A

Some useful facts from calculus

A.1 Croft's lemma

Croft's lemma comes in handy (in a slightly generalized version) when one is trying to establish regularity of the semigroup of a continuous-time Markov process by the use of skeletons; the usual statement is the following:

Lemma A.1 (Croft's lemma). *Let $f : (A, \infty) \to \mathbb{C}$ be a continuous map for some $A \geq 0$ such that for every $t > A$, we have*

$$\lim_{n\to\infty} f(nt) = l,$$

where l is a real number. Then

$$\lim_{x\to\infty} f(x) = l.$$

In Chapter 4 we use a version of this result near 0:

Lemma A.2 (Croft's lemma near 0). *Let $v : (0, B) \to \mathbb{C}$ be a continuous map for some $B \leq \infty$ such that for every $t \in (0, B)$, we have*

$$\lim_{n\to\infty} v(\frac{t}{n}) = l,$$

where l is a real number. Then

$$\lim_{h\to 0^+} v(h) = l.$$

Both versions are clearly equivalent (consider $v(t) := f(\frac{1}{t})$); let us prove the second one:

Proof. Of course (considering $v - l$), we may take $l = 0$. For fixed $\epsilon > 0$ and $p \in \mathbb{N}$, consider the set

$$C_p := \{t \in (0, l) : \forall n \geq p : |v(\frac{t}{n})| \leq \epsilon\}.$$

J.-F. Collet, *Discrete Stochastic Processes and Applications,*
Universitext, https://doi.org/10.1007/978-3-319-74018-8

The sets $(C_p)_{p\in\mathbb{N}}$ form an increasing sequence, and since v is continuous on the interval $(0,l)$, they are all closed. On the other hand, for all $t \in (0,l)$, the convergence of $v(\frac{t}{n})$ tells us that $t \in C_p$ for some p large enough, and therefore, we have

$$(0,l) = \bigcup_{p\geq 1} C_p.$$

Baire's theorem (see, for instance, [32]) now tells us that at least one set C_p must have nonempty interior, i.e., must contain an interval (a,b). This means that for all $t \in (a,b)$ and $n \geq p$ we have $|v(\frac{t}{n})| \leq \epsilon$, or in other words, for all $x \in (\frac{a}{n}, \frac{b}{n})$ we have $|v(x)| \leq \epsilon$. Summarizing, we have shown the following:

$$\forall x \in \bigcup_{n\geq p} (\frac{a}{n}, \frac{b}{n}) : \quad |v(x)| \leq \epsilon.$$

Consider now the sequence of intervals $I_n := (\frac{a}{n}, \frac{b}{n})$ as n increases. The left endpoint $\frac{a}{n}$ clearly converges to 0, and these intervals eventually overlap. More precisely, the intervals I_n and I_{n+1} overlap if and only if $\frac{b}{n+1} > \frac{a}{n}$, or equivalently if and only if $a < (1+\frac{1}{n})b$. Since $a < b$, this is satisfied for all n large enough, say $n \geq n_0$. We may take $n_0 \geq p$, thus guaranteeing that the union $\cup_{n\geq n_0} I_n$ contains the interval $(0, \frac{a}{n_0})$. Setting $x_0 := \frac{a}{n_0}$, we get $|v(x)| \leq \epsilon$ for all $x \in (0, x_0)$, which shows the desired convergence to 0. □

When discretizing a continuous-time Markov chain in Chapter 4 we make use of the following generalization:

Lemma A.3. *Let $v : (0,B) \to \mathbb{C}$ be a continuous map for some $B \leq \infty$ such that for all $t \in (0,B)$ we have*

$$\lim_{n\to\infty} v(\frac{t}{n}) = \alpha(t),$$

where α is monotonic, right continuous, or left continuous. Then α has to be constant: $\alpha(t) = l$ for all $t \in (0,B)$, and

$$\lim_{h\to 0^+} v(h) = l.$$

Proof. In view of Lemma A.2, we need to show only that α is constant; let us first show that for every rational number $r \in (0,1]$ and every number $s \in (0,B)$, we have $\alpha(rs) = \alpha(s)$. Writing $r = \frac{k}{q}$, we have

$$v(r\frac{s}{kn}) \to \alpha(rs)$$

(consider the sequence (kn) and $t = rs$), and on the other hand,

$$v(r\frac{s}{kn}) = v(\frac{s}{qn}) \to \alpha(s),$$

(consider the sequence (qn)), from which we get $\alpha(rs) = \alpha(s)$.

Now let $t_1, t_2 \in (0, B)$ be any two numbers with $t_1 < t_2$. We may use a sequence of rational numbers (r_n) decreasing to $\frac{t_1}{t_2}$ and a sequence of rational numbers (s_n) decreasing to $\frac{t_1}{t_2}$, with $r_n, s_n \leq 1$. If α is nonincreasing, we have

$$\alpha(t_2) = \alpha(s_n t_2) \leq \alpha(t_1) \leq \alpha(r_n t_2) = \alpha(t_2),$$

and therefore $\alpha(t_1) = \alpha(t_2)$; obviously, this argument (with reversed inequalities) is also valid if α is nondecreasing. If α is right continuous, then letting $n \to \infty$ in the upper bound, we see that $\alpha(t_1) \leq \alpha(t_2)$; therefore α is nondecreasing, and we are done. Similarly, if α is left continuous, we conclude that it has to be nonincreasing. □

A.2 A characterization of exponential functions and distributions

Let us begin with the following characterization of exponential functions, which we shall need in Chapter 4:

Lemma A.4. *Assume that u and α are two maps defined on the interval $[0, \infty) \to \mathbb{R}$ such that*

$$\forall t > 0 : \quad \lim_{n \to \infty} u(\frac{t}{n})^n = \alpha(t).$$

Assume, moreover, that the map α is monotonic, right continuous, or left continuous. Then α is given by

$$\forall t > 0 : \quad \alpha(t) = \alpha(1)^t. \tag{A.1}$$

Proof. We are first going to show that

$$\forall t > 0, \forall k \in \mathbb{N} : \quad \alpha(kt) = \alpha(t)^k. \tag{A.2}$$

Let $k \geq 1$ be a fixed integer; considering the extracted sequence $n_k := kn$, as $n \to \infty$ we obtain

$$u(\frac{t}{nk})^{nk} \to \alpha(t).$$

On the other hand, we have

$$u(\frac{t}{nk})^n \to \alpha(\frac{t}{k}),$$

and therefore,

$$\alpha(\frac{t}{k})^k = \alpha(t),$$

which on changing t into kt yields (A.2).

Let us first prove (A.1) for rational numbers: for nonzero $q \in \mathbb{N}$, taking $t = \frac{1}{q}$ and $k = q$, we have $\alpha(\frac{1}{q}) = \alpha(1)^{\frac{1}{q}}$. Thus for all $k, q \in \mathbb{N}$ with $q \neq 0$, we obtain

$$\alpha(\frac{k}{q}) = \alpha(\frac{1}{q})^k = \alpha(1)^{\frac{k}{q}}.$$

We may now use monotonicity to move on to real numbers: for $t > 0$, let r_k be a sequence of rational numbers decreasing to t, and s_k a sequence of rational numbers increasing to t. If α is (for instance) increasing, then for all k we have

$$\alpha(s_k) = \alpha(1)^{s_k} \leq \alpha(t) \leq \alpha(r_k) = \alpha(1)^{r_k}.$$

Taking the limit $k \to \infty$ yields $\alpha(t) = \alpha(1)^t$. If α is left continuous, it suffices to use the sequence s, and the sequence r if α is right continuous. □

A byproduct of the proof is that every function α satisfying (A.2) must be of the form (A.1); this can be used to prove the following characterization of exponential and geometric distributions (commonly referred to as the *memoryless property*):

Corollary A.5. *1. A positive absolutely continuous random variable X satisfies the relation*

$$\forall t, s \geq 0 : \quad P(X > t + s) = P(X > t)P(X > s) \tag{A.3}$$

if and only if it follows an exponential distribution.

2. A positive random variable X taking values in $\mathbb{N}$ satisfies (A.3) if and only if it follows a geometric distribution.

Proof. In either case, note that $P(X > t)$ is a nonincreasing function of t. If X is exponential and λ is its parameter, you may check that

$$P(X > t) = e^{-\lambda t},$$

which implies (A.3); for necessity, note that (A.3) implies that $\alpha(t) := P(X > t$ satisfies (A.2); therefore, if we put $\lambda := -\ln P(X > 1)$, then Lemma A.4 gives us

$$P(X < t) = 1 - e^{-\lambda t}.$$

Thus by differentiating, we obtain $f_X(t) = \lambda e^{-\lambda t}$. For the discrete case, if X is geometric of parameter p, you may check that

$$\forall t > 0 : \quad P(X > t) = (1 - p)^t,$$

and thus (A.3) is satisfied. Conversely, if (A.3) is true, then by setting $p := P(X = 1)$, Lemma A.4 gives us

$$\forall t \geq 1 : \quad P(X > t) = (1 - p)^t.$$

We may now obtain the probability mass function by taking a difference:

$$\forall t \geq 1 : \quad P(X = t) = P(X > t - 1) - P(X > t) = p(1 - p)^{t-1}.$$

Therefore, X is geometric. □

Another consequence of Lemma A.4 is the well-known result that if a continuous function f satisfies the so-called functional equation for the exponential,

$$f(t+s) = f(t)f(s), \tag{A.4}$$

then it is an exponential function; if in (A.4) we replace equality by an inequality, we have a comparison result, which we shall use in Chapter 4:

Lemma A.6. *Assume that f is a map defined on some interval $[0,T)$ with $T \leq \infty$, that it has a right derivative at 0, and that it satisfies the following inequality:*

$$\forall t, s/0 \leq t+s < T: \quad f(t+s) \geq f(t)f(s).$$

Assume, moreover, that $f(0) = 1$. Then

$$\forall t \geq 0: \quad f(t) \geq \exp(tf'(0)) \geq 1 + tf'(0). \tag{A.5}$$

Note that since $\exp(tf'(0))$ is the solution to the functional equation (A.4), this says that every function satisfying the corresponding inequality is necessarily larger than this solution, whence the name *comparison* result.

Proof. For all $t \in [0,T)$ and $n \in \mathbb{N}$ we have $f(t) \geq f(\frac{t}{n})^n$; since f is right continuous at 0, it follows that for fixed t, if n is large enough, then $f(\frac{t}{n})$ is close to 1, so we may manipulate logarithms without a second thought; then as $n \to \infty$ we have

$$\ln f(t) \geq n \ln f(\frac{t}{n}) \sim n(f(\frac{t}{n}) - 1) = t\frac{f(\frac{t}{n}) - 1}{\frac{t}{n}} \to tf'(0).$$

This proves the first half of inequality (A.5), and the second half follows from the classical inequality $e^h \geq 1 + h$ for all $h \geq 0$. □

A.3 Countable sums

If $(z_i)_{n\in\mathbb{N}}$ is a sequence of complex numbers, we all know what is meant by the sum of the associated series (whenever the limit exists):

$$\sum_{n=0}^{\infty} z_n := \lim_{N\to\infty} \sum_{n=0}^{N} z_n.$$

If, on the other hand, we have a countable set $\mathcal{S}$ of complex numbers, how can we define the sum $\sum_{z\in\mathcal{S}} z$ of all its elements? The surprising fact is that if we define a numbering of the elements of $\mathcal{S}$, i.e., if we choose a one-to-one map $z : i \in \mathbb{N} \mapsto z_i \in \mathcal{S}$ (after all, countable sets are exactly those for which such a map may be found), then the sum $\sum_{i=0}^{\infty} z_i$ depends on the choice of the map z. More precisely, if $\pi : \mathbb{N} \to \mathbb{N}$ is a bijective map and $(z_i)_{n\in\mathbb{N}}$ is a sequence of complex numbers, in general the two following sums will differ:

$$\sum_{i=0}^{\infty} z_i \neq \sum_{j=0}^{\infty} z_{\pi(j)}$$

(you should play around with the sequence $(-1)^n$ to convince yourself of this). In other words, the value we get for the sum depends on the order in which we throw in the elements of $\mathcal{S}$. The good news is that this does not happen when we deal with nonnegative numbers:

Theorem A.7. *Let $(z_i)_{n\in\mathbb{N}}$ be a sequence of nonnegative numbers, $\mathcal{S} := \{z_i,\, i \in \mathbb{N}\}$ the set of all its terms, and $\mathcal{F}$ the collection of all finite subsets of $\mathcal{S}$. Define the sum of all elements in $\mathcal{S}$ as*

$$\sum_{x\in\mathcal{S}} x := \sup\{\sum_{x\in F} x, \quad F \in \mathcal{F}\}.$$

Then

$$\sum_{x\in\mathcal{S}} x = \sum_{i=0}^{\infty} z_i$$

(whether the series is finite or infinite).

