})$.

Theorem B.35. *If $\mathscr{A}$ is an algebra, then $\mathscr{M}(\mathscr{A}) = \sigma(\mathscr{A})$.*

A similar result holds for a slightly different notion of class. A collection $\mathscr{D} \subset \mathscr{P}(\Omega)$ is a **Dynkin system** if (i) $\varnothing \in \mathscr{D}$, (ii) $A, B \in \mathscr{D}$ with $A \subset B$ implies $B \setminus A \in \mathscr{D}$, and (iii) for any sequence $(A_n)_{n\geq 1} \subset \mathscr{D}$ such that $A_n \uparrow A$, one has $A \in \mathscr{D}$. Again, there always exists a smallest Dynkin system generated by a collection $\mathscr{S}$, denoted $\delta(\mathscr{S})$. A collection $\mathscr{C} \subset \mathscr{P}(\Omega)$ is **$\cap$-stable** if $A, B \in \mathscr{C}$ implies $A \cap B \in \mathscr{C}$.

Theorem B.36. *If $\mathscr{C}$ is $\cap$-stable (in particular, if $\mathscr{C}$ is an algebra), then $\delta(\mathscr{C}) = \sigma(\mathscr{C})$.*

This result can be used to determine when two measures are identical.

Corollary B.37. *Let $(\Omega, \mathscr{F})$ be a measurable space. Let $\mathscr{C}$ be a collection of sets which is $\cap$-stable and which generates $\mathscr{F}$: $\mathscr{F} = \sigma(\mathscr{C})$. If μ and ν are two probability measures on $(\Omega, \mathscr{F})$ which coincide on $\mathscr{C}$ ($\mu(C) = \nu(C)$ for all $C \in \mathscr{C}$), then $\mu = \nu$.*

B.5.2 Measurable functions

Let $(\Omega, \mathscr{F})$ and $(\Omega', \mathscr{F}')$ be two measurable spaces. A map $f : \Omega \to \Omega'$ is **$\mathscr{F}/\mathscr{F}'$-measurable** if $f^{-1}(B') \in \mathscr{F}$ for each $B' \in \mathscr{F}'$.

Let $(\Omega', \mathscr{F}')$ be a measurable space. For any set Ω, given an arbitrary map $h : \Omega \to \Omega'$, we denote by $\sigma(h)$ the smallest σ-algebra on Ω with respect to which h is $\mathscr{F}/\mathscr{F}'$-measurable: $\sigma(h) \stackrel{\text{def}}{=} \{h^{-1}(B') : B' \in \mathscr{F}'\}$. $\sigma(h)$ is called the σ-algebra **generated by h**.

Lemma B.38 (Doob–Dynkin lemma)**.** *Let $(\Omega, \mathscr{F})$, $(\Omega', \mathscr{F}')$ be measurable spaces, where $\mathscr{F} = \sigma(h)$ for some $h : \Omega \to \Omega'$. For any $\mathscr{F}/\mathscr{B}(\mathbb{R})$-measurable map $g : \Omega \to \mathbb{R}$, there exists an $\mathscr{F}'/\mathscr{B}(\mathbb{R})$-measurable map $\varphi : \Omega' \to \mathbb{R}$ such that $g = \varphi \circ h$.*

Proof. See [186, Lemma 1.13]. □

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

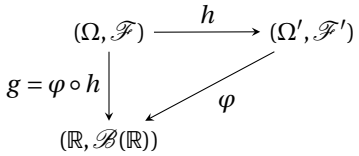


Figure B.3: The setting of Lemma B.38.

B.6 Integration

Let $\mathscr{F} \subset \mathscr{P}(\Omega)$ be a σ-algebra. When integrating real-valued functions, it is convenient to include the possibility of these taking the values $\pm\infty$. Let therefore $\overline{\mathbb{R}} \stackrel{\text{def}}{=} \mathbb{R} \cup \{\pm\infty\}$, together with the σ-algebra $\mathscr{B}(\overline{\mathbb{R}})$ containing sets of the form B, $B \cup \{+\infty\}$, $B \cup \{-\infty\}$, or $B \cup \{+\infty\} \cup \{-\infty\}$, where $B \in \mathscr{B}(\mathbb{R})$. An $\mathscr{F}/\mathscr{B}(\overline{\mathbb{R}})$-measurable function $f : \Omega \to \overline{\mathbb{R}}$ will simply be called **measurable**: $f^{-1}(I) \in \mathscr{F}$ for all $I \in \mathscr{B}(\overline{\mathbb{R}})$. To be measurable, f needs only satisfy $f^{-1}(\{\pm\infty\}) \in \mathscr{F}$ and $f^{-1}((-\infty, x]) \in \mathscr{F}$ for all $x \in \mathbb{R}$.

Integration is first defined for non-negative functions. A measurable function $\varphi : \Omega \to \mathbb{R} \cup \{+\infty\}$ is **simple** if it takes a finite set of values; it can therefore be written as a finite linear combination

$$\varphi = \sum_{k=1}^{n} a_k 1_{E_k},$$

where $E_k = \{\omega \in \Omega : \varphi(\omega) = a_k\} \in \mathscr{F}$, where $\varphi(\mathbb{R}) = \{a_1, \dots, a_n\} \subset \mathbb{R} \cup \{+\infty\}$. A measurable map $f : \Omega \to [0, +\infty]$ can always be written as a limit of an increasing sequence of simple functions $\varphi_n \uparrow f$. The **integral of φ with respect to μ** is

$$\int \varphi \, \mathrm{d}\mu \stackrel{\text{def}}{=} \sum_{k=1}^{n} a_k \mu(E_k).$$

In this definition, we make the convention that $0 \cdot \infty = 0$. If $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ is measurable and nonnegative, its **integral with respect to μ** is

$$\int f \, \mathrm{d}\mu \stackrel{\text{def}}{=} \sup\Big\{ \int \varphi \, \mathrm{d}\mu : \varphi \text{ simple}, 0 \le \varphi \le f \Big\}.$$

For an arbitrary measurable function f, let $f^+ \stackrel{\text{def}}{=} f \mathbf{1}_{\{f \ge 0\}}$, $f^- \stackrel{\text{def}}{=} (-f)^+$. We say that f is **integrable** if $\int f^+ \, \mathrm{d}\mu < \infty$ and $\int f^- \, \mathrm{d}\mu < \infty$. The set of integrable function is denoted by $L^1(\mu)$ (we sometimes omit the measure when it is clear from the context). The **integral** of $f \in L^1(\mu)$ is

$$\int f \, \mathrm{d}\mu \stackrel{\text{def}}{=} \int f^+ \, \mathrm{d}\mu - \int f^- \, \mathrm{d}\mu.$$

In this book, we also use alternative notations for $\int f \, \mathrm{d}\mu$, such as $\mu(f)$ or $\langle f \rangle_\mu$. We list below a few properties of the integral.

- If $f \ge 0$ and $\int f \, \mathrm{d}\mu < \infty$, then f is μ-almost everywhere finite.
- If $f \ge 0$ and $\int f \, \mathrm{d}\mu = 0$, then $f = 0$ μ-almost everywhere.
- If $f, g \in L^1(\mu)$, $\int (f + g) \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\mu$.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

- $f \in L^1(\mu)$ if and only if $\int |f| \, \mathrm{d}\mu < \infty$, and when this occurs, $|\int f \, \mathrm{d}\mu| \le \int |f| \, \mathrm{d}\mu$.
- If $0 \le f \le g$ μ-almost everywhere, then $\int f \, \mathrm{d}\mu \le \int g \, \mathrm{d}\mu$.
- If $f, g \in L^1(\mu)$, $f = g$ μ-almost everywhere, then $\int f \, \mathrm{d}\mu = \int g \, \mathrm{d}\mu$.

Theorem B.39 (Monotone Convergence Theorem)**.** *Let $(f_n)_{n\ge1}$ be a sequence of nonnegative measurable functions such that $f_n \le f_{n+1}$ μ-almost everywhere. Then*

$$\int \lim_{n\to\infty} f_n \, \mathrm{d}\mu = \lim_{n\to\infty} \int f_n \, \mathrm{d}\mu .$$

Theorem B.40 (Dominated Convergence Theorem)**.** *Let $(f_n)_{n\ge1} \subset L^1(\mu)$. Assume there exists $g \in L^1(\mu)$ such that $|f_n| \le g$ μ-almost everywhere, for all $n \ge 1$. If $f_n \to f$ μ-almost everywhere, then $f \in L^1(\mu)$ and*

$$\int f \, \mathrm{d}\mu = \lim_{n\to\infty} \int f_n \, \mathrm{d}\mu .$$

Exercise B.15. *Let $(\xi_k)_{k\ge1}$ be a sequence of real functions defined on some open interval $I \subset \mathbb{R}$ and $x_0 \in I$. Let that sequence be such that, for each k, $\lim_{x\to x_0} \xi_k(x)$ exists. Assuming there exists a summable sequence $(\epsilon_k)_{k\ge1} \subset \mathbb{R}_{\ge0}$ such that $\sup_{x\in I} |\xi_k(x)| \le \epsilon_k$, show that*

$$\lim_{x\to x_0} \sum_k \xi_k(x) = \sum_k \lim_{x\to x_0} \xi_k(x) . \tag{B.14}$$

Let μ, ν be two finite measures. ν is **absolutely continuous with respect to μ** if, for all $A \in \mathscr{F}$, $\mu(A) = 0$ implies $\nu(A) = 0$; we then write $\nu \ll \mu$. When both $\mu \ll \nu$ and $\nu \ll \mu$, the measures are said to be **equivalent**. If there exists $A \in \mathscr{F}$ such that $\mu(A) = 0$, $\nu(A^{\mathrm{c}}) = 0$, then μ and ν are **singular**.

Theorem B.41 (Radon–Nikodým's theorem)**.** *Let μ and ν be two finite measures such that $\nu \ll \mu$. There exists a measurable function $f \ge 0$ such that*

$$\forall B \in \mathscr{F}, \quad \nu(B) = \int_B f \, \mathrm{d}\mu .$$

*f is called the **Radon–Nikodým derivative of ν with respect to μ** and is often denoted $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$.*

Any two versions of the Radon–Nikodým derivative coincide μ-almost everywhere; this is a consequence of the following lemma.

Lemma B.42. *Let $f, g \in L^1(\mu)$ be such that $\int_B f \, \mathrm{d}\mu = \int_B g \, \mathrm{d}\mu$ for all $B \in \mathscr{F}$. Then $f = g$ almost everywhere.*

The Radon–Nikodým derivative enjoys properties similar to that of the ordinary derivative. First, if $\nu_1, \nu_2 \ll \mu$, then $\nu_1 + \nu_2 \ll \mu$ and

$$\frac{\mathrm{d}(\nu_1 + \nu_2)}{\mathrm{d}\mu} = \frac{\mathrm{d}\nu_1}{\mathrm{d}\mu} + \frac{\mathrm{d}\nu_2}{\mathrm{d}\mu} . \tag{B.15}$$

Then, a property similar to the chain rule holds: if ν, μ, ρ satisfy $\nu \ll \mu \ll \rho$, then

$$\frac{\mathrm{d}\nu}{\mathrm{d}\rho} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \frac{\mathrm{d}\mu}{\mathrm{d}\rho} . \tag{B.16}$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

B.6.1 Product spaces

Given two measurable spaces $(\Omega, \mathscr{F})$, $(\Omega', \mathscr{F}')$, we can consider the product

$$\Omega \times \Omega' \stackrel{\text{def}}{=} \{(\omega, \omega') : \omega \in \Omega, \omega' \in \Omega'\},$$

equipped with the **product σ-algebra** $\mathscr{F} \otimes \mathscr{F}'$, generated by the algebra of finite unions of **rectangles**, that is, sets of the form $A \times A'$ with $A \in \mathscr{F}$ and $A \in \mathscr{F}'$. If μ is a measure on $(\Omega, \mathscr{F})$ and μ' is a measure on $(\Omega', \mathscr{F}')$, we can define, for a rectangle,

$$(\mu \otimes \mu')(A \times A') \stackrel{\text{def}}{=} \mu(A)\mu'(A').$$

Using Theorem B.33, it can be shown that, when μ and ν are σ-finite, $\mu \otimes \mu'$ has a unique extension to $\mathscr{F} \otimes \mathscr{F}'$; we call it the **product measure**.

Theorem B.43 (Theorem of Fubini–Tonelli)**.** *If μ and μ' are σ-finite and if $F : \Omega \times \Omega' \to \mathbb{R}_{\geq 0}$ is $\mathscr{F} \otimes \mathscr{F}'$-measurable, then the functions*

$$\omega \mapsto \int_{\Omega'} F(\omega, \omega')\mu'(\mathrm{d}\omega') \quad \textit{and} \quad \omega' \mapsto \int_{\Omega} F(\omega, \omega')\mu(\mathrm{d}\omega)$$

are $\mathscr{F}$- and $\mathscr{F}'$-measurable, respectively. Moreover,

$$\begin{aligned}\int_{\Omega\times\Omega'} F \,\mathrm{d}(\mu \otimes \mu') &= \int_{\Omega} \Big\{ \int_{\Omega'} F(\omega, \omega')\, \mu'(\mathrm{d}\omega') \Big\} \mu(\mathrm{d}\omega) \\ &= \int_{\Omega'} \Big\{ \int_{\Omega} F(\omega, \omega')\, \mu(\mathrm{d}\omega) \Big\} \mu'(\mathrm{d}\omega')\end{aligned}$$

The above construction extends to the product of an arbitrary finite number of σ-finite measurable spaces: $(\Omega_1, \mathscr{F}_1), \ldots, (\Omega_n, \mathscr{F}_n)$.

B.7 Lebesgue measure

The Lebesgue measure is first constructed on the real line, by extending to all Borel sets the basic notion of *length* of bounded intervals:

$$\ell([a, b)) \stackrel{\text{def}}{=} b - a.$$

(Unbounded intervals are defined to have measure $+\infty$.) This allows to define a natural measure on the algebra of finite unions of such intervals, that can be extended to all Borel sets $\mathscr{B}(\mathbb{R})$ using Theorem B.33. The resulting σ-finite measure ℓ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is called the **Lebesgue measure**.

On $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$, equipped with the Borel σ-algebra $\mathscr{B}(\mathbb{R}^n)$, the Lebesgue measure is defined as the product measure, that is, it is first defined on parallelepipeds

$$\ell^n\big([a_1, b_1) \times [a_2, b_2) \times \cdots \times [a_n, b_n)\big) \stackrel{\text{def}}{=} \prod_{i=1}^{n} (b_i - a_i),$$

and then extended. The Lebesgue measure is **translation invariant**, $\ell^n(B + \mathbf{x}) = \ell^n(B)$ for all $B \in \mathscr{B}(\mathbb{R}^n)$, $\mathbf{x} \in \mathbb{R}^n$, and enjoys the following **scaling property**: $\ell^n(\alpha B) = \alpha^n \ell^n(B)$ for all $B \in \mathscr{B}(\mathbb{R}^n)$ and all scaling factor $\alpha > 0$.

One usually writes $\mathrm{d}\mathbf{x}$ instead of $\ell^n(\mathrm{d}\mathbf{x})$. For instance, the integration of a function $f : \mathbb{R}^n \to \mathbb{R}$ with respect to ℓ^n is written $\int f(\mathbf{x})\, \mathrm{d}\mathbf{x}$.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

B.8 Probability

We remind the reader of some basic elements from Probability Theory. The books by Kallenberg [186] or Grimmett and Stirzaker [152] provide good references.

In probability theory, a **probability space** is a triple $(\Omega, \mathscr{F}, P)$, where $(\Omega, \mathscr{F})$ is a measurable space and $P : \mathscr{F} \to [0,1]$ is a probability measure. Each $\omega \in \Omega$ is to be interpreted as the outcome of a random experiment and each measurable set $A \in \mathscr{F}$ is interpreted as an **event**, with $P(A)$ measuring the a priori likeliness of the occurence of A when sampling some $\omega \in \Omega$.

B.8.1 Random variables and vectors

A measurable map $X : \Omega \to \overline{\mathbb{R}}$ is called a **random variable**. The **distribution of** $X : \Omega \to \overline{\mathbb{R}}$ is the probability measure P_X on $\overline{\mathbb{R}}$ defined by $P_X(I) \stackrel{\text{def}}{=} P(X \in I)$, for all $I \in \mathscr{B}(\overline{\mathbb{R}})$. The **cumulative distribution function of** $X : \Omega \to \overline{\mathbb{R}}$ is $F_X(x) \stackrel{\text{def}}{=} P(X \le x)$, $x \in \mathbb{R}$. X has a **density (with respect to the Lebesgue measure)** if there exists a measurable function $f_X : \mathbb{R} \to \mathbb{R}_{\ge 0}$ such that

$$P(X \in B) = \int_B f_X(x)\,\mathrm{d}x, \quad \forall B \in \mathscr{B}(\mathbb{R}).$$

The integral of a random variable $X \in L^1(P)$ is denoted

$$E[X] \stackrel{\text{def}}{=} \int X\,\mathrm{d}P,$$

and is called the **expectation** of X with respect to P. The **variance** of X is then defined by

$$\mathrm{Var}(X) \stackrel{\text{def}}{=} E\big[(X - E[X])^2\big].$$

We list here a few inequalities that are used frequently:

- **Jensen's inequality**: If $X \in L^1(P)$ and if $\phi : \mathbb{R} \to \mathbb{R}$ is convex and such that $\phi(X) \in L^1(P)$, then $\phi(E[X]) \le E[\phi(X)]$. When ϕ is strictly convex, equality holds if and only if X is almost surely constant.
- **Markov's inequality**: for all non-negative $X \in L^1(P)$ and all $\lambda > 0$,

$$P(X \ge \lambda) \le \frac{E[X]}{\lambda}. \tag{B.17}$$

- **Chebyshev's inequality**: for all X and all $\lambda > 0$,

$$P(|X - E[X]| \ge \lambda) \le \frac{\mathrm{Var}(X)}{\lambda^2}. \tag{B.18}$$

- **Chernov's inequality**: for all X and all $\lambda > 0$,

$$P(X \ge \lambda) \le \inf_{t>0} \frac{E[e^{tX}]}{e^{t\lambda}}. \tag{B.19}$$

There are various ways by which a sequence of random variables $(X_n)_{n\ge 1}$ can *converge* to a limiting random variable X.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

- **$(X_n)_{n\geq 1}$ converges to X almost surely** if there exists $C \in \mathscr{F}$, $P(C) = 1$, such that $X_n(\omega) \to X(\omega)$ for all $\omega \in C$.

- **$(X_n)_{n\geq 1}$ converges to X in probability** if, for all $\epsilon > 0$, $P(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$.

- Let $p \geq 1$. **$(X_n)_{n\geq 1}$ converges to X in L^p** if $E[|X|^p] < \infty$, $E[|X_n^p|] < \infty$, for all n, and $E[|X_n - X|^p] \to 0$ when $n \to \infty$.

- **$(X_n)_{n\geq 1}$ converges to X in distribution** if $F_{X_n}(x) \to F_X(x)$ when $n \to \infty$, for all x at which F_X is continuous.

Almost sure convergence and convergence in L^p both imply convergence in probability, which in turn implies convergence in distribution. The remaining implications do not hold in general.

B.8.2 Independence

Two events $A, B \in \mathscr{F}$ are **independent** if $P(A \cap B) = P(A)P(B)$. A collection of events $(A_i)_{i\in I}$ is **independent** if $P(\bigcap_{j\in J} A_j) = \prod_{j\in J} P(A_j)$ for all $J \subset I$ finite.

A collection of random variables $(X_i)_{i\in I}$ is **independent** if the collection of events $(\{X_i \leq \alpha_i\})_{i\in I}$ is independent for all $(\alpha_i)_{i\in I} \subset \mathbb{R}$. If, moreover, all the variables X_i have the same distribution, we say that $(X_i)_{i\in I}$ is **i.i.d. (independent, identically distributed)**.

We now state two central results of Probability Theory.

Theorem B.44 (Law of Large Numbers)**.** *Let $(X_n)_{n\geq 1} \subset L^1(P)$ be an i.i.d. sequence. Then, as $n \to \infty$,*

$$\frac{X_1 + \cdots + X_n}{n} \to E[X_1] \quad P\text{-almost surely.}.$$

Remember that X is a **standard normal** random variable, $X \sim \mathscr{N}(0,1)$, if it has a density with respect to the Lebesgue measure $\mathrm{d}t$, given by $\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$; in particular, its cumulative distribution function is

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, \mathrm{d}t, \quad \forall x \in \mathbb{R}.$$

Theorem B.45 (Central Limit Theorem)**.** *Let $(X_n)_{n\geq 1}$ be an i.i.d. sequence with $m \stackrel{\text{def}}{=} E[X_1] < \infty$ and $\sigma^2 \stackrel{\text{def}}{=} \mathrm{Var}(X_1) < \infty$. Then, as $n \to \infty$,*

$$\frac{(X_1 - m) + \cdots + (X_n - m)}{\sigma\sqrt{n}} \to \mathscr{N}(0,1) \quad \textit{in distribution.}$$

In particular, for all $a < b$,

$$P\Big(a \leq \frac{(X_1 - m) + \cdots + (X_n - m)}{\sigma\sqrt{n}} \leq b\Big) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{t^2}{2}}\, \mathrm{d}t.$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

B.8.3 Moments and cumulants of random variables

Let X be a random variable. If $r \in \mathbb{N}$, the ***r*th moment** of X is defined by

$$m_r(X) \stackrel{\text{def}}{=} E[X^r],$$

provided the expectation exists. The **moment generating function** associated to X is the function

$$t \mapsto M_X(t) \stackrel{\text{def}}{=} E[e^{tX}], \quad t \in \mathbb{R}.$$

If $M_X(t)$ possesses a convergent MacLaurin expansion, then all its moments exist and can be recovered from the formula

$$m_r(X) = \frac{\mathrm{d}^r}{\mathrm{d}t^r} M_X(t)|_{t=0}.$$

Under suitable conditions, the moments $(m_r(X))_{r\geq 1}$ completely characterize the distribution of X.

Theorem B.46. *Assume that all moments $m_r(X)$, $r \geq 1$, exist. If there exists some $\epsilon > 0$ such that $\sum_{r\geq 1} \frac{1}{r!} m_r(X) t^r$ converges for all $t \in (-\epsilon, \epsilon)$, then $M_X(t) = \sum_{r\geq 1} \frac{1}{r!} m_r(X) t^r$ on that interval and any random variable Y with $m_r(Y) = m_r(X)$, for all $r \geq 1$, has the same distribution as X.*

Proof. See [186, Exercise 10, Chapter 5]. □

Let us now consider the **cumulant generating function** (also known as the **log-moment generating function**) $C_X(t) = \log M_X(t)$. The coefficients of its MacLaurin expansion (if it has one) are called the cumulants of the random variable X: for $r \in \mathbb{N}$, the ***r*th cumulant** of X is defined by

$$c_r(X) \stackrel{\text{def}}{=} \frac{\mathrm{d}^r}{\mathrm{d}t^r} C_X(t)|_{t=0}.$$

Cumulants possess a variety of other names, depending on the context. When $r \geq 2$, they are also called **semi-invariants**, thanks to the following remarkable property: for any $a, b \in \mathbb{R}$,

$$c_r(aX + b) = a^r c_r(X). \tag{B.20}$$

(This of course doesn't hold for $r = 1$, since $c_1(aX + b) = ac_1(X) + b$.) In statistical mechanics, cumulants are often called **Ursell functions**, **truncated correlation functions** or **connected correlation functions**.

Exercise B.16. *Show that cumulants can be expressed in terms of moments using the following recursion formula:*

$$c_r = m_r - \sum_{m=1}^{r-1} \binom{r-1}{m-1} c_m m_{r-m}.$$

In particular,

$$c_1 = m_1, \quad c_2 = m_2 - m_1^2, \quad c_3 = m_3 - 3m_2 m_1 + 2m_1^3, \quad \ldots$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

Cumulants of a random variable X characterize the distribution of X whenever its moments do. The advantage of cumulants compared to moments, in addition to their satisfying (B.20), is the way they act on sums of independent random variables: if X and Y are independent random variables, then

$$c_r(X+Y) = c_r(X) + c_r(Y)\,.$$

This follows immediately from the identity $C_{X+Y} = C_X + C_Y$.

B.8.4 Characteristic function

The **characteristic function** of a random variable X is defined by

$$\varphi_X(t) \stackrel{\text{def}}{=} E[e^{\mathrm{i}tX}] \stackrel{\text{def}}{=} E[\cos(tX)] + \mathrm{i}E[\sin(tX)]\,.$$

Note that, since

$$E[|e^{\mathrm{i}tX}|] = 1\,,$$

the characteristic function is well defined for all random variables. If there exists $\epsilon > 0$ such that the moment generating function $M_X(t)$ is finite for all $|t| < \epsilon$, then $\varphi_X(-\mathrm{i}t) = M_X(t)$.

Characteristic functions owe their name to the fact that they characterize the distribution of a random variable: $\varphi_X = \varphi_Y$ if and only if X and Y have the same distribution.

If $X_1, \ldots, X_n$ are independent random variables, then

$$\varphi_{X_1+\cdots+X_n}(s) = \varphi_{X_1}(s)\cdots\varphi_{X_n}(s)\,.$$

Theorem B.47 (Lévy's continuity theorem)**.** *Let X_n be a sequence of random variables. Assume that $\varphi(t) \stackrel{\text{def}}{=} \lim_{n\to\infty}\varphi_{X_n}(t)$ exists, for all $t \in \mathbb{R}$, and that φ is continuous at $t = 0$. Then there exists a random variable X such that $\varphi_X = \varphi$ and X_n converges to X in distribution.*

B.8.5 Conditional Expectation

Conditional expectation is a fundamental concept in probability theory and plays a central role in our study of infinite-volume Gibbs measures in Chapter 6. Before giving its formal definition, we motivate it starting from the simplest possible case.