Proof. Every partial sum is obviously less than $\sum_{x\in\mathcal{S}}$; on the other hand, if F is finite, then for n large enough we have $F \subset \{1, \dots, \}$, so that

$$\sum_{x\in F} x \leq \sum_{i=0}^{n} z_i \leq \sum_{i=0}^{\infty} z_i.$$

□

The sum $\sum_{x\in\mathcal{S}}$ as we just defined it depends only on the set $\mathcal{S}$; therefore, any other numbering of its elements would give the same value for the sum of the series.

In Chapter 4 we often encounter functions defined as sums of series; the main tool for establishing continuity of such sums is the following result, which is a special case of the Lebesgue dominated convergence theorem:

Theorem A.8. *Let $(u_n)_{n\in\mathbb{N}}$ be a sequence of maps defined on some interval $[0, T)$ with $T \leq \infty$ such that each u_n is right continuous at 0 that satisfies the following "domination" inequality:*

$$\forall t \in [0, T), \forall n \in \mathbb{N} : \quad |u_n(t)| \leq v_n,$$

where $(v_n)_{n\in\mathbb{N}}$ is a sequence of nonnegative numbers such that the series $\sum v_n$ is finite. Then the sum $\sum_n u_n(t)$ is right continuous at 0:

$$\lim_{t\to 0^+} \sum_{n=0}^{\infty} u_n(t) = \sum_{n=0}^{\infty} u_n(0).$$

The proof is a very easy $\frac{\epsilon}{2} + \frac{\epsilon}{2}$ exercise, which we leave to the reader. More interestingly, the essence of this result is that the tail of the infinite sum is controlled by the domination inequality, which ensures that the infinite sum behaves exactly as a finite one would. Also, note that we formulated our result here as a statement on right continuity at 0, but it may be easily adapted to prove (two-sided) continuity of the sum of a series of continuous functions at a given point.

A.4 Right continuous and right constant functions

For a set T, the notion of continuous maps with values in T depends on the topology we use. Recall that the *discrete topology* on T is the one for which all sets are open. This means that all one-point sets $\{x\}$ are open, and as a consequence, every convergent sequence of points in T has to be constant after a certain point. In other words, if $t_n \in T \to t \in T$, then for some n_0 we have $t_n = t_{n_0}$ for all $n \geq n_0$. It turns out that the right-constancy condition we saw in Chapter 4 is exactly equivalent to right continuity for this discrete topology:

Lemma A.9. *Let T be any set endowed with the discrete topology and I an open interval of $\mathbb{R}$. A map $f : I \to T$ is right continuous if and only if it is right constant, meaning that for every $x \in I$, there exists some $h > 0$ such that $x + h \in I$, and f is constant on the interval $[x, x + h]$.*

Proof. Trivially, right constancy implies right continuity; to prove the converse, let us argue by contradiction, i.e., assume that f is right continuous and no h satisfies the above condition, and consider a sequence $h_n \to 0$. Then we may find a subsequence (call it still h_n) for which each term $f(x + h_n)$ differs from the preceding one $f(x + h_{n-1})$. On the other hand, since f is right continuous at x, we must have $f(x + h_n) \to f(x)$, which implies that the sequence $f(x + h_n)$ has to be eventually constant, a contradiction. □

To put it in more graphical terms: take any point on the graph of f; then to the right of this point, the graph of x will locally be a horizontal segment. Note that this result says nothing about the width of this step: it may be very small, and we may even have situations in which the width of consecutive steps forms a sequence going to 0; a typical example is given by

$$f(x) := n \quad \forall x \in [\frac{1}{2^n}, \frac{1}{2^n} + \frac{1}{2^{n+1}}), \forall n \in \mathbb{N}.$$

Obviously, right constancy will imply right continuity whatever topology we use on T; however, the converse implication is not true in general. Consider, for instance, the countable set

$$T := \{0\} \cup \{\frac{1}{n}, n \geq 1\}$$

endowed with the absolute-value distance (inherited from $\mathbb{R}$). Define the map $f : [0, \infty) \to T$ by

$$\forall x \in [\frac{1}{2^{n+1}}, \frac{1}{2^n}) : \quad f(x) = \frac{1}{n},$$

and $f(0) = 0$. Then you may check as an exercise that f is right continuous on $[0, \infty)$, although it does not satisfy the right constancy property at 0. Finally, note that there is no contradiction with Lemma A.9: the sequence $\frac{1}{n}$ does not converge to 0 in the discrete topology on T.

A.5 The Gamma function

The Euler Gamma function is defined as follows:

$$\forall z > 0 : \quad \Gamma(z) := \int_0^\infty t^{z-1} e^{-t}\, dt. \tag{A.6}$$

In fact, this definition makes sense for every complex number z with $\Re z > 0$, but in this text we have no use for that. As is well known, this is a generalization of the factorial in the sense that for every positive integer n we have $\Gamma(n) = (n-1)!$; the asymptotics for z large are provided by the famous Stirling formula:

$$z \to \infty : \quad \Gamma(z+1) \sim \sqrt{2\pi z}(\frac{z}{e})^z. \tag{A.7}$$

Another important property of the Gamma function that we will need is its log-convexity (see Problem 6.12 for a proof, and the monograph [1] for other properties of the Γ function). This means that for every set of n nonnegative numbers $\lambda_1, \ldots, \lambda_n$ satisfying $\lambda_1 + \cdots + \lambda_n = 1$ and every set of nonnegative numbers $x_1, \ldots, x_n$, we have

$$\ln \Gamma(\sum \lambda_i x_i) \leq \sum \lambda_i \ln \Gamma(x_i).$$

As a particular case, if we take $X_i = n_i + 1$ (where n_i is a nonnegative integer) and $\lambda_i = \frac{1}{n}$, we obtain the following inequality, which we shall make use of in the study of random walks:

$$\prod_{i=1}^n n_i! \geq \big[\Gamma(1 + \sum_{i=1}^n \frac{n_i}{d})\big]^n. \tag{A.8}$$

A.6 The Laplace transform

This is a minimal reminder of a few properties of the Laplace transform that we make use of in the text; as usual, we take a user-oriented approach and do not seek general or sharp results (for an excellent introduction to this topic we

recommend taking a look at [33]). Recall that if $f : [0, \infty) \to \mathbb{R}$ is a function, its Laplace transform is the map Lf defined by

$$(Lf)(z) := \int_0^\infty e^{-zt} f(t)\, dt$$

(it may be checked that $(Lf)(z)$ as defined here makes sense for every complex number with real part greater than some number α that depends on f; but here we shall focus on only the algebraic properties of the Laplace transform without getting into summability issues). The first elementary property of the Laplace transform is its behavior with respect to differentiation, which is what makes it so useful as a tool for constant-coefficient differential equations. More precisely, if f is differentiable, by a very easy integration by parts you may show that

$$(Lf')(z) = z(Lf)(z) - f(0).$$

Another key property is its behavior with respect to the convolution product. Recall that if f and g are two maps defined on $[0, \infty)$, we may extend them to $\mathbb{R}$ by adding the value 0, i.e., replacing them by $f\mathbb{1}_{[0,\infty)}$ and $g\mathbb{1}_{[0,\infty)}$ respectively. Then their convolution product $f * g$ which is defined by

$$(f * g)(t) := \int_{\mathbb{R}} f(t - s)g(s)\, ds$$

reduces to an integral on a bounded interval:

$$(f * g)(t) = \int_0^t f(t - s)g(s)\, ds.$$

The Laplace transform turns convolution products into usual products: for any two maps f, g and $t > 0$ we have

$$(L(f * g))(t) = Lf(t)Lg(t)$$

(again this is an easy computation; if in doubt, check [33] for details). Let us now examine some special cases that we will need in Chapter 5. If a is a positive real number, let ϕ_a be the function defined by $\phi_a(t) := e^{-at}$. By a straightforward computation we have

$$(L\phi_a)(z) = \frac{1}{z + a}.$$

In the study of birth–death processes we will need the corresponding result for an arbitrary number of simple poles. Let us begin with the case of two factors: if a and b are two positive numbers with $a \neq b$, then you may check that

$$(\phi_a * \phi_b)(t) = \frac{e^{-bt} - e^{-at}}{a - b}.$$

This is therefore the inverse Laplace transform of the rational function $\frac{1}{(z+a)(z+b)}$; here is the n-term statement (see, for instance, [27], p. 224 or [33], p. 39):

Theorem A.10. *Let $n \geq 2$ be an integer, let $a_1, \ldots, a_n$ be n distinct positive numbers, and define*

$$F_n(z) := \prod_{k=1}^{n} \frac{1}{(z + a_k)};$$

then the inverse Laplace transform of F_n is

$$f_n(t) = \sum_{k=1}^{n} e^{-a_k t} \prod_{j=1, j\neq k}^{n} \frac{1}{(a_j - a_k)}. \tag{A.9}$$

This means that the right-hand side of (A.9) is the convolution product $(\phi_{a_1} * \cdots * \phi_{a_n})(t)$.

Proof. The idea is to reduce the proof to the scalar case using partial fractions. The partial fraction decomposition of F_n is

$$F_n(z) = \frac{A_1}{z + a_1} + \cdots + \frac{A_n}{z + a_n},$$

where the (so-called) residue A_k is given by

$$A_k = \lim_{z \to -a_k} (z + a_k) F_n(z),$$

which means

$$A_k = \prod_{j=1, j\neq k}^{n} \frac{1}{a_j - a_k}.$$

The representation (A.9) follows immediately by linearity of the (inverse) Laplace transform. □

The case in which all the a_k are equal will be needed in the study of the Poisson process; if G_n is defined by

$$G_n(z) := \frac{1}{(z + a)^n},$$

then it is easy to check that its inverse Laplace transform is given by

$$g_n(t) = \frac{t^{n-1}}{(n-1)!} e^{-at}. \tag{A.10}$$

One way to show this is to check directly that the Laplace transform of g_n is G_n (in computing the integral you will stumble on the Euler gamma function; if necessary, see Appendix B for details); the other is to show that the n-fold convolution product $\phi_a * \cdots * \phi_a$ is equal to g_n, which is easily done by induction. This is an exercise in calculus; however, it has a probabilistic substance: it is equivalent to the statement that the sum of n exponential variables has a gamma distribution; again see Appendix B on the gamma distribution if more details are needed.