In elementary probability, the conditional probability of an event A with respect to an event B with $P(B) > 0$ is defined by

$$P(A\,|\,B) \stackrel{\text{def}}{=} \frac{P(A\cap B)}{P(B)}\,.$$

This defines a new probability measure $P(\cdot\,|\,B)$ under which random variables can be integrated, yielding a **conditional expectation given B**: for $X \in L^1(P)$,

$$E[X\,|\,B] \stackrel{\text{def}}{=} \int X(\omega)P(\mathrm{d}\omega\,|\,B)\,.$$

Often, one is more interested in considering the conditional expectation with respect to a *collection* of events, associated to some *partial information* in a random experiment.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Example B.48. Consider an experiment in which two dice are rolled, modeled by two independent random variables X_1, X_2 on a probability space Ω, taking values in $\{1,2,\dots,6\}$. Assume that some partial information is given about the outcome of the sum $S = X_1 + X_2$, namely whether $S > 5$ or $S \le 5$. Given this partial information, the expectation of X_1 is $E[X_1 \,|\, S > 5]$ if $\{S > 5\}$ occurred and $E[X_1 \,|\, S \le 5]$ if $\{S \le 5\}$ occurred. It thus appears natural to encode this information in a *random variable*

$$\omega \mapsto E[X_1 \,|\, S > 5]\mathbf{1}_{\{S>5\}}(\omega) + E[X_1 \,|\, S \le 5]\mathbf{1}_{\{S\le 5\}}(\omega)\,. \qquad \diamond$$

This example leads to a first generalization of conditional expectation, as follows. Let $(B_k)_k \subset \mathscr{F}$ be a countable partition of Ω: $B_k \cap B_{k'} = \varnothing$ if $k \neq k'$ and $\bigcup_k B_k = \Omega$. This means that, for each outcome ω of the experiment, exactly one event B_k occurs. For convenience, let $\mathscr{B} \subset \mathscr{F}$ denote the sub-σ-algebra containing the events which are unions of sets B_k. The occurrence of some $B \in \mathscr{B}$ provides some information on the occurrence of some events B_k.

Now if we also assume that $P(B_k) > 0$ for all k, we can define, for $X \in L^1(P)$,

$$E[X \,|\, \mathscr{B}](\omega) \stackrel{\text{def}}{=} \sum_k E[X \,|\, B_k]\mathbf{1}_{B_k}(\omega)\,.$$

Exercise B.17. *Show that, as a random variable on Ω, $E[X \,|\, \mathscr{B}]$ satisfies the following properties:*

$$\omega \mapsto E[X \,|\, \mathscr{B}](\omega) \qquad \text{is } \mathscr{B}\text{-measurable,} \tag{B.21}$$

$$E\big[E[X \,|\, \mathscr{B}]\mathbf{1}_B\big] = E[X\mathbf{1}_B] \qquad \text{for all } B \in \mathscr{B}\,. \tag{B.22}$$

In particular,

$$E\big[E[X \,|\, \mathscr{B}]\big] = E[X]\,.$$

The above definition, although natural, is not yet suited to our needs, its main defect being the necessity to assume that $P(B_k) > 0$. Indeed, the theory of infinite-volume Gibbs measures, exposed in Chapter 6, requires conditioning on a fixed configuration outside a finite region, an event that always has zero probability. We therefore need a definition of conditional expectation which allows to condition with respect to events of zero probability.

It turns out that (B.21)-(B.22) characterize $E[X \,|\, \mathscr{B}]$ in an essentially unique manner. This can be used to define conditional expectation in much greater generality:

Lemma B.49. *Let $(\Omega, \mathscr{F}, P)$ be a probability space. Consider $X \in L^1(P)$ and a sub-σ-algebra $\mathscr{G} \subset \mathscr{F}$. There exists a random variable $Y \in L^1(P)$ for which the following conditions hold:*

1. *Y is $\mathscr{G}$-measurable.*
2. *For all $G \in \mathscr{G}$, $E[Y\mathbf{1}_G] = E[X\mathbf{1}_G]$.*

If Y' is another variable satisfying these properties, then $P(Y \neq Y') = 0$. Any of them is called a ***version of the conditional expectation of X with respect to $\mathscr{G}$*** *and is denoted by $E[X \,|\, \mathscr{G}]$.*

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

We list the main properties of conditional expectation. In view of the almost-sure uniqueness, all the properties are to be understood as holding almost surely. All the random variables below are assumed to be integrable.

1. $E[a_1X_1 + a_2X_2 \,|\, \mathscr{G}] = a_1E[X_1 \,|\, \mathscr{G}] + a_2E[X_2 \,|\, \mathscr{G}]$.
2. If $X \le X'$, then $E[X \,|\, \mathscr{G}] \le E[X' \,|\, \mathscr{G}]$.
3. $|E[X \,|\, \mathscr{G}]| \le E[|X| \,|\, \mathscr{G}]$.
4. (**Tower property**) If $\mathscr{G} \subset \mathscr{H}$, then
$$E\big[E[X \,|\, \mathscr{G}] \,\big|\, \mathscr{H}\big] = E[X \,|\, \mathscr{G}] = E\big[E[X \,|\, \mathscr{H}] \,\big|\, \mathscr{G}\big].$$
5. If Z is $\mathscr{G}$-measurable, then $E[XZ \,|\, \mathscr{G}] = ZE[X \,|\, \mathscr{G}]$.

Conditional expectation can be characterized equivalently in the following way:

Lemma B.50. *Let $X \in L^1(P)$, $\mathscr{G} \subset \mathscr{F}$ a sub-σ-algebra. Then $E[X \,|\, \mathscr{G}]$ is the (almost sure) unique $\mathscr{G}$-measurable random variable with the property that*
$$E\big[(X - E[X \,|\, \mathscr{G}])Z\big] = 0 \quad \textit{for all } \mathscr{G}\textit{-measurable } Z \in L^1(P). \tag{B.23}$$

Remark B.51. The above definition provides a nice geometrical interpretation of the conditional expectation of a random variable with finite variance. Let us denote by $L^2(P)$ the (real) vector space of all random variables such that $E[X^2] < \infty$ (or, equivalently, with finite variance). The space $L^2(P)$ is a Hilbert space for the inner product $(X, Y) \mapsto E[XY]$. (B.23) can then be interpreted as stating that the vector $X - E[X \,|\, \mathscr{G}]$ is orthogonal to the linear subspace $\big\{Z \in L^2(P) : Z \text{ is } \mathscr{G}\text{-measurable}\big\}$. This implies that $E[X \,|\, \mathscr{G}]$ coincides with the orthogonal projection of X on this subspace; see Figure B.4. ◇

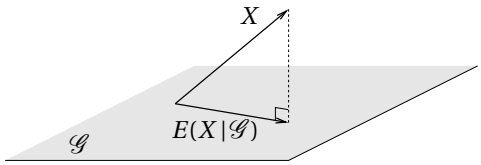


Figure B.4: Restricted to random variables with finite variance, the conditional expectation $E(X \,|\, \mathscr{G})$ corresponds to the orthogonal projection of X onto the linear subspace of all $\mathscr{G}$-measurable random variables.

Finally, we will occasionally need the following classical result whose proof can be found in [351].

Theorem B.52 (Backward martingale convergence)**.** *Let $X \in L^1$ and let $\mathscr{G}_n$ be a decreasing sequence of σ-algebras, $\mathscr{G}_n \supset \mathscr{G}_{n+1}$, and set $\mathscr{G}_\infty \stackrel{\text{def}}{=} \bigcap_n \mathscr{G}_n$. Then,*
$$E[X \,|\, \mathscr{G}_n] \to E[X \,|\, \mathscr{G}_\infty] \quad \textit{in } L^1 \textit{ and almost surely.}$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

B.8.6 Conditional probability

Let $\mathscr{G} \subset \mathscr{F}$. The **conditional probability of $A \in \mathscr{F}$ with respect to $\mathscr{G}$** is defined by the (almost surely unique) random variable

$$P(A|\mathscr{G})(\omega) \stackrel{\text{def}}{=} E[\mathbf{1}_A|\mathscr{G}](\omega)\,.$$

By definition, $P(A|\mathscr{G})$ inherits many of the properties of the conditional expectation. In particular it is, up to almost-sure equivalence, the unique $\mathscr{G}$-measurable random variable for which

$$P(A\cap G) = \int_G P(A|\mathscr{G})\,\mathrm{d}P, \qquad \forall G \in \mathscr{G}\,.$$

Remember that, by linearity of the conditional expectation, one has, for disjoint events $A, B \in \mathscr{F}$, $P(A\cup B|\mathscr{G})(\omega) = P(A|\mathscr{G})(\omega) + P(B|\mathscr{G})(\omega)$. It is important to notice that even though this equality holds for P-almost all ω, the set of such ωs depends in general on A and B. Since there are usually uncountably many events in $\mathscr{F}$, one should therefore not expect, for a fixed ω, for $P(\cdot|\mathscr{G})(\omega)$ to define a probability measure on $(\Omega,\mathscr{F})$. This leads to the following definition. A map $\widehat{P}(\cdot|\mathscr{G})(\cdot) : \mathscr{F} \times \Omega \to [0,1]$ is called a **regular conditional probability with respect to** $\mathscr{G}$ if (i) for each $\omega \in \Omega$, $\widehat{P}(\cdot|\mathscr{G})(\omega)$ is a probability distribution on $(\Omega,\mathscr{F})$, (ii) for each $A \in \mathscr{F}$, $\widehat{P}(A|\mathscr{G})(\cdot)$ is a version of $P(A|\mathscr{G})$.

Regular conditional probabilities exist under fairly general assumptions, which can be found for example in [186].

The Gibbs measures constructed and studied in Chapter 6 are examples of regular conditional probabilities. Indeed, when $\mu \in \mathscr{G}(\pi)$ is conditioned with respect to the values taken by the spins outside a finite region Λ, the kernel $\pi_\Lambda(\cdot|\omega)$ is a version of $\mu(\cdot|\mathscr{F}_{\Lambda^c})(\omega)$. But by definition, $\pi_\Lambda(\cdot|\omega)$ is a probability measure for each $\omega \in \Omega$. See the comments of Section 6.3.1. ⋄

B.8.7 Random vectors

Most of what was said for random variables can be adapted to the case of measurable functions taking values in a space of larger dimension: $\mathbf{X} : \Omega \to \mathbb{R}^n$ is a **random vector** if it is $\mathscr{F}/\mathscr{B}(\mathbb{R}^n)$-measurable. The **distribution** of $\mathbf{X}$ is the probability measure $P_{\mathbf{X}}$ on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$ defined by $P_{\mathbf{X}}(B) \stackrel{\text{def}}{=} P(\mathbf{X} \in B)$, $B \in \mathscr{B}(\mathbb{R}^n)$. The expectation $E[\mathbf{X}]$ is to be understood as coordinate-wise integration. A random vector has a **density (with respect to the Lebesgue measure)** if there exists a measurable $f_{\mathbf{X}} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ such that

$$P(\mathbf{X} \in B) = \int_B f_{\mathbf{X}}(\mathbf{x})\,\mathrm{d}\mathbf{x} \quad \forall B \in \mathscr{B}(\mathbb{R}^n)\,.$$

Random variables $X_1, \ldots, X_n$ with density $f_{(X_1,\ldots,X_n)}$ are independent if and only if

$$f_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) = f_{X_1}(x_1)\cdots f_{X_n}(x_n)\,.$$

The different types of *convergence* defined earlier for random variables have direct analogues for random vectors. Moreover, an equivalent version of Theorem B.47 holds.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

B.9 Gaussian vectors and fields

In this section, we recall some basic definitions and properties related to Gaussian fields. A good reference is the first chapter of Le Gall's book [212].

B.9.1 Basic definitions and properties

We already defined a normal $\mathcal{N}(0,1)$ earlier. More generally, given $m \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_{\geq 0}$, a random variable X is called a **Gaussian with mean m and variance σ^2**, $X \sim \mathcal{N}(m,\sigma^2)$, if it admits the density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/2\sigma^2}.$$

As is easily verified, $E[X] = m$ and $\mathrm{Var}(X) = \sigma^2$.

Exercise B.18. *Show that $X \sim \mathcal{N}(m,\sigma^2)$ if and only if its characteristic function is given by*

$$E[e^{\mathrm{i}tX}] = \exp\bigl(-\tfrac{1}{2}\sigma^2 t^2 + \mathrm{i}mt\bigr).$$

Exercise B.19. *Let $X_1,\dots,X_n$ be independent Gaussian random variables with $X_i \sim \mathcal{N}(m_i,\sigma_i^2)$ and let $t_1,\dots,t_n \in \mathbb{R}$. Show that*

$$\sum_{i=1}^n t_i X_i \sim \mathcal{N}(t_1 m_1 + \cdots + t_n m_n, t_1^2\sigma_1^2 + \cdots + t_n^2\sigma_n^2).$$

Let us now introduce $\mathbb{R}^n$-valued *Gaussian vectors*. As before, elements of $\mathbb{R}^n$ will be denoted using bold letters: $\mathbf{x},\mathbf{y},\dots$ and the scalar product will be denoted $\mathbf{x}\cdot\mathbf{y}$.