B

Some useful facts from probability

B.1 Some limit theorems in probability theory

B.1.1 Continuity of probability from above and below

We will repeatedly have to compute the probability of some event involving an infinite number of times by approximating it by events involving finitely many times only; the main technical device to do this is the so-called property of *continuity of probability*:

Theorem B.1. *1. (continuity from below of probability) If $A_1 \subset A_2 \subset \cdots \subset A_n \cdots$ is an increasing sequence of events, then $P(A_k)$ has a limit as $k \to \infty$, given by*

$$\lim_{k\to\infty} P(A_k) = P(\bigcup_{n=1}^{\infty} A_n).$$

2. (continuity from above of probability) If $A_1 \supset A_2 \supset \cdots \supset A_n \cdots$ is a decreasing sequence of events, then $P(A_k)$ has a limit as $k \to \infty$, given by

$$\lim_{k\to\infty} P(A_k) = P(\bigcap_{n=1}^{\infty} A_n).$$

Proof. To prove continuity from below, we write A_n and $\bigcup_{j=1}^{\infty} A_j$ as disjoint unions:

$$A_n = A_1 \cup (A_2 \setminus A_1) \cup \cdots \cup (A_n \setminus A_{n-1}),$$
$$\bigcup_{j=1}^{\infty} A_j = A_1 \cup (A_2 \setminus A_1) \cup \cdots \cup (A_j \setminus A_{j-1}) \cup \cdots .$$

By σ-additivity we obtain

J.-F. Collet, *Discrete Stochastic Processes and Applications*,
Universitext, https://doi.org/10.1007/978-3-319-74018-8

$$P(\bigcup_{j=1}^{\infty} A_j) = P(A_1) + \sum_{j=2}^{\infty} P(A_j \setminus A_{j-1})$$

$$= \lim_{n\to\infty} [P(A_1) + \sum_{j=2}^{n} P(A_j \setminus A_{j-1})] = \lim_{n\to\infty} P(A_n).$$

Continuity from above follows immediately if we apply continuity from below to the decreasing sequence $(\Omega \setminus A_n)_{n\in\mathbb{N}}$:

$$P(\Omega \setminus \bigcap_{n=1}^{\infty} A_n) = P(\bigcup_{n=1}^{\infty} [\Omega \setminus A_n]) = \lim_{n\to\infty} P(\Omega \setminus A_n).$$

□

A first consequence that we will need in Chapter 4 is the right continuity of the cumulative distribution function of a random variable:

Corollary B.2. *For a random variable X, its cumulative distribution function F defined by $F_X(t) := P(X \leq t)$ is a right continuous function. More generally, every function of t of the form $P(X \leq t, A|B)$, where A and B are two events with $P(B) \neq 0$, is right continuous.*

Proof. For a sequence (t_n) decreasing to t, the event $\{X \leq t\}$ may be represented as the decreasing intersection

$$\{X \leq t\} = \bigcap_{n\geq 1} \{X \leq t_n\}.$$

Therefore, continuity from above tells us that $F_X(t) = \lim_{n\to\infty} F_X(t_n)$, which exactly means that F_X is right continuous. The very same argument may be applied to the function $P(X \leq \cdot, A|B)$. □

Continuity from above is often used in the case that the complete intersection has zero probability. Here is a typical example, which we use in Chapter 4:

Corollary B.3. *Let X be a random variable satisfying $P(X > 0) = 1$. Then*

$$\lim_{h\to 0^+} P(X < h) = 0.$$

Proof. For any sequence h_k decreasing to zero, continuity from above applied to the decreasing sequence of events $\{X < h_k\}$ immediately gives the result. □

An equivalent statement near ∞ is the following:

Corollary B.4. *Let X be a random variable satisfying $P(X < \infty) = 1$. Then*

$$\lim_{n\to\infty} P(X \leq n) = 1.$$

Proof. Apply continuity from below to the sequence $\{X \leq n\}$, or apply the previous result to $\frac{1}{X}$. □

B.1.2 Three notions of convergence of sequences of random variables

If $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables on the same probability space, we say that

1. X_n converges to X *in probability* if for every $\epsilon > 0$,
$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0;$$
2. X_n converges to X *almost surely* if
$$P(\lim_{n\to\infty} X_n = X) = 1;$$
3. X_n converges to X *in distribution* if for every x,
$$\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x).$$

The corresponding notation is respectively

$$X_n \xrightarrow{p} X, \quad X_n \xrightarrow{a.s.} X, \quad X_n \xrightarrow{d} X.$$

Convergence in distribution is just pointwise convergence of the cumulative distribution functions; note that if the probability space is discrete, this is equivalent to convergence of the probability mass function, i.e., to the requirement that

$$\forall x \in \mathcal{S} : \; P(X_n = x) \to P(X = x).$$

It is well known that almost sure convergence implies convergence in probability; conversely, the well-known monotone convergence theorem asserts that if a sequence of variables $(X_n)_{n\in\mathbb{N}}$ is increasing (meaning $X_n(\omega) \leq X_{n+1}(\omega)$ for all ω) and converges to X in probability, then it also converges to X almost surely.

B.2 Exponential and related distributions

B.2.1 Some properties of the exponential distribution

Recall that for $\lambda > 0$, the exponential distribution is defined over $[0, \infty)$ by the density $f_\lambda(t) = \lambda e^{-\lambda t}$; we write $T \sim \mathcal{E}(\lambda)$ to indicate that a variable T has this distribution; you should have no difficulty checking that in this case, the expected value of T is $\frac{1}{\lambda}$. In the study of the Poisson process we shall need the following two properties of independent exponentials:

Proposition B.5. *If $X \sim \mathcal{E}(\lambda)$ and $Y \sim \mathcal{E}(\mu)$ are two independent exponential variables, then*

1. $\lambda X \sim \mathcal{E}(1)$;
2. $\min(X, Y) \sim \mathcal{E}(\lambda + \mu)$,
3. $P(X < Y) = \frac{\lambda}{\lambda+\mu}$.

The first two facts are easily established by determining the cumulative distribution functions (of λX and $\min(X, Y)$ respectively), and the third one by computing the integral of the joint density $f_{X,Y}(x, y) = \lambda e^{-\lambda x} \mu e^{-\mu y}$ over the set $\{y \in [0, \infty), 0 < x < y\}$. In investigating the explosion time of a continuous-time Markov chain in Chapter 4, we will require the following result on sums of exponential variables:

Lemma B.6. *Let $(X_k)_{k\in\mathbb{N}}$ be a collection of i.i.d. $\mathcal{E}(1)$ variables; then their sum almost surely converges to infinity:*

$$\sum_{k=1}^{n} X_k \overset{a.s.}{\to} \infty \quad \text{as} \quad \text{n} \to \infty.$$

Proof. Let $M > 0$ be fixed; for every $n \geq 2M$, writing $S_n := \sum_{k=1}^{n} X_k$, we have the inclusion of events

$$\{S_n \leq M\} \subset \{S_n - n \leq -\frac{n}{2}\} \subset \{|S_n - n| \geq \frac{n}{2}\};$$

hence

$$P(S_n \leq M) \leq P(|S_n - n| \geq \frac{n}{2}).$$

Using Chebyshev's inequality (note that $E(S_n) = n$ and $var(S_n) = n$), we obtain

$$P(S_n \leq M) \leq \frac{n}{\frac{n^2}{4}} = \frac{4}{n}.$$

This shows that S_n converges to ∞ in probability. Thus by the monotone convergence theorem, it also converges almost surely. □

B.2.2 The Gamma distribution

If $T_1, \ldots, T_k$ are k i.i.d. variables of exponential variables with parameter λ, then the sum $T_1 + \cdots + T_k$ has the following density:

$$f_{k,\lambda}(t) = \lambda^k \frac{t^{k-1}}{(k-1)!} e^{-\lambda t}, \quad t > 0. \tag{B.1}$$

This can be checked directly by convolution. More precisely, you should try to show by induction that

$$f_{k,\lambda} = f_\lambda * \cdots * f_\lambda,$$

with k terms on the right-hand side. Relation (B.1) defines the so-called $\Gamma(k, \lambda)$ distribution (sometimes also called the *Erlang distribution*), here presented in the shape-rate parametrization. The obvious generalization of (B.1)

to the case in which k is not an integer is obtained by replacing the factorial with the Γ function:

$$f_{k,\lambda}(t) = \lambda^k \frac{t^{k-1}}{\Gamma(k)} e^{-\lambda t}, \quad t > 0.$$

This makes sense as soon as k is a positive number, and definition (A.6) immediately shows that this is indeed a probability density; however, when k is not an integer, the interpretation as a density of a sum of exponential variables is lost.

B.2.3 The truncated exponential distribution

In our study of the "bus paradox" in Chapter 3 we come across the truncated exponential distribution, which is defined and characterized as follows:

Definition B.7. *A truncated exponential variable is a variable of the form $Y = \min(X, t)$, where X is an exponential variable of parameter $\lambda > 0$, for some $\lambda > 0$, $t > 0$. In this case, we write $Y \sim \mathcal{TE}(\lambda, t)$.*

It is immediate to check that the cumulative distribution of Y is given by

$$P(Y \leq s) = \begin{cases} 1 & \text{for} \quad s \geq t, \\ 1 - e^{-\lambda s} & \text{for} \quad s < t. \end{cases}$$

Equivalently, this means that the probability measure of Y is given by

$$P_Y = \lambda e^{-\lambda s} \mathbb{1}_{[0,t]} dx + e^{-\lambda t} \delta_t$$

(where dx is the Lebesgue measure on $\mathbb{R}$, and δ_t is the Dirac measure at t). For statements that do not involve measure theory, this is also equivalent to the fact that for every interval A, the probability $P(Y \in A)$ is given by

$$P(Y \in A) = e^{-\lambda t} \mathbb{1}_A(t) + \int_{A \cap [0,t]} \lambda e^{-\lambda s} \, ds,$$

and also that for every test function α we have

$$E(\alpha(Y)) = e^{-\lambda t} \alpha(t) + \int_0^t \alpha(s) \lambda e^{-\lambda s} \, ds. \tag{B.2}$$

Note that Y is a variable that is neither discrete nor absolutely continuous: in the definition of Y the whole mass of the exponential variable X beyond t has been transferred to the single point t. This is accounted for by the first term in the above expression for $P(Y \in A)$ or $E(\alpha(Y))$.

B.2.4 Binomial coefficients, binomial and related distributions

B.2.5 The Vandermonde convolution identity

In a few places in the text we make use of the following relation, known as *Vandermonde's identity*:

$$\binom{n+m}{l} = \sum_{k=0}^{l} \binom{n}{k}\binom{m}{l-k}. \tag{B.3}$$

Proof. Depending on your taste, this may be proved algebraically or combinatorially. For the algebraic proof we use the identity

$$(1+x)^{n+m} = (1+x)^n(1+x)^m.$$

Expanding all three terms by the binomial formula, we obtain

$$\sum_{l=0}^{n+m}\binom{n+m}{l}x^l = [\sum_{k=0}^{n}\binom{n}{k}x^k][\sum_{j=0}^{m}\binom{m}{j}x^j] = \sum_{k=0}^{n}\sum_{l=k}^{n+m}\binom{n}{k}\binom{m}{l-k}x^l,$$

where the last expression was obtained by the change of summation index $l = k + j$. Using Fubini's theorem (draw a picture if necessary), this last sum may be rewritten as follows:

$$\sum_{k=0}^{n}\sum_{l=k}^{n+m}\binom{n}{k}\binom{m}{l-k}x^l = \sum_{l=0}^{n+m}[\sum_{k=0}^{l}\binom{n}{k}\binom{m}{l-k}]x^l.$$

Identification with the expansion of $(1+x)^{n+m}$ yields (B.3). To prove (B.3) combinatorially, note that the general term on the right-hand side counts the number of ways to draw k of objects sequentially from a pool of n objects and then $l-k$ objects from another pool of m objects; summing over k, all we did was select l objects from $l+m$ objects (the union of both pools). □