Definition B.53. *A random vector $\mathbf{X}:\Omega\to\mathbb{R}^n$ is **Gaussian** if the random variable $\mathbf{t}\cdot\mathbf{X}$ is Gaussian for each $\mathbf{t}\in\mathbb{R}^n$.*

By Exercise B.18, this is equivalent to requiring that, for all $\mathbf{t}\in\mathbb{R}^n$,

$$E[e^{\mathrm{i}\mathbf{t}\cdot\mathbf{X}}] = \exp\bigl(-\tfrac{1}{2}\mathrm{Var}(\mathbf{t}\cdot\mathbf{X}) + \mathrm{i}E[\mathbf{t}\cdot\mathbf{X}]\bigr) = \exp\bigl(-\tfrac{1}{2}\mathbf{t}\cdot\Sigma\mathbf{t} + \mathrm{i}\,\mathbf{m}\cdot\mathbf{t}\bigr),$$

where $\mathbf{m} = (m_1,\dots,m_n)$ with $m_i = E[X_i]$ and Σ is the $n\times n$ matrix with elements $\Sigma(i,j) = \mathrm{Cov}(X_i,X_j)$. $\mathbf{m}$ and Σ are called the **mean** and **covariance matrix** of $\mathbf{X}$. We write in this case $\mathbf{X}\sim\mathcal{N}(\mathbf{m},\Sigma)$.

Since $1 \geq |E[e^{\mathrm{i}\mathbf{t}\cdot\mathbf{X}}]| = \exp(-\frac{1}{2}\mathbf{t}\cdot\Sigma\mathbf{t})$, we have $\mathbf{t}\cdot\Sigma\mathbf{t}\geq 0$: the covariance matrix is nonnegative-definite.

Lemma B.54. *Let Σ be an $n\times n$ nonnegative-definite symmetric matrix. Then there exists an $n\times n$ matrix A such that $\Sigma = AA^\intercal$ (where $A^\intercal$ denotes the transpose of A). Moreover, if Σ is invertible, then so is A.*

Proof. Let us denote by $\lambda_1,\dots,\lambda_n$ the eigenvalues of Σ; observe that $\lambda_i\geq 0$, $i = 1,\dots,n$, since Σ is nonnegative-definite. Symmetry of Σ implies the existence of an orthogonal matrix O such that $\Sigma = O^\intercal DO$, where $D = \mathrm{diag}(\lambda_1,\dots,\lambda_n)$. Let $D^{1/2} \stackrel{\mathrm{def}}{=} \mathrm{diag}(\sqrt{\lambda_1},\dots,\sqrt{\lambda_n})$ and $A = O^\intercal D^{1/2}$. It then follows that $AA^\intercal = O^\intercal D^{1/2}D^{1/2}O = O^\intercal DO = \Sigma$, as required.

If Σ is invertible, then $\lambda_i > 0$, $i = 1,\dots,n$. This implies that $D^{1/2}$ is invertible. Since O is also invertible, it follows that so is A. □

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

Exercise B.20. *Show that the components of a Gaussian vector $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Sigma)$ are independent if and only if Σ is diagonal.*

Exercise B.21. *Show that $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Sigma)$ if and only if $\mathbf{X} = A\mathbf{Y} + \mathbf{m}$ with $\mathbf{Y}$ a random vector with independent $\mathcal{N}(0,1)$ components and A an $n \times n$ matrix such that $AA^\intercal = \Sigma$.*

Proposition B.55. *Let Σ be a positive-definite symmetric $n \times n$ matrix and $\mathbf{m} \in \mathbb{R}^n$. Then $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Sigma)$ if and only if it possesses the following density with respect to the Lebesgue measure $\mathrm{d}\mathbf{x}$:*

$$\mathbf{x} \mapsto \frac{1}{(2\pi)^{n/2}\sqrt{|\det \Sigma|}} \exp\bigl(-\tfrac{1}{2}(\mathbf{x}-\mathbf{m})\cdot\Sigma^{-1}(\mathbf{x}-\mathbf{m})\bigr).$$

Proof. Using Exercise B.21, $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \Sigma)$ if and only $\mathbf{X} = A\mathbf{Y} + \mathbf{m}$, with $\Sigma = AA^\intercal$ and $\mathbf{Y}$ a Gaussian random vector with i.i.d. $\mathcal{N}(0,1)$ components. The density of $\mathbf{Y}$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}} \exp\bigl(-\tfrac{1}{2}\|\mathbf{y}\|_2^2\bigr).$$

Note that, by Lemma B.54, A is invertible. The claim therefore follows from the change of variable formula, $f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{Y}}(A^{-1}(\mathbf{x}-\mathbf{m}))\,|\det \Sigma|^{-1/2}$, where we have used the fact that the absolute value of the Jacobian of the transformation is equal to $|\det(A^{-1})| = |\det A|^{-1} = |\det \Sigma|^{-1/2}$. □

Exercise B.22. *Use the method exposed in the previous proof to prove* (8.61).

B.9.2 Convergence of Gaussian vectors

The following result shows that limits of convergent sequences of Gaussian random vectors are themselves Gaussian.

Proposition B.56. *Let $(\mathbf{X}^{(k)})_{k\geq 1}$ be a sequence of Gaussian random vectors, with mean $\mathbf{m}^{(k)}$ and covariance matrix $\Sigma^{(k)}$. Then $\mathbf{X}^{(k)}$ converges to a random vector $\mathbf{X}$ in distribution if and only if the limits $\mathbf{m} = \lim_{k\to\infty} \mathbf{m}^{(k)}$ and $\Sigma = \lim_{k\to\infty} \Sigma^{(k)}$ both exist. In that case, $\mathbf{X}$ is also a Gaussian vector, with mean $\mathbf{m}$ and covariance matrix Σ.*

Proof. Assume that $\mathbf{m} = \lim_{k\to\infty} \mathbf{m}^{(k)}$ and $\Sigma = \lim_{k\to\infty} \Sigma^{(k)}$ exist. Then,

$$\lim_{k\to\infty} E[e^{\mathrm{i}\mathbf{t}\cdot\mathbf{X}^{(k)}}] = \lim_{k\to\infty} \exp\bigl(-\tfrac{1}{2}\mathbf{t}\cdot\Sigma^{(k)}\mathbf{t} + \mathrm{i}\mathbf{m}^{(k)}\cdot\mathbf{t}\bigr) = \exp\bigl(-\tfrac{1}{2}\mathbf{t}\cdot\Sigma\mathbf{t} + \mathrm{i}\mathbf{m}\cdot\mathbf{t}\bigr)$$

exists and is continuous at $\mathbf{t} = 0$. It thus follows from Levy's continuity theorem (n-dimensional version of Theorem B.47) that the sequence $(\mathbf{X}^{(k)})_{k\geq 1}$ converges in distribution and that the limit is a Gaussian random vector with mean $\mathbf{m}$ and covariance matrix Σ.

Assume now that $\mathbf{X}^{(k)} \to \mathbf{X}$ in distribution. The characteristic function of $\mathbf{X}$ satisfies, for any $\mathbf{t} \in \mathbb{R}^n$,

$$E[e^{\mathrm{i}\mathbf{t}\cdot\mathbf{X}}] = \lim_{k\to\infty} E[e^{\mathrm{i}\mathbf{t}\cdot\mathbf{X}^{(k)}}] = \lim_{k\to\infty} \exp\bigl(-\tfrac{1}{2}\mathbf{t}\cdot\Sigma^{(k)}\mathbf{t} + \mathrm{i}\mathbf{m}^{(k)}\cdot\mathbf{t}\bigr).$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

In particular, choosing $\mathbf{t} = t\mathbf{e}_i$, this yields

$$\lim_{k\to\infty} \exp\bigl(-\tfrac{1}{2}t^2\Sigma^{(k)}(i,i)\bigr) = |E[e^{\mathrm{i}t\mathbf{X}\cdot\mathbf{e}_i}]| \le 1\,.$$

This implies that the limits $\lim_{k\to\infty}\Sigma^{(k)}(i,i)$ $(i=1,\dots,n)$ exist in $[0,\infty]$; moreover, the value $+\infty$ can be excluded, since it would contradict the continuity at $\mathbf{t}=0$ of the characteristic function of $\mathbf{X}$.

Similarly, letting $\mathbf{t} = t(\mathbf{e}_i+\mathbf{e}_j)$, we obtain the existence of $\lim_{k\to\infty}\Sigma^{(k)}(i,j)$ for all $i\neq j$. This in turn implies the existence and continuity of

$$\lim_{k\to\infty} \exp\bigl(\mathrm{i}\mathbf{m}^{(k)}\cdot\mathbf{t}\bigr) = \exp\bigl(-\tfrac{1}{2}\mathbf{t}\cdot\Sigma\mathbf{t}\bigr)E[e^{\mathrm{i}\mathbf{t}\cdot\mathbf{X}}]\,.$$

Consequently, $\lim_{k\to\infty}\mathbf{m}^{(k)}$ also exists. This proves the claim. □

Definition B.57. *A collection of random variables $\varphi = (\varphi_i)_{i\in S}$ indexed by a countable set S is **a Gaussian random field** (or simply **Gaussian field**) if all its finite-dimensional distributions are Gaussian, that is, if*

$$E[e^{\mathrm{i}\sum_{i\in S}t_i\varphi_i}] = \exp\Bigl(-\tfrac{1}{2}\sum_{i,j\in S}t_i\,t_j\,\mathrm{Cov}(\varphi_i,\varphi_j) + \mathrm{i}\sum_{i\in S}t_iE[\varphi_i]\Bigr)\,,$$

for all $(t_i)_{i\in S}$ taking only finitely many nonzero values.

As follows from the definition, Proposition B.56 and the Kolmogorov extension theorem, a sequence of Gaussian random fields $\varphi^{(k)}$ on S converges to a random field φ on S if and only if the limits

$$\lim_{k\to\infty}E[\varphi_i^{(k)}], \qquad \lim_{k\to\infty}\mathrm{Cov}(\varphi_i^{(k)},\varphi_j^{(k)})$$

exist for all $i,j\in S$. Moreover, in that case, $(\varphi_i)_{i\in S}$ is Gaussian with

$$E[\varphi_i] = \lim_{k\to\infty}E[\varphi_i^{(k)}], \qquad \mathrm{Cov}(\varphi_i,\varphi_j) = \lim_{k\to\infty}\mathrm{Cov}(\varphi_i^{(k)},\varphi_j^{(k)})\,,$$

for all $i,j\in S$.

B.9.3 Gaussian fields and independence

For $T\subset S$, let $\mathscr{F}_T \stackrel{\mathrm{def}}{=} \sigma(\varphi_j,\, j\in T)$ (defined as the smallest σ-algebra on Ω such that each φ_j, $j\in T$, is measurable).

Proposition B.58. *Let $\varphi=(\varphi_i)_{i\in S}$ be a Gaussian field and $T\subset S$. Then $\mathscr{F}_T$ and $\mathscr{F}_{S\setminus T}$ are independent if and only if $\mathrm{Cov}(\varphi_i,\varphi_j)=0$ for all $i\in T$, $j\in S\setminus T$.*

Proof. See [212, Section 1.3]. □

B.10 The total variation distance

There are various ways by which one can measure the similarity of two probability measures. The simplest is the total variation distance.

Definition B.59. *The **total variation distance** between two probability measures μ and ν on $(\Omega, \mathscr{F})$ is defined by*

$$\|\mu - \nu\|_{TV} \stackrel{\text{def}}{=} 2 \sup_{A \in \mathscr{F}} |\mu(A) - \nu(A)| .$$

We warn the reader that some authors define $\|\mu - \nu\|_{TV}$ without the factor 2.

Lemma B.60. *Let μ and ν be two probability measures on $(\Omega, \mathscr{F})$, with $\mu \ll \nu$. Then,*

$$\|\mu - \nu\|_{TV} = \left\langle \left| 1 - \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right| \right\rangle_\nu = \sup_{f : \|f\|_\infty \le 1} \left| \langle f \rangle_\mu - \langle f \rangle_\nu \right| .$$

Proof. If $\rho = \mathrm{d}\mu / \mathrm{d}\nu$, then

$$\mu(A) - \nu(A) = \int_A (\rho - 1) \mathrm{d}\nu \le \int (\rho - 1)_+ \mathrm{d}\nu .$$

Since the inequality is saturated for $A = \{\rho \ge 1\}$, we get

$$\sup_{A \in \mathscr{F}} \{\mu(A) - \nu(A)\} = \int (\rho - 1)_+ \mathrm{d}\nu .$$

In the same way,

$$\sup_{A \in \mathscr{F}} \{\nu(A) - \mu(A)\} = \int (\rho - 1)_- \mathrm{d}\nu .$$

But since $\int (\rho - 1)_+ \mathrm{d}\nu - \int (\rho - 1)_- \mathrm{d}\nu = \int (\rho - 1) \mathrm{d}\nu = 0$, this gives

$$\sup_{A \in \mathscr{F}} \{\mu(A) - \nu(A)\} = \sup_{A \in \mathscr{F}} \{\nu(A) - \mu(A)\} = \sup_{A \in \mathscr{F}} |\mu(A) - \nu(A)| .$$

We conclude that

$$\int |\rho - 1| \mathrm{d}\nu = \int (\rho - 1)_+ \mathrm{d}\nu + \int (\rho - 1)_- \mathrm{d}\nu = 2 \sup_{A \in \mathscr{F}} |\mu(A) - \nu(A)| ,$$

which proves the first identity. The second is a consequence of the first:

$$\sup_{f : \|f\|_\infty \le 1} \left| \int f \mathrm{d}\mu - \int f \mathrm{d}\nu \right| = \sup_{f : \|f\|_\infty \le 1} \left| \int f (\rho - 1) \mathrm{d}\nu \right| = \int |\rho - 1| \mathrm{d}\nu = \|\mu - \nu\|_{TV} ,$$

the supremum being achieved by the function $\mathbf{1}_{\{\rho \ge 1\}} - \mathbf{1}_{\{\rho < 1\}}$. □

B.11 Shannon's Entropy

Shannon's Entropy $\mathsf{S}_{\mathrm{Sh}}(\cdot)$ is the central object for the implementation of the Maximum Entropy Principle, which was used in Chapter 1 to motivate the Gibbs distribution. In this section, we show that $\mathsf{S}_{\mathrm{Sh}}(\cdot)$ is unique, up to a multiplicative constant, among a class of functions $\mathsf{S} : \mathscr{M}_1(\Omega) \to \mathbb{R}$ satisfying a certain set of conditions, one of which being to be maximal for the uniform distribution. We follow the approach of Khinchin [191].