B.2.6 Obtaining the Poisson distribution from the Binomial distribution

We all know that the binomial distribution $\mathcal{B}(n,p)$ has expectation np. If we want to let n get extremely large while maintaining this expectation finite, it is natural to take $p = \frac{n}{\lambda}$, where λ is a positive constant. In the language of Bernoulli trials, this means that we play an infinitely large number of times, with a vanishingly small chance of success. The limiting distribution in these asymptotics turns out to be the Poisson distribution of parameter λ:

Lemma B.8. *Let $\lambda > 0$ be fixed, and let X_n be a sequence of binomial variables with $X_n \sim \mathcal{B}(n, \frac{\lambda}{n})$. Then as $n \to \infty$, the sequence (X_n) converges in distribution to a Poisson variable of parameter λ:*

$$X_n \xrightarrow{d} X \sim \mathcal{P}(\lambda).$$

Proof. Since we are dealing with variables taking values in $\mathbb{N}$, it suffices to show that

$$\forall k \in \mathbb{N}: \quad \lim_{n\to\infty} P(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

This is immediate using the fact that $(1 - \frac{\lambda}{n})^n \to e^{-\lambda}$.
□

B.2.7 The negative binomial distribution

If we toss a coin that gives success with probability p (some people call this an infinite sequence of independent Bernoulli trials), our first success occurs at trial number n if and only if we obtained a sequence of $n-1$ failures followed by a success. If we denote by X the number of the trial that brings our first success, this means that for all $n \geq 1$ we have

$$P(X = n) = p(1-p)^{n-1}.$$

This is the famous *geometric distribution of parameter p*. In this case, we shall write $X \sim \mathcal{G}(p)$. The fact that all probabilities $P(X = k)$ for $k \geq 1$ add up to 1 is a consequence of the formula for the geometric series,

$$\frac{1}{1-t} = 1 + t + \cdots + t^n + \cdots, \tag{B.4}$$

applied to $t = 1 - p$. The *negative binomial distribution* appears when we consider further successes; for $k \geq 1$, the number Y of the trial that brings our kth success is distributed as follows:

$$P(Y = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}, \quad n \geq k.$$

The abbreviated notation to indicate that Y follows the *negative binomial distribution of parameters k* and p is $Y \sim \mathcal{NB}(k, p)$. How do we make sure that all probabilities add up to 1? If we differentiate (B.4) k times, after some elementary manipulation we obtain

$$\frac{1}{(1-t)^{k+1}} = \sum_{n=0}^{\infty} \binom{n}{k} t^{n-k}.$$

This (when applied to $t = 1 - p$) immediately yields

$$\sum_{n=k}^{\infty} P(Y = n) = 1.$$

B.3 Order statistics

If $X := (X_1, \ldots, X_n)$ is a random vector, for each $\omega \in \Omega$ we may order the collection of the values $X_1(\omega), \ldots, X_n(\omega)$ in a nondecreasing fashion. If we denote by $X_{(1)}(\omega), \ldots, X_{(n)}(\omega)$ the corresponding value, this gives us a random vector $(X_{(1)}, \ldots, X_{(n)})$, whose components are called the *order statistics* of X. In particular, the first and last order statistics are the minimum and maximum of the components of X:

$$X_{(1)} = \min_i X_i, \quad X_{(n)} = \max_i X_i.$$

In the study of the Poisson process we will require the following result on the distribution of the order statistics of a random sample (for a proof, see, for instance, [4], Proposition 13.15, p. 285):

Theorem B.9. *If the X_i are i.i.d. continuous variables with common density f, then the joint density of all the order statistics of X is given by*

$$f_{(X_{(1)},\ldots,X_{(n)})}(x_1, \ldots, x_n) = n! f(x_1) \cdots f(x_n) \mathbb{1}_{x_1 < x_2 < \cdots < x_n}.$$

C

Some useful facts from linear algebra

Summary. In this appendix (assuming here that you have at least a working knowledge of the notion of eigenvalues and eigenvectors) I collect without proof the basic facts about linear algebra that are used in Chapter 2; for more on these topics, see, for instance, [23], [15], or [25]. Denote by $\mathcal{M}_n(\mathbb{R})$ the set of all $n \times n$ square matrices with real entries. The main raison d'être for matrices is obviously to multiply vectors; due to the fact that real matrices may have nonreal eigenvalues, it is sometimes necessary to consider the action of a matrix $A \in \mathcal{M}_n(\mathbb{R})$ on $\mathbb{C}^n$ rather than just $\mathbb{R}^n$. Therefore, at the risk of sounding slightly pedantic, we have to consider $\mathcal{M}_n(K)$, where K may be either $\mathbb{R}$ or $\mathbb{C}$.

C.1 Matrix norms

Definition C.1. *A matrix norm on $\mathcal{M}_n(K)$ is a map $|\cdot|$ that to every matrix $A \in \mathcal{M}_n(K)$ assigns a nonnegative number $|A|$ and enjoys the following properties:*

separation: $|A| = 0$ *only if* $A = 0$;
homogeneity: $\forall \lambda \in K, A \in \mathcal{M}_n(K) : |\lambda A| = |\lambda| A$;
subadditivity: $\forall A, B \in \mathcal{M}_n(K) : |A + B| \leq |A| + |B|$;
submultiplicativity: $\forall A, B \in \mathcal{M}_n(K) : |AB| \leq |A||B|$.

The first three requirements tell us that the map $|\cdot|$ is a norm on the vector space $\mathcal{M}_n(K)$ (see, for instance, [25] if a brushup on vector norms is needed).

The *Frobenius norm* (also called the *Hilbert–Schmidt norm* or *Schur norm*) of a matrix is defined by

$$|A|_F := \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}}. \tag{C.1}$$

It is easy to check that this is indeed a matrix norm on $\mathcal{M}_n(K)$.

The *conjugate matrix* A^* of a matrix $A \in \mathcal{M}_n(\mathbb{C})$ is the complex conjugate transpose of A:

J.-F. Collet, *Discrete Stochastic Processes and Applications*,
Universitext, https://doi.org/10.1007/978-3-319-74018-8

$$A^* = \bar{A}^t.$$

A matrix $A \in \mathcal{M}_n(\mathbb{C})$ is said to be *unitary* if it is invertible and its inverse is its conjugate:

$$AA^* = A^*A = I_n.$$

Unfortunately (like many other areas of mathematics, elementary linear algebra suffers from an overabundance of equivalent terms), if a matrix A has real entries, then the word *orthogonal* is preferred: $A \in \mathcal{M}_n(\mathbb{R})$ is said to be orthogonal if

$$AA^t = A^tA = I_n.$$

The Frobenius norm is invariant under multiplication by unitary matrices: if A is any complex matrix and U is a unitary matrix, then

$$|AU|_F = |A|_F.$$

A given norm $|\cdot|$ on K^n may be used to define a norm on $\mathcal{M}_n(K)$ as follows (note that here we are using the same notation $|\cdot|$ whether the argument is a vector or a matrix): if for $A \in \mathcal{M}_n(K)$ we put

$$|A| := \sup_{x \in K^n} \frac{|Ax|}{|x|},$$

then it is easy to check that this does indeed define a matrix norm on $\mathcal{M}_n(K)$, which is said to be *induced by*, or *subordinate to*, the norm on K^n.

The two most important induced norms on $\mathcal{M}_n(K)$ are the following:

Theorem C.2. *On $\mathbb{C}^n$, the ∞-norm and 1-norm are defined by*

$$\forall x \in \mathbb{C}^n : \quad |x|_\infty := \max_i |x_i|, \quad |x|_1 := \sum_i |x_i|.$$

The induced norms on $\mathcal{M}_n(K)$ are respectively the so-called max-absolute-row-sum norm,

$$|A|_\infty = \max_i \sum_j |a_{ij}|,$$

and max-absolute-column-sum norm,

$$|A|_1 = \max_j \sum_i |a_{ij}|.$$

C.2 Eigenvalues and spectral radius

The Schur triangularization theorem (see, for instance, [25], p. 508 for a proof) tells us that every square matrix is unitarily similar to an upper triangular matrix:

Theorem C.3 (Schur's triangularization theorem). *For every square matrix $A \in \mathcal{M}_n(\mathbb{C})$ there exists a unitary matrix $U \in \mathcal{M}_n(\mathbb{C})$ such that the matrix U^*AU is upper triangular.*

Note that in this case, the matrices A and U^*AU must have the same characteristic polynomial. This implies that the diagonal of the matrix U^*AU is made up of the eigenvalues of A, and for each $\lambda \in \sigma(A)$ the number of times it appears on the diagonal is its algebraic multiplicity. An important consequence of Schur's theorem is that it tells us what the spectrum becomes when we compute powers, or more generally polynomials, of a square matrix:

Theorem C.4. *For every square matrix $A \in \mathcal{M}_n(\mathbb{C})$ and complex-valued polynomial P, the spectrum of $P(A)$ is simply the set*

$$\sigma(P(A)) = \{P(\lambda),\, \lambda \in \sigma(A)\}.$$

Indeed, it suffices to notice that if $U^*AU = T$ is the Schur form of A, then $U^*P(A)U = P(T)$ is also upper triangular.

How is the spectral radius related to matrix norms? First, for every subordinate norm we have the bound

$$\rho(A) \leq |A| \tag{C.2}$$

(to prove this, choose an eigenvector x and bound the quantity $|Ax|$). Second, the *spectral radius formula*, a.k.a. Gelfand's formula, asserts that for every matrix norm, the spectral radius $\rho(A)$ is given by

$$\rho(A) = \lim_{k\to\infty} |A^k|^{\frac{1}{k}}. \tag{C.3}$$

Finally, we will need the notion of *convergent matrix*, as well as its characterization in terms of the spectral radius:

Theorem C.5. *For a square matrix A, the limit of A^k as $k \to \infty$ is 0 if and only if $\rho(A) < 1$; such matrices are called convergent matrices.*

For a proof, see, for instance, [25], p. 617. Note that we do not need to be any more precise when we say that A^k converges to 0: since the set of $n \times n$ matrices is finite-dimensional, all norms are equivalent, and this convergence is equivalent to the convergence of $A^k(i,j)$ to 0 for all i, j. Also note that some texts use a slightly different terminology and use *convergent* to describe any matrix A such that A^k has a limit. Here is why convergent matrices are so important:

Theorem C.6. *If A is a convergent matrix, then the matrix $I - A$ is invertible, and its inverse is given by the sum of the series*

$$(I - A)^{-1} = I + A + \cdots + A^n + \cdots .$$

Proof. Let $\epsilon > 0$ be small enough so as to have $\rho(A) + \epsilon < 1$; the spectral radius formula tells us that for k large enough, we have $|A^k| \leq (\rho(A) + \epsilon)^k$, which means that our series is normally convergent. Now, for every n, the partial sum of the series commutes with $I - A$, and

$$(I - A)(I + A + \cdots + A^n) = I - A^{n+1}.$$

The right-hand side converges to I as $n \to \infty$, which shows that the sum S of the series satisfies $S(I - A) = (I - A)S = I$. □

In Chapter 2 we make use of the power method, which is based on the following two results:

Theorem C.7. *Let λ be a simple eigenvalue of a square matrix A, and let x, y be respectively right and left eigenvectors associated with λ. Then $yx \neq 0$, and the projector onto $N(A - \lambda I)$ along $R(A - \lambda I)$ is $\frac{xy}{yx}$.*

(For a proof, see [25], p. 518.)