Consider a random experiment modeled by some probability space $(\Omega, \mathscr{F}, P)$. Consider a partition A of Ω into a finite number of events, called **atoms**. When A is

a partition with k atoms, we will write $\mathsf{A} = \{A_1, \ldots, A_k\}$. For convenience, we allow some atoms to be empty.

We should consider such a partition as corresponding to some partial information about the outcome $\omega \in \Omega$ of the experiment. For example, when throwing a dice, $\mathsf{A} = \{A_1, A_2\}$, where $A_1 = \{\text{the outcome is even}\}$ $A_2 = \{\text{the outcome is odd}\}$.

Our aim is to define the *unpredictability* of the outcome of the measurement, corresponding to a partition A. Since the probability P is fixed, the unpredictability associated to the partition $\mathsf{A} = \{A_1, \ldots, A_k\}$ will be defined through a function $\mathsf{S}(P(A_1), \ldots, P(A_k))$, usually denoted simply $\mathsf{S}(\mathsf{A})$ or $\mathsf{S}(A_1, \ldots, A_k)$ and called a function of the partition A. Notice that, since k is arbitrary, we are actually looking for a *collection* of functions.

Below, we define four conditions, most of which will be natural in terms of unpredictability, and then show that the only function that satisfies these conditions is, up to a positive multiplicative constant, the Shannon Entropy.

The first three assumptions are natural. First, for all partitions $\mathsf{A} = \{A_1, \cdots, A_k\}$,

$$\mathsf{S}(A_1, \cdots, A_k) \quad \text{is continuous in } (P(A_1), \ldots, P(A_k)). \tag{U1}$$

Second, we assume that unpredictability is not sensitive to the presence of atoms that have zero probability; for a partition $\mathsf{A} = \{A_1, \ldots, A_k\}$,

$$P(A_k) = 0 \quad \text{implies} \quad \mathsf{S}(A_1, \cdots, A_{k-1}, A_k) = \mathsf{S}(A_1, \ldots, A_{k-2}, A_{k-1} \cup A_k). \tag{U2}$$

Third, as discussed in Section 1.2.2, we want the unpredictability to be maximal for partitions whose atoms have equal probabilities. Namely, call a partition $\mathsf{U} = \{U_1, \ldots, U_m\}$ **uniform** if $P(U_i) = P(U_j) = \frac{1}{m}$ for all i, j.

Among partitions with m atoms, S is maximal for the uniform partitions. (U3)

To motivate the fourth assumption, we introduce some more terminology. A partition A is **finer** than a partition B if each atom of B is a union of atoms of A. When realizing the random experiment, if A is finer than B, information about the outcome $\omega \in \Omega$ can be revealed in two stages: first, by revealing the atom B_j such that $B_j \ni \omega$, and then, given B_j, one reveals the atom A_i such that $B_j \supset A_i \ni \omega$.

The unpredictability associated to the first stage is measured by $\mathsf{S}(\mathsf{B})$. After observing the result of the first stage, one should update our probability measure: assuming that the atom B_j occurred in the first stage, the relevant probability measure is $P(\cdot \,|\, B_j)$. The unpredictability of the second stage is thus measured by

$$\mathsf{S}(\mathsf{A} \,|\, B_j) \stackrel{\text{def}}{=} \mathsf{S}\big(P(A_1 \,|\, B_j), \ldots, P(A_n \,|\, B_j)\big).$$

Averaging over the possible outcomes B_j of the first stage, we are led to define the entropy of the second stage by

$$\mathsf{S}(\mathsf{A} \,|\, \mathsf{B}) \stackrel{\text{def}}{=} \sum_j P(B_j) \mathsf{S}(\mathsf{A} \,|\, B_j).$$

Now, the unpredictability of the complete experience should not depend on the way the experiment was conducted (in one stage or in two stages). It is therefore natural to assume that

$$\mathsf{S}(\mathsf{A}) = \mathsf{S}(\mathsf{B}) + \mathsf{S}(\mathsf{A} \,|\, \mathsf{B}). \tag{U4}$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Lemma B.61. *The **Shannon entropy**, defined by*

$$\mathsf{S}_{\mathrm{Sh}}(\mathsf{B}) \stackrel{\mathrm{def}}{=} -\sum_{j=1}^{k} P(B_j)\log P(B_j) \tag{B.24}$$

for all partitions $\mathsf{B} = \{B_1, \ldots, B_k\}$*, satisfies* (U1)-(U4).

Proof. (U1) and (U2) are clearly satisfied, (U3) has been shown in Lemma 1.9, and (U4) can be verified by a straightforward computation. □

Theorem B.62. *Let* $\mathsf{S}(\cdot)$ *be a function on finite partitions satisfying* (U1)–(U4)*. Then there exists a constant* $\lambda > 0$ *such that*

$$\mathsf{S}(\cdot) = \lambda \mathsf{S}_{\mathrm{Sh}}(\cdot)\,.$$

We express (U4) in a slightly different way, better suited for the computations to come. For two arbitrary partitions A, B, consider the **composite** partition

$$\mathsf{A} \vee \mathsf{B} \stackrel{\mathrm{def}}{=} \{A \cap B : A \in \mathsf{A}, B \in \mathsf{B}\}\,.$$

Then, (U2) implies that $\mathsf{S}(\mathsf{A} \vee \mathsf{B} \,|\, \mathsf{B}) = \mathsf{S}(\mathsf{A} \,|\, \mathsf{B})$ and (U4) can be used in the following form:

$$\mathsf{S}(\mathsf{A} \vee \mathsf{B}) = \mathsf{S}(\mathsf{B}) + \mathsf{S}(\mathsf{A} \,|\, \mathsf{B})\,. \tag{U4}$$

Notice that if $P(A \cap B) = P(A)P(B)$ for all $A \in \mathsf{A}$ and all $B \in \mathsf{B}$, then $\mathsf{S}(\mathsf{A} \,|\, \mathsf{B}) = \mathsf{S}(\mathsf{A})$ and (U4) implies

$$\mathsf{S}(\mathsf{A} \vee \mathsf{B}) = \mathsf{S}(\mathsf{A}) + \mathsf{S}(\mathsf{B})\,. \tag{B.25}$$

We will first start by proving a version of Theorem B.62 for uniform partitions U. Below, $|\mathsf{U}|$ denotes the number of atoms in U.

Proposition B.63. *Let* $\mathsf{S}(\cdot)$ *be a function defined on uniform partitions, which is monotone increasing in* $|\mathsf{U}|$ *and which is additive in the sense that if* $P(U \cap U') = P(U)P(U')$ *for all* $U \in \mathsf{U}$ *and all* $U' \in \mathsf{U}'$*. Then,*

$$\mathsf{S}(\mathsf{U} \vee \mathsf{U}') = \mathsf{S}(\mathsf{U}) + \mathsf{S}(\mathsf{U}')\,. \tag{B.26}$$

Then there exists $\lambda > 0$ *such that* $\mathsf{S}(\mathsf{U}) = \lambda \log|\mathsf{U}|$ *for all uniform partitions* U*.*

Proof. Since $\mathsf{S}(\cdot)$ is constant on partitions with the same number of atoms, we define $L(k) \stackrel{\mathrm{def}}{=} \mathsf{S}(\mathsf{U})$ if $|\mathsf{U}| = k$. By the assumption, $L(k)$ is increasing in k. Let then $\mathsf{U}^1, \ldots, \mathsf{U}^n$ be independent partitions, each containing k atoms. On the one hand, $\mathsf{U}^1 \vee \cdots \vee \mathsf{U}^n$ is also a uniform partition that contains k^n atoms and, therefore, $\mathsf{S}(\mathsf{U}^1 \vee \cdots \vee \mathsf{U}^n) = L(k^n)$. On the other hand, by (B.26),

$$\mathsf{S}(\mathsf{U}^1 \vee \cdots \vee \mathsf{U}^n) = \sum_{j=1}^{n} \mathsf{S}(\mathsf{U}^j) = n\,L(k)\,,$$

and so $L(k^n) = nL(k)$. We verify that $L(\cdot)$ is necessarily of the form $L(k) = \lambda \log k$, for some $\lambda > 0$. Namely, fix two arbitrary integers $k, \ell \geq 2$. Choose some large integer $m \geq 1$ and find some integer n so that $k^n \leq \ell^m < k^{n+1}$. On the one hand, $n \log k \leq m \log \ell < (n+1)\log k$. On the other hand, the monotonicity of $L(\cdot)$ implies that

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

$nL(k) = L(k^n) \le L(\ell^m) = mL(\ell)$ and, similarly, $mL(\ell) \le (n+1)L(k)$, which with the previous set of inequalities gives

$$\Big|\frac{L(\ell)}{L(k)} - \frac{\log \ell}{\log k}\Big| \le \frac{1}{m}\,.$$

Since m was arbitrary, this shows that $L(k)/\log k$ does not depend on k and must be equal to a constant. □

Proof of Theorem B.62: Assume $\mathsf{S}(\cdot)$ satisfies (U1)–(U4). Let us temporarily denote by $\mathsf{S}_k(\cdot)$ the function $\mathsf{S}(\cdot)$ when restricted to partitions with k atoms. Using (U2), followed by (U3),

$$\mathsf{S}_k\big(\tfrac{1}{k},\dots,\tfrac{1}{k}\big) = \mathsf{S}_{k+1}\big(\tfrac{1}{k},\dots,\tfrac{1}{k},0\big) \le \mathsf{S}_{k+1}\big(\tfrac{1}{k+1},\dots,\tfrac{1}{k+1}\big)\,.$$

Note that (B.25) guarantees (B.26). This shows that, when restricted to uniform partitions, $\mathsf{S}(\cdot)$ satisfies the hypotheses of Proposition B.63, yielding the existence of a constant $\lambda > 0$ such that $\mathsf{S}(\mathsf{U}) = \lambda \log|\mathsf{U}|$ for all uniform partitions U.

Let us now consider an arbitrary partition $\mathsf{B} = \{B_1,\dots,B_k\}$. By (U1), we can safely assume that the probabilities $P(B_j) \in \mathbb{Q}$. If we consider a collection of integers $w_1,\dots,w_k$ such that $P(B_j) = \frac{w_j}{Z}$, where $Z = w_1 + \cdots + w_k$, the partition B can be reinterpreted as follows. Consider a collection of Z labeled balls, each of a specific color, among k different colors. Assume that there are exactly w_j balls of color j, $j = 1,\dots,k$. A ball is sampled at random, uniformly. Then clearly, the color of the ball sampled has color j with probability $\frac{w_j}{Z} = P(B_j)$. We therefore reinterpret B_j as the event "the sampled ball has color j" and use this to compute $\mathsf{S}(\mathsf{B})$.

In this same experiment, consider now the partition $\mathsf{A} = \{A_1,\dots,A_Z\}$ defined by $A_i = \{\text{the ball } i \text{ was sampled}\}$.

Since A is finer than B we have $\mathsf{A} \vee \mathsf{B} = \mathsf{A}$ and, since A is uniform, $\mathsf{S}(\mathsf{B} \vee \mathsf{A}) = \mathsf{S}(\mathsf{A}) = \lambda \log Z$.

Now, observe that

$$P(A_i \,|\, B_j) = \begin{cases} \frac{1}{w_j} & \text{if } i \in B_j\,, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, using (U2), $\mathsf{S}(\mathsf{A}\,|\,B_j) = \lambda \log w_j = \lambda \log P(B_j) + \lambda \log Z$, and so

$$\mathsf{S}(\mathsf{A}\,|\,\mathsf{B}) = \lambda \sum_{j=1}^{k} P(B_j) \log P(B_j) + \lambda \log Z\,.$$

This proves the claim, since assumption (U4) implies

$$\mathsf{S}(\mathsf{B}) = \mathsf{S}(\mathsf{A}) - \mathsf{S}(\mathsf{A}\,|\,\mathsf{B}) = -\lambda \sum_{j=1}^{k} P(B_j) \log P(B_j)\,. \qquad \square$$

B.12 Relative entropy

B.12.1 Definition, basic properties

We have seen that, when μ, ν are two probability measures such that $\mu \ll \nu$, then there exists a nonnegative measurable function $\mathrm{d}\mu/\mathrm{d}\nu$, the Radon–Nikodým derivative of μ with respect to ν, such that $\mu(A) = \int_A \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\mathrm{d}\nu$ for all $A \in \mathscr{F}$.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Definition B.64. *The **relative entropy** $\mathsf{h}(\mu\,|\,\nu)$ of μ with respect to ν is defined as*

$$\mathsf{h}(\mu\,|\,\nu) \stackrel{\text{def}}{=} \begin{cases} \langle \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \rangle_\nu, & \text{if } \mu \ll \nu, \\ \infty & \text{otherwise.} \end{cases}$$

Since $x \log x \geq -e^{-1}$ on $\mathbb{R}_{>0}$, $\mathsf{h}(\mu\,|\,\nu)$ is always well defined (but can be equal to $+\infty$).