Theorem C.8. *If $\rho(A) \neq 0$ is a dominant eigenvalue of A, then as $k \to \infty$, the matrix $(\frac{A}{\rho(A)})^k$ converges to the projector onto $N(A - \rho(A)I)$ along $R(A - \rho(A)I)$.*

Proof. This is a special case of a more general result that you will find in [25] (see 7.10.34 on p. 630). Let us give a simple proof based on Theorem C.5. Let Π denote the projector as in the statement of the result; recall that Π commutes with A. Writing any x as $x = \Pi x + (I - \Pi)x$ and applying $(\frac{A}{\rho(A)})^k$, we obtain

$$(\frac{A}{\rho(A)})^k = \Pi + [\frac{A}{\rho(A)}(I - \Pi)]^k. \tag{C.4}$$

If $\mu \neq 0$ is an eigenvalue of $A(I - \Pi)$ and x is an associated eigenvector, applying Π to the relation $A(I - \Pi)x = \mu x$, we obtain $x = (I - \Pi)x$, and thus $Ax = \mu x$. Since $\rho(A)$ is dominant, this implies $\mu < \rho(A)$. This means that the spectral radius of $\frac{A}{\rho(A)}(I - \Pi)$ is strictly less than 1, so by Theorem C.5, the second term in (C.4) converges to 0. □

C.3 Monotone matrices and M matrices

Results of Perron and Frobenius (Theorems 2.15 and 2.27 in Chapter 2) highlight the importance of nonnegative matrices. In this connection it seems well worth investigating the class of invertible matrices with a nonnegative inverse. Here is a nice characterization, Collatz's theorem:

Theorem C.9. *For a square matrix A with real entries, the following two conditions are equivalent:*

$$A \text{ is invertible and } A^{-1} \geq 0, \tag{C.5}$$

$$Ax \geq 0 \implies x \geq 0. \tag{C.6}$$

Proof. The second condition is immediate from the first, as can be seen by multiplying the inequality $Ax \geq 0$ by A^{-1}. Conversely, if the second condition is true, then $Ax = 0$ implies that both Ax and $A(-x)$ are nonnegative, and thus using the condition twice, we get $x \geq 0$ and $-x \geq 0$, thus $x = 0$, which means that A is invertible. To show that $A^{-1} \geq 0$, denoting the ith basis vector by e_i, we note that $AA^{-1}e_i = e_i \geq 0$ implies that the ith column vector $A^{-1}e_i$ of the matrix A^{-1} is nonnegative, i.e., $A^{-1} \geq 0$. □

This motivates the following definition:

Definition C.10. *A square matrix A is said to be monotone if*

$$Ax \geq 0 \implies x \geq 0.$$

How can we prove that a given square matrix is monotone? As in Theorem 2.2 and Lemma 2.8, the "weight" of the diagonal of A plays an important role. Here is the relevant notion:

Definition C.11. *A square matrix A is said to be diagonally dominant if*

$$\forall i: \quad |A(i,i)| \geq \sum_{j \neq i} |A(i,j)|;$$

it is said to be strictly diagonally dominant if all inequalities are strict.

If we add a condition on the signs of the entries, this turns out to ensure monotonicity:

Theorem C.12. *Assume that A is strictly diagonally dominant, that all its diagonal terms are positive, and that all its off-diagonal terms are nonpositive:*

$$\forall i: \quad A(i,i) > 0, \quad \forall j \neq i: \quad A(i,j) \leq 0.$$

Then A is monotone.

Proof. Let D be the diagonal matrix made up of the diagonal elements of A, i.e., $D(i,i) = A(i,i)$. Obviously, D is monotone, and writing

$$A = D - (D - A) = D[I - D^{-1}(D - A)],$$

it suffices to show that $I - D^{-1}(D - A)$ is monotone. Putting $B := D^{-1}(D - A)$, in view of Theorem C.9 we need to show that $I - B$ is invertible and has a nonnegative inverse. Note that $B > 0$, as a consequence of the fact that A is strictly diagonally dominant. Therefore, in view of Theorem C.6, it suffices to show that the series $I + B + \cdots + B^n + \cdots$ converges; this will imply both existence and positivity of the inverse $(I - B)^{-1}$. We shall now show that $\rho(B) < 1$ by considering the row sums of B. Dividing the inequality

$$\sum_{j \neq i} |A(i,j)| < A(i,i)$$

by $A(i,i)$, we obtain

$$\sum_{j \neq i} |B(i,j)| < 1,$$

and thus by Theorem 2.13, we have $\rho(B) < 1$. □

Closely related to this result is the notion of M matrices:

Definition C.13. *A square matrix A is said to be an M matrix if it is monotone and all its off-diagonal elements are nonpositive: $A(i,j) \leq 0$ for $i \neq j$.*

It is quite easy to find monotone matrices that are not M matrices (see, for instance, Problem 2.5 in Chapter 2 for a characterization of 2×2 monotone matrices that are M matrices). For various characterizations of M matrices, see [25], pp. 626 and 639. The following is an important consequence of Theorem C.12 that we will need in Chapter 2 for the study of the PageRank algorithm:

Corollary C.14. *1. Let B be a nonnegative matrix. Then for all $r > \rho(B)$, the matrix $A := rI - B$ is an M matrix.*
2. As a consequence, if P is a Markov matrix, then for all $\alpha \in (0,1)$, the matrix $I - \alpha P$ is an M matrix.

Proof. The off-diagonal elements of A are clearly negative; thus we need to show only that A is monotone, which by Theorem C.9 means showing that it is invertible with nonnegative inverse. Dividing by r, let us consider instead the matrix $I - \frac{B}{r}$. Since $r > \rho(B)$, we have $\rho(\frac{B}{r}) < 1$, and therefore, as in the proof of Theorem C.12, this implies that the matrix $I - \frac{B}{r}$ is invertible, and by Theorem C.6, its inverse is given by the sum of a matrix series in which all terms are nonnegative. This means that $I - \frac{B}{r}$ is monotone, and therefore so is A. The second point follows immediately, since by Theorem 2.1, we know that $\rho(P) = 1$. □

For the sake of completeness let us mention that the converse of the first point in the conclusion is also true, that is, that every M matrix A has a representation of the form $A = rI - B$ with $B \geq 0$ and $r > \rho(B)$ (see, for instance, [25], p. 639, for a proof). Indeed, in some texts it is this representation that is taken as a definition of M matrices.

C.4 Permutation matrices

In Chapter 2 we investigate the properties of the directed graph $G(A)$ associated with a square matrix A; the relevant properties of the graph (such as strong connectivity or lack thereof) are not exactly intrinsic to the matrix, since the definition of $G(A)$ uses a numbering of the vertices. To put it bluntly, it does not matter how you label two vertices; what matters is whether they are connected. The mathematical concept relevant here is that of *permutation*

matrix. If $\sigma \in \mathfrak{S}_n$ is a permutation, we can use σ to define a linear map L_σ on vectors by simply permuting the entries. More precisely, for $x \in \mathbb{R}^n$, the vector $L_\sigma(x)$ is defined by $(L_\sigma(x))(j) = x(\sigma(j))$. The permutation matrix P_σ is the associated matrix acting on row vectors. This motivates the following definition:

Definition C.15. *For a permutation $\sigma \in \mathfrak{S}_n$, the permutation matrix P_σ is the matrix defined by*

$$P_\sigma(i,j) := \delta_{i\sigma(j)}.$$

You may check that this is consistent with our definition of the map L_σ. Note that in some texts, the map L_σ is not used explicitly, and a permutation matrix is just defined as a matrix that in each row and each column has exactly one nonzero entry, which is 1.

We leave it to you as an (easy) exercise to check the following.

Theorem C.16. *Some algebraic properties of permutation matrices are as follows:*

1. *product:*
$$P_{\sigma_1} P_{\sigma_2} = P_{\sigma_2} P_{\sigma_1} = P_{\sigma_1\sigma_2};$$
2. *inverse:*
$$P_\sigma^{-1} = P_\sigma^t = P_{\sigma^{-1}};$$
3. *multiplication of square matrices:*
$$(P_\sigma A)(i,j) = A(\sigma^{-1}(i), j); \quad (AP_\sigma)(i,j) = A(i, \sigma(j));$$
4. *permutation similarity:*
$$(P_\sigma^t A P_\sigma)(i,j) = A(\sigma(i), \sigma(j)). \tag{C.7}$$

At this point you should try to play with a simple permutation, say of three elements, and check by an example the formulas in item 3 above (the first one, for example, says that row k of the matrix A becomes row $\sigma(k)$ of the matrix $P_\sigma A$; what is the corresponding statement for columns?).

C.5 Matrix exponentials

The most elementary (and classical) way to define the exponential of a square matrix is based on the power series expansion of the exponential function for complex numbers:

$$\forall z \in \mathbb{C}: \quad e^z = 1 + z + \frac{z^2}{2} + \cdots + \frac{z^n}{n!} + \cdots.$$

Here is the precise statement:

Theorem C.17. *For an $n \times n$ square matrix A with complex entries, the series*

$$I_n + A + \frac{A^2}{2} + \cdots + \frac{A^k}{k!} + \cdots$$

is normally convergent in every matrix norm. Generalizing the notation used for $n = 1$, we denote by e^A the sum of this series. Then:

1. *If B is another $n \times n$ matrix such that $AB = BA$, we have $e^{A+B} = e^A e^B = e^B e^A$.*
2. *As a consequence, e^A is invertible, and its inverse is e^{-A}.*

Note that since e^A is defined as a limit of polynomials in A, it commutes with A; more generally, if A and B commute, then so do e^A and every polynomial in B.

Proof. The fact that the series is normally convergent comes from the bound $|A^k| \leq |A|^k$, so our matrix exponential is well defined as its sum. To prove its algebraic properties, note that since A and B commute, we may apply the binomial formula to expand all powers of $A + B$; therefore,

$$\begin{aligned} e^{A+B} &= \sum_{n=0}^{\infty} \frac{(A+B)^n}{n!} = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} A^k B^{n-k} \\ &= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{1}{k!(n-k)!} A^k B^{n-k} = \sum_{k=0}^{\infty} \frac{1}{k!} A^k \sum_{l=0}^{\infty} \frac{1}{l!} B^l \\ &= e^A e^B. \end{aligned}$$

□

One of the main uses of matrix exponentials is the resolution of linear differential equations (in $\mathbb{R}^n$) with constant coefficients. Let us begin with a basic differentiability result:

Lemma C.18. *For every $n \times n$ matrix A, the map $t \in \mathbb{R} \mapsto e^{tA}$ is differentiable everywhere on $\mathbb{R}$, and*

$$\frac{d}{dt} e^{tA} = A e^{tA} = e^{tA} A.$$

Proof. Since the radius of convergence of the power series is infinite, we may differentiate it termwise, which immediately gives the result. □

An interesting consequence of commuting matrices is the following:

Lemma C.19. *For two $n \times n$ matrices A and B, the following three conditions are equivalent:*

$$AB = BA; \tag{C.8}$$

$$\forall t \in \mathbb{R} : \quad Be^{tA} = e^{tA} B; \tag{C.9}$$

$$\exists \epsilon > 0 \quad \forall t \in (0, \epsilon) : \quad Be^{tA} = e^{tA} B. \tag{C.10}$$

Proof. If A and B commute, then every polynomial in A commutes with B, which shows that e^{tA} and B commute for all $t \in \mathbb{R}$, so clearly (C.8) implies (C.9) and (C.10). To show that (C.10) implies (C.8), taking derivatives, we obtain for all $t \in (0, \epsilon)$ the relation

$$BAe^{tA} = Ae^{tA}B,$$

which on taking the limit $t \to 0$ yields $AB = BA$. □

Here comes an interesting application to linear ordinary differential equations with constant coefficients:

Theorem C.20. *For a given $n \times n$ matrix A, consider the following Cauchy problem in $\mathcal{M}_{p\times n}(\mathbb{C})$:*

$$\begin{cases} \frac{dX}{dt} = AX, \\ X(0) = X_0, \end{cases} \tag{C.11}$$

in which $X_0 \in \mathcal{M}_{p\times n}(\mathbb{C})$ is given. The unique solution to this Cauchy problem is given by

$$X(t) = e^{tA}X_0.$$

Similarly, if we take $Y_0 \in \mathcal{M}_{n\times p}(\mathbb{C})$, the Cauchy problem

$$\begin{cases} \frac{dY}{dt} = YA, \\ Y(0) = Y_0, \end{cases} \tag{C.12}$$

is well posed in $\mathcal{M}_{n\times p}(\mathbb{C})$ and has the unique solution $Y(t) = Y_0e^{tA}$.