Lemma B.65. *$\mathsf{h}(\mu\,|\,\nu) \geq 0$, with equality if and only if $\mu = \nu$.*

Proof. We can assume that $\mathsf{h}(\mu\,|\,\nu) < \infty$. Since $\Psi(x) = x \log x$ is strictly convex on $(0,\infty)$, Jensen's inequality implies that

$$\mathsf{h}(\mu\,|\,\nu) = \big\langle \Psi(\tfrac{\mathrm{d}\mu}{\mathrm{d}\nu}) \big\rangle_\nu \geq \Psi\big(\langle \tfrac{\mathrm{d}\mu}{\mathrm{d}\nu} \rangle_\nu\big) = \Psi(1) = 0\,.$$

Moreover, Jensen's inequality is an equality if and only if $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ is almost surely a constant, and the latter can only be 1. □

Proposition B.66. *1. $(\mu,\nu) \mapsto \mathsf{h}(\mu\,|\,\nu)$ is convex.*

2. $\mu \mapsto \mathsf{h}(\mu\,|\,\nu)$ is strictly convex.

To prove this proposition, we will need the following elementary inequality.

Exercise B.23. *Let a_i, b_i, $i = 1,\dots,n$, be nonnegative real numbers. Set $A \stackrel{\text{def}}{=} \sum_{i=1}^n a_i$ and $B \stackrel{\text{def}}{=} \sum_{i=1}^n b_i$. Then,*

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq A \log \frac{A}{B}\,,$$

with equality if and only if there exists λ such that $a_i = \lambda b_i$ for all $1 \leq i \leq n$. Hint: *use Lemma B.65.*

Proof of Proposition B.66. 1. Let $\alpha \in (0,1)$ and take four probability measures μ_1, μ_2, ν_1, ν_2. Set $\mu \stackrel{\text{def}}{=} \alpha\mu_1 + (1-\alpha)\mu_2$ and $\nu \stackrel{\text{def}}{=} \alpha\nu_1 + (1-\alpha)\nu_2$. We need to prove that

$$\mathsf{h}(\mu\,|\,\nu) \leq \alpha \mathsf{h}(\mu_1\,|\,\nu_1) + (1-\alpha)\mathsf{h}(\mu_2\,|\,\nu_2)\,. \tag{B.27}$$

We can assume that $\mu_i \ll \nu_i$, $i = 1,2$, so that the right-hand side is finite and the following Radon–Nikodým derivatives are well defined: for $i = 1,2$,

$$f_i \stackrel{\text{def}}{=} \frac{\mathrm{d}\mu_i}{\mathrm{d}\nu}\,, \quad g_i \stackrel{\text{def}}{=} \frac{\mathrm{d}\nu_i}{\mathrm{d}\nu}\,, \quad h_i \stackrel{\text{def}}{=} \frac{\mathrm{d}\mu_i}{\mathrm{d}\nu_i}\,, \quad \phi \stackrel{\text{def}}{=} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,.$$

With these notations, (B.27) can be rewritten, thanks to (B.16),

$$\begin{aligned} \langle \phi \log \phi \rangle_\nu &\leq \alpha \langle h_1 \log h_1 \rangle_{\nu_1} + (1-\alpha)\langle h_2 \log h_2 \rangle_{\nu_2} \\ &= \Big\langle \alpha f_1 \log \frac{f_1}{g_1} + (1-\alpha) f_2 \log \frac{f_2}{g_2} \Big\rangle_\nu\,. \end{aligned} \tag{B.28}$$

By (B.15), we have $\alpha f_1 + (1-\alpha) f_2 = \phi$ and $\alpha g_1 + (1-\alpha) g_2 = 1$, so Exercise B.23 implies that

$$\alpha f_1 \log \frac{f_1}{g_1} + (1-\alpha) f_2 \log \frac{f_2}{g_2} = \alpha f_1 \log \frac{\alpha f_1}{\alpha g_1} + (1-\alpha) f_2 \log \frac{(1-\alpha) f_2}{(1-\alpha) g_2} \geq \phi \log \phi\,,$$

pointwise in Ω. Integrating this inequality with respect to ν yields (B.28).

2. This claim follows immediately from the corresponding properties of the function $x \mapsto x \log x$. □

B.12.2 Two useful inequalities

Pinsker's inequality

The relative entropy is a measure of the similarity of two measures μ and ν. However, it is not a metric, as it is not even symmetric in its two arguments. Actually, even its symmetrized version $\mathsf{h}(\nu\,|\,\mu)+\mathsf{h}(\mu\,|\,\nu)$ fails to be a metric, as it violates the triangle inequality. Nevertheless, smallness of the relative entropy between two measures allows one to control their total variation distance (see Section B.10).

Lemma B.67 (Pinsker's inequality)**.** *Let μ and ν be two probability measures on the same measurable space, with $\mu \ll \nu$. Then*

$$\|\mu-\nu\|_{TV} \le \sqrt{2\mathsf{h}(\mu\,|\,\nu)}\,. \tag{B.29}$$

Proof. Notice that, by applying Jensen's inequality,

$$\begin{aligned}(1+x)\log(1+x)-x &= x^2\int_0^1 \mathrm{d}t\int_0^t \mathrm{d}s\,\frac{1}{1+xs}\\ &\ge \tfrac{1}{2}x^2\,\frac{1}{1+x\int_0^1\mathrm{d}t\int_0^t\mathrm{d}s\,2s} = \frac{x^2}{2(1+\frac{x}{3})}\,.\end{aligned} \tag{B.30}$$

Let $m \stackrel{\text{def}}{=} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}-1$. Then $\langle m\rangle_\nu = 0$ and, using (B.30),

$$\mathsf{h}(\mu\,|\,\nu) = \langle (1+m)\log(1+m)\rangle_\nu = \langle (1+m)\log(1+m)-m\rangle_\nu \ge \Big\langle \frac{m^2}{2(1+\frac{m}{3})}\Big\rangle_\nu .$$

But, using Lemma B.60 and the Cauchy-Schwartz inequality,

$$\big(\|\mu-\nu\|_{TV}\big)^2 = \langle |m|\rangle_\nu^2 = \Big\langle \frac{|m|}{(1+\frac{m}{3})^{1/2}}\big(1+\tfrac{m}{3}\big)^{1/2}\Big\rangle_\nu^2 \le \Big\langle \frac{m^2}{1+\frac{m}{3}}\Big\rangle_\nu \big\langle 1+\tfrac{m}{3}\big\rangle_\nu .$$

Since $\big\langle 1+\frac{m}{3}\big\rangle_\nu = 1$, this proves (B.29). □

An exponential inequality

Pinsker's inequality (Lemma B.67) allows one to control the differences $|\mu(A)-\nu(A)|$ uniformly in $A\in\mathscr{F}$ in terms of the relative entropy between the two measures. Sometimes, however, we need to control *ratio* of such probabilities. The following result can then be useful.

Lemma B.68. *Let μ and ν be two equivalent probability measures on some measurable space $(\Omega,\mathscr{F})$. If $\nu(A)>0$, then*

$$\frac{\mu(A)}{\nu(A)} \ge \exp\left(-\frac{\mathsf{h}(\nu\,|\,\mu)+e^{-1}}{\nu(A)}\right).$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Proof. From Jensen's inequality and the inequality $x \log x \geq -e^{-1}$, which holds for all $x > 0$, we can write

$$\begin{aligned}\log \frac{\mu(A)}{\nu(A)} = \log \frac{\langle \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \mathbf{1}_A \rangle_\nu}{\langle \mathbf{1}_A \rangle_\nu} &= \log \langle \tfrac{\mathrm{d}\mu}{\mathrm{d}\nu} \,|\, A \rangle_\nu \\ &\geq \langle \log \tfrac{\mathrm{d}\mu}{\mathrm{d}\nu} \,|\, A \rangle_\nu = -\frac{\langle \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \mathbf{1}_A \rangle_\mu}{\nu(A)} \\ &\geq -\frac{\langle \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \rangle_\mu + e^{-1}}{\nu(A)} .\end{aligned}$$

□

B.13 The symmetric simple random walk on $\mathbb{Z}^d$

Good references for these topics are the books by Spitzer [319], Lawler [209] and Lawler and Limic [211].

Let $(\xi_n)_{n\geq 1}$ be an i.i.d. sequence of random vectors uniformly distributed in the set $\{j \in \mathbb{Z}^d : j \sim 0\}$. The **simple random walk on $\mathbb{Z}^d$ started at $i \in \mathbb{Z}^d$** is the random process $(X_n)_{n\geq 0}$ with $X_0 = i$ and defined by

$$X_n \stackrel{\text{def}}{=} i + \sum_{k=1}^{n} \xi_k .$$

We denote the distribution of this process by $\mathbb{P}_i$.

B.13.1 Stopping times and the strong Markov property

For each $n \geq 0$, we consider the σ-algebra $\mathscr{F}_n \stackrel{\text{def}}{=} \sigma(X_0, \ldots, X_n)$. A random variable T with values in $\mathbb{Z}_{\geq 0} \cup \{+\infty\}$ is a **stopping time** if $\{T \leq n\} \in \mathscr{F}_n$ for all n, that is, if the occurrence of the event $\{T \leq n\}$ can be decided by considering only the first n steps of the walk. Given a stopping time T, let $\mathscr{F}_T$ denote the σ-algebra containing all events A such that $A \cap \{T \leq n\} \in \mathscr{F}_n$ for all n. That is, $\mathscr{F}_T$ contains all events that depend only on the part of the trajectory of the random walk up to time T.

We then have the following result.

Theorem B.69 (Strong Markov property)**.** *Let T be a stopping time. Then, on the event $\{T < \infty\}$, the random process $(X_{T+n} - X_T)_{n\geq 0}$ has the same distribution as a simple random walk started at 0 and is independent of $\mathscr{F}_T$.*

B.13.2 Local Limit Theorem

Theorem B.70. *There exists $\rho > 0$ such that, for any $i = (i_1, \ldots, i_d) \in \mathbb{Z}^d$ such that $\sum_{k=1}^d i_k$ and n have the same parity and $\|i\|_2 < \rho n$,*

$$\mathbb{P}_0(X_n = i) = 2(2\pi n/d)^{-d/2} \exp\Big(-\frac{d\|i\|_2^2}{2n} + O\big(n^{-1}\big) + O\big(\|i\|_2^4 n^{-3}\big)\Big). \tag{B.31}$$

Proof. See, for example, [211, Theorem 2.3.11], using the fact that the random walk $(X_{2n})_{n\geq 0}$ is aperiodic (see [211, Theorem 2.1.3] for a similar argument). □

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

B.13.3 Recurrence and transience

Given $A \subset \mathbb{Z}^d$, we consider the first entrance times in A, $\tau_A \stackrel{\text{def}}{=} \inf\{n \geq 0 : X_n \in A\}$ and $\tau_A^+ \stackrel{\text{def}}{=} \inf\{n \geq 1 : X_n \in A\}$, with the usual convention that $\inf \varnothing = +\infty$. When $A = \{k\}$, we write simply τ_k, τ_k^+.

Definition B.71. *The random walk is **recurrent** if $\mathbb{P}_0(\tau_0^+ < \infty) = 1$. Otherwise, it is **transient**.*

Theorem B.72. *The simple random walk on $\mathbb{Z}^d$ is transient if and only if*

$$\int_{[-\pi,\pi]^d} \Big\{1 - \frac{1}{2d} \sum_{j \sim 0} \cos(p \cdot j)\Big\}^{-1} \mathrm{d}p < \infty. \tag{B.32}$$

Proof. Let $(X_n)_{n\geq 0}$ be the walk starting at 0 and let $p \stackrel{\text{def}}{=} \mathbb{P}_0(\tau_0^+ < \infty)$. First observe that, by the strong Markov property, the number N_0 of returns of the walk to 0 satisfies, for all $k \geq 0$, $\mathbb{P}_0(N_0 = k) = p^k(1-p)$. In particular,

$$\sum_{n\geq 1} \mathbb{P}_0(X_n = 0) = \mathbb{E}_0\Big[\sum_{n\geq 1} \mathbf{1}_{\{X_n = 0\}}\Big] = \mathbb{E}_0[N_0]$$

if finite if and only if $p < 1$, that is, if and only if X is transient. We show that the convergence of this series is equivalent to (B.32).