Proof. It follows immediately from Lemma C.18 that $e^{tA}X_0$ is a solution to (C.11). The right-hand side AX is a Lipschitz continuous function of X (whatever matrix norm we use), so by the well-known existence and uniqueness result on ordinary differential equations, our system (C.11) is well posed (see, for instance, [14] for details), which means that $e^{tA}X_0$ is the unique solution. However, in this simple case, uniqueness may be proved directly without having to appeal to the general theory. Indeed, if $X(t)$ is a solution, from (C.18) you can check that $\frac{d}{dt}[e^{-tA}X(t)]$ vanishes identically. This means that $e^{-tA}X(t)$ is constant, and thus using the initial condition, we see that $X(t) = e^{tA}X_0$. This proves the first half of our theorem regarding (C.11); the other half (regarding (C.12)) follows immediately by transposition. □

In the form (C.11), the two most common cases of application of this result are $p = 1$, in which we solve a linear system of n ordinary differential equations in n unknown functions, and $p = n$, in which case the solution to the problem is an $n \times n$ matrix. In this case we may ask whether the solution at an arbitrary time will commute with A if and only if it does so at time 0:

Corollary C.21. *Let A and X_0 be two given $n \times n$ matrices. Then the Cauchy problems*

$$\begin{cases} \frac{dX}{dt} = AX, \\ X(0) = X_0, \end{cases} \tag{C.13}$$

and

$$\begin{cases} \frac{dX}{dt} = XA, \\ X(0) = X_0, \end{cases} \tag{C.14}$$

are both well posed; they are equivalent (meaning both have the same solution) if and only if the matrices A and X_0 commute.

Proof. We already know that both problems are well posed, with respective solutions $e^{tA}X_0$ and X_0e^{tA}, and we know from Lemma C.19 that these are equal for all t if and only if $AX_0 = X_0A$. □

In the framework of complex-valued functions of one real variable, it is well known that under suitable regularity assumptions, the exponential map e^z is the unique solution to the Cauchy problem $y' = y$, $y(0) = 1$. Let us now indicate a generalization of this result to matrix-valued maps:

Theorem C.22. *For some $T \leq \infty$, let $P : [0, T) \to \mathcal{M}_{n\times n}(\mathbb{C})$ be a matrix-valued map satisfying $P(0) = I_n$ and the so-called semigroup property:*

$$\forall t \geq 0, s > 0 : \quad P(t + s) = P(t)P(s). \tag{C.15}$$

Let us assume that P is right differentiable at 0, i.e., that the right derivative

$$A := \lim_{h\to 0^+} \frac{P(t) - I}{h}$$

exists; by this we mean that this limit exists componentwise, and that $A(i, j)$ is finite for all i, j. Then for all $t \geq 0$ we have $P(t) = e^{tA}$.

Proof. We are going to show that the map P is differentiable on the open interval $(0, T)$, and that for all $t > 0$, we have

$$P'(t) = P(t)A = AP(t).$$

In view of Theorem C.20, this will yield the desired representation. First note that from (C.15) we have the commutation relation $P(s)P(t) = P(t)P(s)$. Since P is right differentiable at 0, it is right continuous at 0, i.e., $P(s) \to I$ as $s \to 0$. Therefore, $P(s)$ is invertible for $s > 0$ small enough; we may then use (C.15) to conclude that $P(t)$ is invertible for all $t \in [0, T)$.

Let us first prove continuity on the open interval. From (C.15) we see that for all $t > 0$, the map P is right continuous at t. To check left continuity, for $s > 0$ small enough we have $P(t-s)P(s) = P(t)$; hence $P(t-s) = P(t)P(s)^{-1}$

converges to $P(t)$ as $s \to 0^+$, which shows that P is left continuous, therefore continuous, at t. We may now show differentiability proceeding in a similar fashion, i.e., treating right and left differentiability separately. If $h > 0$, we have

$$\frac{P(t+h) - P(t)}{h} = P(t)\frac{P(h) - I}{h}.$$

As $h \to 0^+$ this converges to $P(t)A$, which shows that the right derivative at t exists and is equal to $P(t)A$. On the other hand, for $h < 0$ we may write

$$\frac{P(t+h) - P(t)}{h} = P(t+h)\frac{P(-h) - I}{-h}.$$

Now, as $h \to 0^-$ this converges to $P(t)A$, which shows that the left derivative at t exists and is equal to $P(t)A$. $\square$

D

An arithmetic lemma

In the course of the proof of Theorem 1.48 we made use of the following fact:

Lemma D.1. *Let A be a set of relatively prime positive integers. Assume that A is closed under addition:*

$$x, y \in A \implies x + y \in A.$$

Then the complement of A in $\mathbb{N}$ is finite.

Let us see how this is proved. We will need the following:

Lemma D.2. *Let A be an infinite subset of $\mathbb{N}$ consisting of relatively prime integers; then there exists a finite subset of A with the same property.*

Proof. Pick some $x \in A$, and let $p_1, \ldots, p_k$ be the prime divisors of x; for every i between 1 and k, p_i cannot divide all elements of A, so let us choose some $x_i \in A$ that p_i does not divide. Then the numbers $x, x_1, \ldots, x_k$ are relatively prime (as an interesting exercise you may want to check that this fact generalizes to any greatest common divisor: every infinite set of positive integers has a finite subset with the same greatest common divisor.) □

We can now prove Lemma D.1:

We begin by remarking that A is closed under multiplication by positive integers, and therefore closed under linear combinations with such coefficients:

$$x, y \in A, n_1, n_2 \in \mathbb{N} \implies n_1 x + n_2 y \in A.$$

Therefore, it suffices to show that every integer larger than some n_0 may be written as a linear combination of this type. From the previous lemma we may choose a finite collection $a_1, \ldots, a_r$ of relatively prime elements of A, and using Bézout's lemma, we have

$$t_1 a_1 + \cdots + t_r a_r = 1$$

J.-F. Collet, *Discrete Stochastic Processes and Applications*,
Universitext, https://doi.org/10.1007/978-3-319-74018-8

for some integers $t_i \in \mathbb{Z}$. In order to obtain an integer n as a linear combination of the a_i's we perform the Euclidian division of n by their sum:

$$n = k \sum_{i=1}^{r} a_i + s, \quad 0 \leq s < \sum_{i=1}^{r} a_i.$$

Thus we obtain

$$n = k \sum_{i=1}^{r} a_i + s \sum_{i=1}^{r} a_i t_i = \sum_{i=1}^{r} (k + st_i) a_i.$$

If n is large enough, all coefficients $k + st_i$ in the above sum are positive, and this provides the desired representation for n. $\square$

E

Table of exponential families

Table E.1. Exponential families: discrete distributions, part 1

Distribution	Binomial	Negative binomial	Geometric	Poisson
$\boldsymbol{P(X=x)}$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\binom{x+r-1}{x}p^x(1-p)^r$	$p(1-p)^{x-1}$	$e^{-\lambda}\frac{\lambda x}{x!}$
$\boldsymbol{x}\in$	$\{1,\cdots n\}$	$\mathbb{N}$	$\mathbb{N}^*$	$\mathbb{N}$
parameter	p	r	p	λ
$\boldsymbol{t(x)}$	x	x	x	x
$\boldsymbol{\theta}$	$\log\frac{p}{1-p}$	$\log p$	$\log(1-p)$	$\log\lambda$
$\boldsymbol{F(\theta)}$	$n\log(1+e^\theta)$	$-r\log(1-e^\theta)$	$-\log(e^{-\theta}-1)$	e^θ
$\boldsymbol{k(x)}$	$\log\binom{n}{x}$	$\log\binom{x+r-1}{x}$	0	$-\log(x!)$
$\boldsymbol{E(X)}$	np	$r\frac{p}{1-p}$	$\frac{1}{p}$	λ
$\boldsymbol{var(X)}$	$np(1-p)$	$r\frac{p}{(1-p)^2}$	$\frac{1-p}{p^2}$	λ

J.-F. Collet, *Discrete Stochastic Processes and Applications,*
Universitext, https://doi.org/10.1007/978-3-319-74018-8

Table E.2. Exponential families: discrete distributions, part 2

Distribution	Zipf	Zeta	Power series	Logarithmic
$\boldsymbol{P(X=x)}$	$\frac{1}{Cx^s}, C := \sum_y y^{-s}$	$\frac{1}{\zeta(s)x^s}$	$u(x)\frac{1}{C\lambda^x}, C := \sum_y u(y)\lambda^y$	$-\frac{1}{\log(1-p)}\frac{p^x}{x}$
$\boldsymbol{x} \in$	$\{1, \cdots N\}$	$\mathbb{N}^*$	$\mathbb{N}$	$\mathbb{N}^*$
parameter	s	s	λ	p
$\boldsymbol{t(x)}$	$-\log x$	$-\log x$	x	x
$\boldsymbol{\theta}$	s	s	$\log\lambda$	$\log p$
$\boldsymbol{F(\theta)}$	$\log(\sum_{x=1}^N x^{-s})$	$\log\zeta(\theta)$	$\log\sum_y u(y)\lambda^y$	$\log(-\log(1-p))$
$\boldsymbol{k(x)}$	0	0	$\log u(x)$	$-\log x$
$\boldsymbol{E(X)}$	*	*	*	$-\frac{p}{1-p}\frac{1}{\log(1-p)}$
$\boldsymbol{var(X)}$	*	*	*	$-p\frac{\log(1-p)-p}{(1-p)^2\log^2(1-p)}$

*Quantities for which no closed formula is available are indicated by a star.

Table E.3. exponential families: absolutely continuous distributions, part 1

Distribution	Exponential	Laplace	Weibull	Pareto
$\boldsymbol{f_X(x)}$	$\lambda e^{-\lambda x}$	$\frac{\lambda}{2}e^{-\lambda\|x\|}$	$\alpha\lambda x^{\alpha-1}\exp(-\lambda x^\alpha)$	$k\frac{x_m^k}{x^{k+1}}$
$\boldsymbol{x} \in$	$[0,\infty)$	$\mathbb{R}$	$[0,\infty)$	$[x_m,\infty)$
parameter	λ	λ	λ	k
$\boldsymbol{t(x)}$	$-x$	$-\|x\|$	$-x^\alpha$	$-\log x$
$\boldsymbol{\theta}$	λ	λ	λ	k
$\boldsymbol{F(\theta)}$	$-\log\theta$	$-\log\frac{\theta}{2}$	$-\log\theta$	$-\log(\theta x_m^\theta)$
$\boldsymbol{k(x)}$	0	0	$(\alpha-1)\log x + \log\alpha$	$-\log x$
$\boldsymbol{E(X)}$	$\frac{1}{\lambda}$	0	$\lambda^{-\frac{1}{k}}\Gamma(1+\frac{1}{\alpha})$	$\frac{k}{k-1}x_m, k>1$
$\boldsymbol{var(X)}$	$\frac{1}{\lambda^2}$	$\frac{2}{\lambda^2}$	$\lambda^{-\frac{2}{k}}[\Gamma(1+\frac{2}{\alpha})-(\Gamma(1+\frac{1}{\alpha}))^2]$	$\frac{k}{(k-2)(k-1)^2}x_m^2, k>2$

Table E.4. exponential families: absolutely continuous distributions, part 2

Distribution	Gamma	Inverse Gamma	Levy	Chi-sqare
$\boldsymbol{f_X(x)}$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1}\exp(-\frac{\beta}{x})$	$\sqrt{\frac{b}{2\pi}}\frac{1}{(x-\mu)^{\frac{3}{2}}}\exp(-\frac{b}{2(x-\mu}$	$\frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1}\exp-(\frac{x}{2})$
$\boldsymbol{x\in}$	$[0,\infty)$	$[0,\infty)$	$[\mu,\infty)$	$[0,\infty)$
parameter	β	β	b	k
$\boldsymbol{t(x)}$	$-x$	$-\frac{1}{x}$	$-\frac{1}{2(x-\mu)}$	$\log x$
$\boldsymbol{\theta}$	β	β	b	$\frac{k}{2}$
$\boldsymbol{F(\theta)}$	$-\alpha\log\theta$	$-\alpha\log\theta$	$-\frac{1}{2}\log\theta$	$\theta\log 2+\log\Gamma(\theta)$
$\boldsymbol{k(x)}$	$(\alpha-1)\log x-\log\Gamma(\alpha)$	$-(\alpha+1)\log x-\log\Gamma(\alpha)$	$-\frac{3}{2}\log(x-\mu)-\log\sqrt{2\pi}$	$-\log x-\frac{x}{2}$
$\boldsymbol{E(X)}$	$\frac{\alpha}{\beta}$	$\frac{\beta}{\alpha-1}, \alpha>1$	∞	k
$\boldsymbol{var(X)}$	$\frac{\alpha}{\beta^2}$	$\frac{\beta^2}{(\alpha-2)(\alpha-1)^2}, \alpha>2$	∞	$2k$

Table E.5. exponential families: absolutely continuous distributions, part 3

Distribution	Chi	Beta	Beta-prime	Gaussian
$\boldsymbol{f_X(x)}$	$\frac{1}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})} x^{k-1}\exp-(\frac{x^2}{2})$	$\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{1}{B(\alpha,\beta)}\frac{x^{\alpha-1}}{(1+x)^{\alpha+\beta}}$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{\lvert x-\mu\rvert^2}{2\sigma^2})$
$\boldsymbol{x\in}$	$[0,\infty)$	$[0,1]$	$[0,\infty)$	$\mathbb{R}$
parameter	k	(α,β)	(α,β)	(μ,σ)
$\boldsymbol{t(x)}$	$\log x$	$(\log x,\log(1-x))$	$(\log x,\log(1+x))$	$(-x^2,x)$
$\boldsymbol{\theta}$	k	(α,β)	$(\alpha,\alpha+\beta)$	$(\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2})$
$\boldsymbol{F(\theta)}$	$\frac{\theta}{2}\log 2+\log\Gamma(\frac{\theta}{2})$	$\log B(\theta_1,\theta_2)$	$\log B(\theta_1,\theta_1+\theta_2)$	$\frac{\theta_2^2}{4\theta_1}-\frac{1}{2}\log\theta_1$
$\boldsymbol{k(x)}$	$-\log(2x)-\frac{x^2}{2}$	$-\log(x(1-x))$	$-\log x$	$-\log\sqrt{2\pi}$
$\boldsymbol{E(X)}$	$\sqrt{2}\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha}{\beta-1}, \beta>1$	μ
$\boldsymbol{var(X)}$	$k-2[\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}]^2$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$	$\frac{\alpha(\alpha+\beta-1)}{(\beta-2)(\beta-1)^2}, \beta>2$	σ^2

Table E.6. exponential families: absolutely continuous distributions, part 4

Distribution	Log-normal	Rayleigh	Wald	Maxwell-Boltzmann
$\boldsymbol{f_X(x)}$	$\frac{1}{x\sqrt{2\pi\sigma^2}}\exp(-\frac{\lvert\log x-\mu\rvert^2}{2\sigma^2})$	$\frac{x}{\sigma^2}\exp(-\frac{x^2}{2\sigma^2})$	$\sqrt{\frac{\lambda}{2\pi x^3}}\exp(-\frac{\lambda\lvert x-\mu\rvert^2}{2x\mu^2})$	$\sqrt{\frac{2}{\pi}}\frac{x^2}{a^3}\exp(-\frac{x^2}{2a^2})$
$\boldsymbol{x\in}$	$[0,\infty)$	$[0,\infty)$	$[0,\infty)$	$[0,\infty)$
parameter	(μ,σ)	σ	(λ,μ)	a
$\boldsymbol{t(x)}$	$(-\log^2 x,\log x)$	$-x^2$	$(-\frac{1}{x},-x)$	$-\frac{x^2}{2}$
$\boldsymbol{\theta}$	$(\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2}-1)$	$\frac{1}{2\sigma^2}$	$(\frac{\lambda}{2},\frac{\lambda}{2\mu^2})$	$\frac{1}{a^2}$
$\boldsymbol{F(\theta)}$	$\frac{(\theta_2+1)^2}{4\theta_1}-\frac{1}{2}\log\theta_1$	$-\log(2\theta)$	$-\frac{\lambda}{\mu}-\frac{1}{2}\log\lambda$	$-\frac{3}{2}\log\theta$
$\boldsymbol{k(x)}$	$-\log\sqrt{2\pi}$	$\log x$	$-\frac{3}{2}\log x$	$2\log x+\log\sqrt{\frac{2}{\pi}}$
$\boldsymbol{E(X)}$	$\exp(\mu+\frac{\sigma^2}{2})$	$\sigma\sqrt{\frac{\pi}{2}}$	μ	$2a\sqrt{\frac{2}{\pi}}$
$\boldsymbol{var(X)}$	$(\exp(\sigma^2)-1)\exp(\sigma^2+2\mu)$	$(2-\frac{\pi}{2})\sigma^2$	$\frac{\mu^3}{\lambda}$	$a^2(3-\frac{8}{\pi})$

Table E.7. exponential families: absolutely continuous distributions, part 5

Distribution	Von Mises	U-power	Half normal	Inverse Chi square
$\boldsymbol{f_X(x)}$	$\frac{\exp(\kappa\cos(x-\mu))}{2\pi I_0(\kappa)}$	$\frac{2k+1}{2\alpha}(\frac{x-\mu}{\alpha})^{2k}$	$\sqrt{\frac{2}{\pi\sigma^2}}\exp(-\frac{x^2}{2\sigma^2})$	$\frac{2^{-\frac{k}{2}}}{\Gamma(\frac{k}{2})}x^{-\frac{k}{2}-1}\exp(-\frac{1}{2x})$
$\boldsymbol{x\in}$	$[\mu-\pi,\mu+\pi]$	$[\mu-\alpha,\mu+\alpha]$	$[0,\infty)$	$[0,\infty)$
parameter	(κ,μ)	k	σ	k
$\boldsymbol{t(x)}$	$(\cos x,\sin x)$	$2\log(x-\mu)$	$-x^2$	$-\log x$
$\boldsymbol{\theta}$	$(\kappa\cos\mu,\kappa\sin\mu)$	k	$\frac{1}{2\sigma^2}$	$\frac{k}{2}$
$\boldsymbol{F(\theta)}$	$\log I_0(\kappa)$	$2\theta\log\alpha-\log(\theta+\frac{1}{2})$	$-\log\theta-\frac{1}{2}\log 2$	$\frac{k}{2}\log 2+\log\Gamma(\frac{k}{2})$
$\boldsymbol{k(x)}$	$-\log(2\pi)$	$-\log\alpha$	$\log\sqrt{\frac{2}{\pi}}$	$-\log x-\frac{1}{2x}$
$\boldsymbol{E(X)}$	μ	μ	$\sigma\sqrt{\frac{2}{\pi}}$	$\frac{1}{k-2},k>2$
$\boldsymbol{var(X)}$	$1-\frac{I_1(\kappa)}{I_0(\kappa)}$	$\frac{2k+1}{2k+3}\alpha^2$	$(1-\frac{2}{\pi})\sigma^2$	$\frac{2}{(k-4)(k-2)^2},k>4$

Solutions to selected problems

Chapter 1:

1.3 For part 3, use $p^{n+1} = p^n p$ and the fact that p is a Markov matrix. For part 4 note that the relation derived in part 3 shows that the sequence is monotone.

1.4 Note that $\{T_1 \geq n, X_0 = 0\}$ means that the chain stays stuck at 0; then use relation (1.1) and the Markov property.

1.5 You should recognize a geometric distribution.

1.7 For part 2, note that we always have $X_{n+1} > X_n$ as long as 5 is not reached, so that $1 \leq T \leq 4$. For each value k we can make a list of the possible trajectories starting at 1 to evaluate $P(T = k|X_0 = 1)$. For part 4, we may remark that the relation derived in part 3 shows that (formally, i.e., without worrying about convergence) for $j \leq N$, the quantity $E(N_j|X_0 = 1)$ is the $(1, j)$-entry of the matrix

$$R := I + p + \cdots + p^k + \cdots .$$

The last row of p has a 1 at the end, so when we compute p^k, we find the same block structure, with Q^k as the upper right block; this shows that $R(1, j) = S(1, j)$ if we define the matrix S by

$$S := I + Q + \cdots + Q^k + \cdots ,$$

which means $S = (I - Q)^{-1}$. Finally, the quantity of question 5 is the waiting time before reaching the absorbing state.

1.12 As usual, the extreme values $k = 0$ and $k = d$ need to be treated separately; if k is not an extreme value, consider the probability of drawing from the left (or right box), and see how k may change. This gives

$$p(k, k-1) = \frac{k}{2d},\ p(k, k) = \frac{1}{2},\ p(k, k+1) = \frac{d-k}{2d}.$$

J.-F. Collet, *Discrete Stochastic Processes and Applications*,
Universitext, https://doi.org/10.1007/978-3-319-74018-8

Chapter 2:

2.1 Note that since $y \neq 0$, we have $xy \neq 0$, and compute the product xAy.

2.3 Show that when one computes the successive powers of A, once a positive element appears, it remains forever: if $A^m(i,j) > 0$, then $A^r(i,j) > 0$ for all $r \geq m$.

2.5 Using the explicit formula for the inverse matrix (in which the determinant appears as a denominator), show that a 2×2 monotone matrix is an M matrix if and only if its determinant is positive.

2.6

1. The column vector $(1,\ldots,1)^t$ should do the trick.
2. We know that for some vector $\bar{x} > 0$, the inequality $\bar{x} > A\bar{x}$ is satisfied; we can choose some $\lambda \in (0,1)$ such that $\lambda\bar{x} > A\bar{x}$ (note how the fact that the inequality is strict plays an essential role here). Then show that for every n we have $\lambda^n\bar{x} > A^n\bar{x}$, and deduce that $A^n\bar{x} \to 0$. Finally, use the fact that $\bar{x} > 0$.
3. Let $\bar{x} > 0$ be as above; pick x such that $(I - A)x \geq 0$. We need to show that $x \geq 0$. Argue by contradiction and consider the set

$$A := \{\alpha \in \mathbb{R} : \quad x + \alpha\bar{x} \geq 0\}.$$

   Show that A is an interval of the form $[\alpha_0, \infty)$ for some $\alpha_0 > 0$, and get a contradiction by showing that $x + \alpha_0\bar{x} > 0$.

2.7 Take a look at the justification of Theorem C.4 based on the Schur form of A.

Chapter 3:

3.2 If we use the method of test functions and compute $E(\phi(T_n))$, then Fubini's formula gives us a sequence of $n-1$ one-dimensional integrals, which may be computed sequentially.

3.3 Each jump of X being exactly of magnitude 1, you may use the equality of events

$$\{X_t \geq n\} = \{T_n \leq t\}$$

to compute $P(T_n \leq t)$ and then differentiate.

3.4 Use the equality of events $\{E_t < s\} = \{T_{X_t} > t - s\}$ and treat the cases $s \geq t$, $s < t$ separately to compute $P(E_t \leq s)$.