Using the identity $(2\pi)^{-d} \int_{[-\pi,\pi]^d} e^{\mathrm{i}p\cdot j}\, \mathrm{d}p = \mathbf{1}_{\{j=0\}}$, for all $j \in \mathbb{Z}^d$, we can rewrite $\mathbb{P}_0(X_n = 0) = (2\pi)^{-d} \int_{[-\pi,\pi]^d} \mathbb{E}_0\big[e^{\mathrm{i}p\cdot X_n}\big]\, \mathrm{d}p$. Now observe that,

$$\mathbb{E}_0\big[e^{\mathrm{i}p\cdot X_n}\big] = \mathbb{E}\big[e^{\mathrm{i}p\cdot(\xi_1+\cdots+\xi_n)}\big] = \big(\tfrac{1}{2d}\sum_{j\sim 0} \cos(p\cdot j)\big)^n \stackrel{\text{def}}{=} \big(\phi_\xi(p)\big)^n.$$

Therefore, for any $\lambda \in (0,1)$,

$$\sum_{n\geq 1} \lambda^n \mathbb{P}_0(X_n = 0) = \int_{[-\pi,\pi]^d} \sum_{n\geq 1} \big(\lambda\phi_\xi(p)\big)^n \frac{\mathrm{d}p}{(2\pi)^d} = \int_{[-\pi,\pi]^d} \frac{\lambda\phi_\xi(p)}{1-\lambda\phi_\xi(p)} \frac{\mathrm{d}p}{(2\pi)^d}.$$

Clearly, $\lim_{\lambda\uparrow 1} \sum_{n\geq 1} \lambda^n \mathbb{P}_0(X_n = 0) = \sum_{n\geq 1} \mathbb{P}_0(X_n = 0)$. It thus only remains for us to show that the limit can be taken inside the integral in the right-hand side. To do that, first observe that $\phi_\xi(p)$ is positive for all $p \in [-\delta,\delta]^d$, as soon as $0 < \delta < \frac{\pi}{2}$. Therefore, by monotone convergence,

$$\lim_{\lambda\uparrow 1} \int_{[-\delta,\delta]^d} \frac{\lambda\phi_\xi(p)}{1-\lambda\phi_\xi(p)} \frac{\mathrm{d}p}{(2\pi)^d} = \int_{[-\delta,\delta]^d} \frac{\phi_\xi(p)}{1-\phi_\xi(p)} \frac{\mathrm{d}p}{(2\pi)^d}.$$

To deal with the integral over $[-\pi,\pi]^d \setminus [-\delta,\delta]^d$, observe that, on this domain, the sequence of functions $\big(\lambda\phi_\xi(p)/\big(1-\lambda\phi_\xi(p)\big)\big)_{0<\lambda<1}$ converges pointwise as $\lambda \uparrow 1$ and is uniformly bounded. Thus, by dominated convergence,

$$\lim_{\lambda\uparrow 1} \int_{[-\pi,\pi]^d\setminus[-\delta,\delta]^d} \frac{\lambda\phi_\xi(p)}{1-\lambda\phi_\xi(p)} \frac{\mathrm{d}p}{(2\pi)^d} = \int_{[-\pi,\pi]^d\setminus[-\delta,\delta]^d} \frac{\phi_\xi(p)}{1-\phi_\xi(p)} \frac{\mathrm{d}p}{(2\pi)^d}$$

and we are done. □

The following corollary, a result originally due to Pólya, shows that the simple random walk behaves very differently in low dimensions ($d = 1, 2$) and in high dimensions ($d \geq 3$).

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

Corollary B.73. *The simple random walk X is transient if and only if $d \geq 3$.*

Proof. A Taylor expansion yields $\cos(x) = 1 - \frac{1}{2}x^2 + \frac{1}{24}x_0^4$ for some $0 \leq x_0 \leq x$. It follows that, for any $x \in [-1,1]$, $1 - \frac{1}{2}x^2 \leq \cos(x) \leq 1 - \frac{11}{24}x^2$. Therefore, changing variables to spherical coordinates, we see that the integral in (B.32) is convergent if and only if

$$\int_0^1 r^{-2} r^{d-1} \, \mathrm{d}r = \int_0^1 r^{d-3} \, \mathrm{d}r < \infty,$$

which is true if and only if $d > 2$. □

By definition, a recurrent random walk returns to its starting point with probability one. The next result quantifies the probability that it manages to travel far away before the first return.

Theorem B.74. *For all $n \geq 1$,*

$$\mathbb{P}_0\big(\tau_{\mathsf{B}(n)^{\mathrm{c}}} < \tau_0^+\big) = \begin{cases} \frac{1}{n+1} & \text{in } d = 1, \\ O\big(\frac{1}{\log n}\big) & \text{in } d = 2. \end{cases}$$

Proof. The first statement is a particular instance of the gambler's ruin estimate; it is discussed, for example, in [209, equation (1.20)]. The second estimate can be found in [209, Proposition 1.6.7]. □

The next result shows that, while a recurrent random walk visits a.s. all vertices, a transient one will a.s. miss arbitrarily large regions on its way to infinity.

Theorem B.75. *For any $r \geq 0$ and any $i \in \mathbb{Z}^d \setminus \mathsf{B}(r-1)$,*

$$\lim_{n\to\infty} \mathbb{P}_i\big(\tau_{\mathsf{B}(n)^{\mathrm{c}}} > \tau^+_{\mathsf{B}(r)}\big) = 1,$$

if and only if X is recurrent.

Proof. See, for example, [209, Chapter 2]. □

B.13.4 Discrete potential theory

The ***n*-step Green function** is defined by

$$G_n(i,j) \stackrel{\text{def}}{=} \mathbb{E}_i\Big[\sum_{k=0}^{n} \mathbf{1}_{\{X_k = j\}}\Big], \qquad i,j \in \mathbb{Z}^d.$$

Let A be a nonempty, proper subset of $\mathbb{Z}^d$. The **Green function in** A is defined by

$$G_A(i,j) \stackrel{\text{def}}{=} \mathbb{E}_i\Big[\sum_{k=0}^{\tau_{A^{\mathrm{c}}}-1} \mathbf{1}_{\{X_k = j\}}\Big], \qquad i,j \in \mathbb{Z}^d.$$

In the transient case, $d \geq 3$, the **Green function** is defined by

$$G(i,j) \stackrel{\text{def}}{=} \mathbb{E}_i\Big[\sum_{n=0}^{\infty} \mathbf{1}_{\{X_n = j\}}\Big], \qquad i,j \in \mathbb{Z}^d.$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

In the recurrent case, $d \le 2$, the **potential kernel** is defined by

$$a(i,j) \stackrel{\text{def}}{=} \lim_{n\to\infty} \{G_n(i,j) - G_n(i,i)\}, \qquad i,j \in \mathbb{Z}^d.$$

We will also use the shorter notations $G_n(i) \equiv G_n(0,i)$, $G_A(i) \equiv G_A(0,i)$, $G(i) \equiv G(0,i)$, $a(i) \equiv a(0,i)$.

Theorem B.76. *1. In $d = 1$,*

$$G_{\mathsf{B}(n)}(0) = a(n+1) = n+1, \qquad \forall n \ge 1.$$

2. In $d = 2$,

$$G_{\mathsf{B}(n)}(0) = \frac{2}{\pi}\log n + O(1), \qquad a(i) = \frac{2}{\pi}\log \|i\|_2 + O(1).$$

3. In $d \ge 3$,

$$G_{\mathsf{B}(n)}(0) = G(0) + O(n^{2-d}), \qquad G(i) = a_d \|i\|_2^{2-d} + O(\|i\|_2^{-d}),$$

where $a_d \stackrel{\text{def}}{=} \frac{d}{2}\Gamma(\frac{d}{2}-1)\pi^{-d/2}$ (Γ denotes here the gamma function).

Proof. The claim in $d = 1$ is proved in [209, Theorem 1.6.4]. Those in $d = 2$ can be found in [209, Theorems 1.6.2 and 1.6.6], and those in higher dimensions are established in [209, Theorem 1.5.4 and Proposition 1.5.8] □

Exercise B.24. *Show that, in $d = 1, 2$,*

$$\lim_{n\to\infty} \big(G_{\mathsf{B}(n)}(i) - G_{\mathsf{B}(n)}(0)\big) = a(i).$$

Finally, we will need the following estimate on the spatial variation of the Green function.

Theorem B.77. *There exists $C < \infty$ such that, for any $A \Subset \mathbb{Z}^2$ and any neighbors $i, j \in \mathbb{Z}^d$,*

$$G_A(i) - G_A(j) \le C.$$

Proof. This follows from [209, Proposition 1.6.3] and the asymptotic behavior of the potential kernel. □

B.14 The isoperimetric inequality on $\mathbb{Z}^d$

In this section, we provide a version of the isoperimetric inequality in $\mathbb{Z}^d$. Given $S \subset \mathbb{Z}^d$, we denote by $\partial_e S \stackrel{\text{def}}{=} \big\{\{i,j\} \in \mathscr{E}_{\mathbb{Z}^d} : i \in S, j \notin S\big\}$ the **edge boundary of S**.

Theorem B.78. *For any $S \Subset \mathbb{Z}^d$,*

$$|\partial_e S| \ge 2d|S|^{(d-1)/d}. \tag{B.33}$$

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Notice that (B.33) is saturated for cubes, for example $S = \mathsf{B}(n)$.

For simplicity, let $|D| \stackrel{\text{def}}{=} \ell^d(D)$ denote the Lebesgue measure of $D \subset \mathbb{R}^d$. The scaling property of the Lebesgue measure then reads $|\lambda D| = \lambda^d |D|$ for all $\lambda > 0$. If $A, B \subset \mathbb{R}^d$, let $A + B \stackrel{\text{def}}{=} \{x + y : x \in A, y \in B\}$.

The following is a weak version of the **Brunn–Minkowski inequality**, adapted from [131, Theorem 4.1]. Let $\mathscr{P}$ denote the collection of all parallelepipeds of $\mathbb{R}^d$ whose faces are perpendicular to the coordinate axes.

Proposition B.79. *If $A, B \subset \mathbb{R}^d$ are finite unions of elements of $\mathscr{P}$, then*

$$|A + B|^{1/d} \geq |A|^{1/d} + |B|^{1/d}. \tag{B.34}$$

Proof. First observe that, for all $x \in \mathbb{R}^d$, $|A+B| = |A+B+x| = |A+(B+x)|$. Therefore, one can always translate A or B in an arbitrary way. In particular, one can always assume A and B to be disjoint.

Now if A and B are arbitrary unions of parallelepipeds, we can express $A \cup B$ as a union $\bigcup_{k=1}^n C_k$, where $C_k \in \mathscr{P}$, and the interior of the C_ks are nonoverlapping (they can, however, share points on their boundaries). We will prove the statement by induction on n.

To prove the claim for $n = 2$, assume that $A \in \mathscr{P}$ has volume $\prod_{i=1}^d a_i$ and that $B \in \mathscr{P}$ is disjoint from A and of volume $\prod_{i=1}^d b_i$. Then, $|A + B| = \prod_{i=1}^d (a_i + b_i)$ and, since $(\prod_{i=1}^d x_i)^{1/d} \leq \frac{1}{d} \sum_{i=1}^d x_i$, see (B.1), we have

$$\Big(\prod_{i=1}^d \frac{a_i}{a_i + b_i}\Big)^{1/d} + \Big(\prod_{i=1}^d \frac{b_i}{a_i + b_i}\Big)^{1/d} \leq \frac{1}{d} \sum_{i=1}^d \frac{a_i}{a_i + b_i} + \frac{1}{d} \sum_{i=1}^d \frac{b_i}{a_i + b_i} = 1,$$

which proves (B.34) for those particular sets.

Let us then suppose that the claim has been proved up to n and assume that A and B are such that their union can be expressed as a union of $n + 1$ nonoverlapping parallelepipeds: $A \cup B = \bigcup_{i=1}^{n+1} C_i$. Since A and B can be assumed to be far apart, A and B can each be expressed using a subset of $\{C_1, \ldots, C_{n+1}\}$. For simplicity, assume that $A = \bigcup_{i=1}^{l} C_i$, $l \geq 2$ and $B = \bigcup_{i=l+1}^{n+1} C_i$.

Observe that C_1 and C_2 can always be separated by some plane π, perpendicular to one of the coordinate axes. Denoting by Π^+ and Π^- the two closed half spaces delimited by π, let $A^\pm \stackrel{\text{def}}{=} A \cap \Pi^\pm$ and $B^\pm \stackrel{\text{def}}{=} B \cap \Pi^\pm$. Again using the fact that B can be translated in an arbitrary manner, we can assume that

$$\frac{|B^\pm|}{|B|} = \frac{|A^\pm|}{|A|}.$$

Now, observe that $A^+ \cup B^+$ and $A^- \cup B^-$ can each be expressed as unions of *at most* n parallelepipeds. We can therefore use the induction hypothesis as follows:

$$\begin{aligned}
|A \cup B| &= |A^+ \cup B^+| + |A^- \cup B^-| \\
&\geq \big(|A^+|^{1/d} + |B^+|^{1/d}\big)^d + \big(|A^-|^{1/d} + |B^-|^{1/d}\big)^d \\
&= |A^+| \Big\{1 + \Big(\frac{|B^+|}{|A^+|}\Big)^{1/d}\Big\}^d + |A^-| \Big\{1 + \Big(\frac{|B^-|}{|A^-|}\Big)^{1/d}\Big\}^d \\
&= |A| \Big\{1 + \Big(\frac{|B|}{|A|}\Big)^{1/d}\Big\}^d \\
&= \big(|A|^{1/d} + |B|^{1/d}\big)^d.
\end{aligned}$$

□

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Proof of Theorem B.78. Remember the notation $\mathscr{S}_0 = [-\frac{1}{2}, \frac{1}{2}]^d$ used for the closed unit cube of $\mathbb{R}^d$. We can always identify $S \Subset \mathbb{Z}^d$ with $A_S \stackrel{\text{def}}{=} \bigcup_{i\in S}\{i + \mathscr{S}_0\} \subset \mathbb{R}^d$. Note that the (Euclidean) boundary of A_S is made of $(d-1)$-dimensional unit cubes which are crossed in their middle by the edges of $\partial_e S$. Notice also that, for all small $\epsilon > 0$, $(A_S + \epsilon\mathscr{S}_0) \setminus A_S$ is a thin layer wrapping A_S, of thickness $\epsilon/2$. We can therefore count the number of edges in $\partial_e S$ by computing the following limit:

$$|\partial_e S| = \lim_{\epsilon \downarrow 0} \frac{|A_S + \epsilon\mathscr{S}_0| - |A_S|}{\epsilon/2}. \tag{B.35}$$

For a fixed $\epsilon > 0$, we use (B.34) as follows:

$$|A_S + \epsilon\mathscr{S}_0| = \left(|A_S + \epsilon\mathscr{S}_0|^{1/d}\right)^d \geq |A_S| + \epsilon d|A_S|^{(d-1)/d}.$$

In the last step, we used $(a+b)^n \geq a^n + na^{n-1}b$ for $a, b \geq 0$, the scaling property of the Lebesgue measure and $|\mathscr{S}_0| = 1$. Using this in (B.35), we get (B.33) since $|A_S| = |S|$. □

Let us finally state an immediate consequence, that is used in Chapter 7.

Corollary B.80. *Let $S \Subset \mathbb{Z}^d$ and write $\partial^{\text{ex}} S \stackrel{\text{def}}{=} \{i \in S^{\text{c}} : d_\infty(i, S) \leq 1\}$. Then,*

$$|\partial^{\text{ex}} S| \geq |S|^{\frac{d-1}{d}}.$$

Proof. Since there can be at most $2d$ edges of $\partial_e S$ incident at a given vertex of $\partial^{\text{ex}} S$, we have $|\partial_e S| \leq 2d|\partial^{\text{ex}} S|$. The conclusion thus follows from (B.33). □

B.15 A result on the boundary of subsets of $\mathbb{Z}^d$

In this section, we provide the tools needed to prove Lemma 7.19.

Consider the set of **$\star$-edges** of $\mathbb{Z}^d$, defined by

$$\mathscr{E}^\star_{\mathbb{Z}^d} \stackrel{\text{def}}{=} \{\{i, j\} \in \mathbb{Z}^d \times \mathbb{Z}^d : \|j - i\|_\infty = 1\}.$$

That is, $\mathscr{E}^\star_{\mathbb{Z}^d}$ contains all edges between pairs of vertices which are corners of the same unit cube in $\mathbb{Z}^d$.

Given $E \subset \mathscr{E}^\star_{\mathbb{Z}^d}$ and a vertex $i \in \mathbb{Z}^d$, we denote by $I(i; E)$ the number of $\star$-edges of E having i as an endpoint. The **boundary** of E is then defined as the set $\partial E \stackrel{\text{def}}{=} \{i \in \mathbb{Z}^d : I(i; E) \text{ is odd}\}$.

A **$\star$-path** between two vertices $i, j \in \mathbb{Z}^d$ is a set $E \subset \mathscr{E}^\star_{\mathbb{Z}^d}$ with $\partial E = \{i, j\}$. A **$\star$-cycle** is a non-empty set $E \subset \mathscr{E}^\star_{\mathbb{Z}^d}$ with $\partial E = \varnothing$.

Two vertices $i, j \in \mathbb{Z}^d$ are **$\star$-connected in** $A \subset \mathbb{Z}^d$ if there exists a $\star$-path between i to j, all of whose $\star$-edges are made of two vertices of A. (A vertex i is always considered to be $\star$-connected to itself.) A set $A \subset \mathbb{Z}^d$ is **$\star$-connected** if all pairs of vertices $i, j \in A$ are $\star$-connected in A. A set $A \subset \mathbb{Z}^d$ is **c-connected** if $A^{\text{c}} \stackrel{\text{def}}{=} \mathbb{Z}^d \setminus A$ is $\star$-connected.

For a set $A \subset \mathbb{Z}^d$, the **$\star$-interior-boundary** is $\partial^{\text{in}}_\star A \stackrel{\text{def}}{=} \{i \in A : \exists j \notin A, \{i, j\} \in \mathscr{E}^\star_{\mathbb{Z}^d}\}$, the **$\star$-exterior-boundary** is $\partial^{\text{ex}}_\star A \stackrel{\text{def}}{=} \{i \notin A : \exists j \in A, \{i, j\} \in \mathscr{E}^\star_{\mathbb{Z}^d}\}$ and the **$\star$-edge-boundary** is $\partial_\star A \stackrel{\text{def}}{=} \{\{i, j\} \in \mathscr{E}^\star_{\mathbb{Z}^d} : i \in A, j \notin A\}$.

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook

Let the set of **$\star$-triangles** be defined by

$$\mathscr{T} \stackrel{\text{def}}{=} \big\{[i,j,k] \stackrel{\text{def}}{=} \big\{\{i,j\},\{j,k\},\{k,i\}\big\} \subset \mathscr{E}^\star_{\mathbb{Z}^d}\big\}.$$

That is, a $\star$-triangle is a cycle built out of three distinct $\star$-edges whose endpoints all belong to the vertices of a common unit cube in $\mathbb{Z}^d$.

In the sequel, it will be convenient to identify a subset $E \subset \mathscr{E}^\star_{\mathbb{Z}^d}$ with the element of $\{0,1\}^{\mathscr{E}^\star_{\mathbb{Z}^d}}$ equal to 1 at each $\star$-edge $e \in E$ and 0 everywhere else. The set $\{0,1\}^{\mathscr{E}^\star_{\mathbb{Z}^d}}$ can be seen as a group for the coordinate-wise addition modulo 2, which we denote by $\oplus$. With this identification, the symmetric difference between two sets E and F can be expressed as $E \triangle F = E \oplus F$. In particular, $E \oplus E = \varnothing$.

The following is a discrete version of (a special case of) the *Poincaré Lemma* of differential topology. Informally, it states that any $\star$-cycle can be realized as the boundary of a surface built out of $\star$-triangles.

Lemma B.81. *Let C be a bounded $\star$-cycle. There exists a finite collection of $\star$-triangles $\mathscr{T}' \subset \mathscr{T}$ such that*

$$C = \bigoplus_{T \in \mathscr{T}'} T.$$

The constructive proof given below uses the following elementary property: if C is a cycle and T is a triangle, then $C \oplus T$ is again a cycle or is empty.

Proof of Lemma B.81: We construct $\mathscr{T}'$ using the following algorithm.

Step 0. Set $\mathscr{T}' = \varnothing$.

Step 1. If C is empty, then stop. Otherwise, go to Step 2.

Step 2. If there exist two $\star$-edges $e = \{i,j\}, e' = \{j,k\}$ in C with $\|k-i\|_\infty = 1$, then:

- $T = [i,j,k]$ is a $\star$-triangle;
- replace $\mathscr{T}'$ by $\mathscr{T}' \cup \{T\}$;
- replace C by $C \oplus T$. Note that the number of $\star$-edges in C decreases at least by 1 in this operation.
- Go to Step 1.

Otherwise go to Step 3.

Step 3. Let us denote by $[C]$ the smallest (with respect to inclusion) parallelepiped $\{a_1,\dots,b_1\} \times \cdots \times \{a_d,\dots,b_d\} \subset \mathbb{Z}^d$, $a_m \le b_m$, such that C is a $\star$-cycle in $[C]$. Let $\ell = \min\{1 \le m \le d : a_m < b_m\}$. Let $e = \{i,j\}, e' = \{j,k\}$ be two $\star$-edges in C such that $j \in \partial^{\text{in}}_\star[C]$ and the ℓth component of j is equal to b_ℓ. Note that, necessarily, $\|k-i\|_\infty = 2$. Let $j' \in [C] \setminus \partial^{\text{in}}_\star[C]$ such that $\|j'-i\|_\infty = \|j'-k\|_\infty = 1$. We add to $\mathscr{T}'$ the two triangles $T_1 = [i,j,j']$ and $T_2 = [j,k,j']$ and replace C by $C \oplus T_1 \oplus T_2$. Note that during this operation, the number of $\star$-edges in C does not increase and either (i) $[C]$ decreases (with respect to inclusion), or (ii) the number of vertices in $C \cap \partial^{\text{in}}_\star[C]$ decreases. Go to Step 1.

The algorithm terminates after finitely many steps, yielding a finite set of triangles $\mathscr{T}'$ such that $C = \bigoplus_{T \in \mathscr{T}'} T$. □

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

Proposition B.82. *Let $A \subset \mathbb{Z}^d$ be $\star$-connected and* c*-connected. Then $\partial^{\text{in}}_\star A$ and $\partial^{\text{ex}}_\star A$ are $\star$-connected.*

The idea used in the proof is due to [333]. It is based on

Lemma B.83. *Let $\partial_\star A = E_1 \cup E_2$ be an arbitrary partition of $\partial_\star A$. Then there exists a $\star$-triangle containing at least one $\star$-edge from both E_1 and E_2.*

Proof of Proposition B.82: To prove that $\partial^{\text{in}}_\star A$ is $\star$-connected, consider an arbitrary partition $\partial^{\text{in}}_\star A = B_1 \cup B_2$. This partition induces a natural partition of $\partial_\star A$: E_k, $k = 1,2$, is the set of all $\star$-edges of $\partial_\star A$ with one endpoint in B_k. By Lemma B.83, there exists a $\star$-triangle containing at least one $\star$-edge of both E_1 and E_2. This implies that there exist $u \in B_1$ and $v \in B_2$ with $\{u,v\} \in \mathscr{E}^\star_{\mathbb{Z}^d}$. Since the partition was arbitrary, the conclusion follows. The same argument can be made for $\partial^{\text{ex}}_\star A$. □

Proof of Lemma B.83: Consider two arbitrary vertices $i \in A$, $j \notin A$. Let π_1 be a $\star$-path between i and j which does not cross E_2 and π_2 a $\star$-path between i and j which does not cross E_1. The existence of such $\star$-paths follows from our assumptions: given any $\star$-edge $\{u,v\} \in \partial_\star A$ with $u \in A$ and $v \notin A$, i is $\star$-connected to u in A (since A is $\star$-connected), while v is $\star$-connected to j in A^{c} (since A^{c} is $\star$-connected).

Since every vertex has an even number of incident $\star$-edges in $\pi_1 \oplus \pi_2$, the latter set is a $\star$-cycle. Therefore, by Lemma B.81, there exists $\mathscr{T}_{\pi_1,\pi_2} \subset \mathscr{T}$ such that

$$\pi_1 \oplus \pi_2 = \bigoplus_{T \in \mathscr{T}_{\pi_1,\pi_2}} T\,. \tag{B.36}$$

Let us denote by $\mathscr{T}'$ the subset of $\mathscr{T}_{\pi_1,\pi_2}$ composed of all $\star$-triangles containing at least one $\star$-edge of E_1 and set $\mathscr{T}'' \stackrel{\text{def}}{=} \mathscr{T}_{\pi_1,\pi_2} \setminus \mathscr{T}'$. Identity (B.36) can then be rewritten as

$$\pi_1 \oplus \bigoplus_{T \in \mathscr{T}'} T = \pi_2 \oplus \bigoplus_{T \in \mathscr{T}''} T \stackrel{\text{def}}{=} F\,. \tag{B.37}$$

Since i and j are the only vertices with an odd number of incident $\star$-edges, F must contain a path $\tilde{\pi}$ between i and j. Removing the latter's $\star$-edges from F, one is left with a cycle $\tilde{C}$, which can be decomposed as $\tilde{C} = \bigoplus_{T \in \tilde{\mathscr{T}}} T$.

By construction, neither π_2, nor any $\star$-triangle in $\mathscr{T}''$ contains a $\star$-edge of E_1. This implies that $\tilde{\pi}$ must contain an odd number of $\star$-edges of E_2 (since each such $\star$-edge connects a vertex of A and a vertex of A^{c}), while each of the $\star$-triangles in $\tilde{\mathscr{T}}$ must contain either 0 or 2. We conclude that F contains an odd number of $\star$-edges of E_2 and therefore $F \cap E_2 \neq \varnothing$.

Returning to (B.37), this implies that at least one of the $\star$-triangles in $\mathscr{T}'$ contains a $\star$-edge of E_2, since π_1 does not contain any $\star$-edge of E_2. However, by definition, every triangle of $\mathscr{T}'$ contain at least one $\star$-edge of E_1. This proves the claim. □

Revised version, August 22 2017
To be published by Cambridge University Press (2017)

www.unige.ch/math/folks/velenik/smbook