3.5 We want to check Definition 3.8; after having checked stationarity and independence of the increments, use total probability (conditioning on the result of α) to determine the distribution of Z_t.

3.6 Using the definition of conditional probability, independence, and stationarity of the increments, show that for $k \le n$,

$$P(X_s = k | X_t = n) = P(X_{t-s} = n-k) \frac{P(X_s = k)}{P(X_t = n)},$$

and then compute the right-hand side.

3.7 Use total probability (conditioning by the value of X_t), and to compute the sum, note that conditional on $X_t = n$, the variable X_t^1 has the binomial distribution $\mathcal{B}(n, p)$.

3.8 To find the correlation of X_t and X_s you may determine $E[(X_t - X_s)^2]$ and then deduce the value of $E[X_t X_s]$.

3.9 The expectation is computed using the law of total probability (conditioning by $N_t = n$); then the expectation of Z_t is obtained by taking the first derivative $\Phi'_{Z_t}(0)$.

3.10 The number of times X_t a fixed ball has changed boxes between 0 and t is $\mathcal{P}(\lambda t)$, and $p(t)$ is the probability that X_t is even. If all balls are initially in the left box, then $N(t) \sim \mathcal{B}(d, p(t))$; for the large-time asymptotics, note that $p(t) \to \frac{1}{2}$. For the general case in which the initial number of balls in the left box is j, we are going to distinguish among the balls present in the left box at time t those coming from the left from those coming from the right: $N(t) = N^l(t) + N^r(t)$, where $N^l(t) \sim \mathcal{B}(j, p(t))$ and $N^r(t) \sim \mathcal{B}(d-j, 1-p(t))$ are independent. In the limit $t \to \infty$, the limit is given by

$$\lim_{t\to\infty} p_k(t) = \frac{1}{2^d} \sum_{l=0}^{k} \binom{j}{l} \binom{d-j}{k-l} = \frac{1}{2^d} \binom{d}{k},$$

the second equality being a consequence of (B.3); so again the limiting distribution is $\mathcal{B}(d, \frac{1}{2})$.

3.11 For X_1^t use the equality of events

$$\{S_1^t > s, X_t = n\} = \{X_{t+s} = n, X_t = n\}.$$

For the third question, the strategy is to formulate everything in terms of the interarrival times S_i; note that $\{X_t = n\} = \{T_n \le t < T_n + S_{n+1}\}$, and

$$\{S_1^t > s_1, X_t = n\} = \{t + s_1 < T_n + S_{n+1}, X_t = n\}$$

(drawing a picture might help). Therefore, by independence of interarrival times, we have

$$\begin{aligned} &P(S_1^t > s_1, \dots, S_k^t > s_k, X_t = n\}) \\ &\quad = P(T_n \le t, t + s_1 < T_n + S_{n+1}) P(S_{n+2} > s_2) \cdots P(S_{n+k} > s_k). \end{aligned} \quad (1)$$

The first term on the right-hand side may be computed using the joint density of T_n and S_{n+1} (just as we did in the proof of Theorem 3.16), and the conclusion follows by summing over $n \in \mathbb{N}$.

Chapter 4:

4.2 For $s < t$, use transitions over the time intervals $[s+h, t+h]$ and $[s,t]$ to write

$$p_{ij}(t+h) - p_{ij}(t) = \sum_k [p_{ik}(s+h) - p_{ik}(s)] p_{kj}(t-s),$$

and then bound this quantity from above.

4.3 Note that over a given time interval $[0,t]$, the value of Y changes if and only if X_t is odd; use this to derive the transition matrix (on the space $\{-1,1\}$)

$$p_t = \frac{1}{2}\begin{pmatrix} 1+\exp(-2\lambda t) & 1-\exp(-2\lambda t) \\ 1-\exp(-2\lambda t) & 1+\exp(-2\lambda t) \end{pmatrix}.$$

4.4 Note that every 2×2 Markov matrix p_t is of the form chosen here; in the forward as in the backward case, there are only two independent differential equations; then by Corollary C.21, both problems are equivalent if and only if the matrices A and $p(0)$ commute, and this is exactly equivalent to the required condition. For a direct proof note that the first system implies (since v' is a multiple of u') that $v - \frac{\mu}{\lambda}u$ is a constant; in other words, we have

$$v = v_0 - \frac{\mu}{\lambda}u_0 + \frac{\mu}{\lambda}u,$$

and (after some basic manipulations) the first system assumes the following form:

$$\begin{cases} u'(t) = -(\lambda+\mu)u - \lambda v_0 + \mu u_0 + \lambda, \\ v'(t) = -(\lambda+\mu)v + \lambda v_0 - \mu u_0 + \mu, \end{cases}$$

which may then be compared to the second system.

4.5 For the last question use the equivalent $\ln(1+h) \sim h$ for h near 0.

4.6 The key point here is that although we have no uniformity condition (compare to Theorem A.8, for instance), the tail of the series may be expressed in terms of the whole sum and the truncated sum. To be more specific, let $|v|_\infty := \sup_n v_n$; writing

$$\sum_{n=0}^{N} u_n(t)v_n \le \sum_{n=0}^{\infty} u_n(t)v_n \le \sum_{n=0}^{N} u_n(t)v_n + |v|_\infty \left(\sum_{n=0}^{\infty} u_n(t) - \sum_{n=0}^{N} u_n(t)\right),$$

you may then let $t \to 0^+$ and conclude as in the proof of Theorem 4.18.

Chapter 5:

5.2 For the Taylor series of f, show by induction on k that for $k \geq 2$,

$$f^{(k)}(x) = -2^k[3 * 5 * \cdots * (2k-3)](1-4x)^{-\frac{2k-1}{2}}.$$

The formula for the quotient of factorials may be proved by induction, or (better) directly by noting the cancellation of all even factors in the numerator. Finally, just compute $\frac{1-f(x)}{2x}$; for completeness, you may even check continuity of the result at 0.

5.3 Begin by showing that if x, y are two integers with $x < y$, then

$$x!y! \geq (x+1)!(y-1)!;$$

then for the general case, use a "perturbation" argument: if some of the n_i are far from $[\frac{n}{d}]$, you may decrease the large ones and increase the small ones to produce a lesser value of the product. For the case $n = 2$, the result (about the uniform binomial distribution) says that the binomial coefficient $\binom{n}{k}$ is maximal when k is near $[\frac{n}{2}]$; note that you may also prove this fact by studying the monotonicity of the sequence $u_k := \binom{n}{k}$ for k between 0 and n. For the multinomial distribution it says that the most probable distribution is that in which all boxes contain (as close as possible to) the same number of elements.

5.4 This is simply the probability p that a path drawn uniformly on the set of $2n$ loops is a simple loop; after simplification one obtains $p = \frac{1}{2n-1}$.

5.6 To recover the Poisson process, note that if $\mu = 0$, the ordinary differential equation is immediately integrated to give

$$\phi(z,t) = e^{\lambda(z-1)t}\phi(z,0).$$

To obtain $p_t(j,\cdot)$ we take the initial data $u_i(0) = \delta_{ij}$; thus $\phi(z,0) = z^j$. We may then expand $\phi(z,t)$ in powers of z:

$$\phi(z,t) = z^j e^{\lambda(z-1)t} = e^{-\lambda t}\sum_{l=0}^{\infty}\frac{(\lambda t)^l}{l!}z^{l+j} = e^{-\lambda t}\sum_{k=j}^{\infty}\frac{(\lambda t)^{k-j}}{(k-j)!}z^k.$$

By identification of the general term in ϕ this gives the expression for $u_k(t) = p_t(j,k)$.

Chapter 6:

6.1 Parametrize the segment $[x, y]$ and apply the one-variable formula.

6.3 Express $Ap \cdot p$ as a double sum and symmetrize it. In other words, write

$$A = \sum_i \sum_j w_{ij} = \sum_i \sum_j w_{ji} = \frac{1}{2} \sum_i \sum_j [w_{ij} + w_{ji}].$$

6.4 Showing that the lower bound obtained in Problem 6.3 is sharper (which means larger) amounts to showing that for all P, Q positive, one has

$$P \log \frac{P}{Q} \geq P - Q.$$

6.6 Use the same symmetrizing trick as in Problem 6.3.

6.8 The relation between $\det A$ and $\det C$ comes from multilinearity of the determinant (spot common factors in each row and each column of A); to get the relation between D_n and D_{n-1}, perform a row manipulation on D_n. This will give an expression of $D_n + D_{n-1}$ as a determinant; perform a row manipulation on this determinant to get the result.

6.9 The symmetrizing trick of Problem 6.3, yet again!

6.13 To obtain (6.23), apply (6.22) to $u - \int u\phi$ and $v - \int v\phi$; then put $\phi := \frac{b}{B}$ in (6.23) to obtain (6.24).

Chapter 7:

7.3 For part 2, use the representation of P as a convex combination of permutation matrices and the fact that H is concave; for part 3, take p uniform; for part 4, use inequality (6.19) with exponents $P(x, y)$.

7.7 For the first question, simply apply Jensen's inequality to the given map, writing $\frac{q_i}{p_i} = \exp(-\ln \frac{p_i}{q_i})$; to get $D_{KL}(p, q)$ large, take, for instance, two Bernoulli variables of parameters p and q, with p fixed and $q \to 0$; finally, the absolute bound for D_{JS} is an immediate consequence of $D_J(p, q) \geq 0$.

7.8 Express $H(X_1, \ldots, X_n)$ by leaving X_n aside and conditioning; then use Cesaro's lemma.

7.10 The first relation is immediate from the Fokker–Planck equation; for the second one, use the following summation by parts formula:

$$\sum_{k=1}^{\infty} a_k(b_{k+1} - b_k) = -a_1 b_1 - \sum_{k=1}^{\infty} b_{k+1}(a_{k+1} - a_k).$$

Chapter 8:

8.6 For part 3, note that z^x increases with x if and only if $z > 1$; thus we are interested in having $\frac{r}{1-r} < 1$, which means $r < \frac{1}{2}$.

References

1. Emil Artin. 2015. *The Gamma function*. Mineola: Courier Dover Publications.
2. Robert B Ash. 1990. *Information theory*. New York: Dover publications.
3. Monica Bianchini, Marco Gori, and Franco Scarselli. 2005. Inside PageRank. *ACM Transactions on Internet Technology (TOIT)* 5 (1): 92–128.
4. Leo Breiman. 1992. *Probability, volume 7 of classics in applied mathematics*. Philadelphia: SIAM.
5. Pierre Brémaud. 2013. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, vol. 31. Berlin: Springer Science & Business Media.
6. Richard A Brualdi. 1982. Matrices eigenvalues, and directed graphs. *Linear and Multilinear Algebra* 11 (2): 143–165.
7. Thomas M Cover, and Joy A Thomas. 2012. *Elements of information theory*. New York: Wiley.
8. GE Crooks. 2008. Inequalities between the Jensen–Shannon and Jeffreys divergences, tech. Technical report, Note 004, 2008. http://threeplusone.com/pubs/technote/CrooksTechNote004.pdf.
9. N Dmitriev, and E Dynkin. 1945. On the characteristic numbers of a stochastic matrix cr acad. *URSS (NS)* 49: 159–162.
10. Nikolai Aleksandrovich Dmitriev, and E Dynkin. 1946. On characteristic roots of stochastic matrices. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya* 10 (2): 167–184.
11. Richard Durrett. 2012. *Essentials of stochastic processes*. Berlin: Springer Science & Business Media.
12. Georg Ferdinand Frobenius. 1912. *Über Matrizen aus nicht negativen Elementen*. Königliche Akademie der Wissenschaften.
13. Jean-Baptiste Hiriart-Urruty, and Claude Lemaréchal. 2012. *Fundamentals of convex analysis*. Berlin: Springer Science & Business Media.
14. Morris W Hirsch, Stephen Smale, and Robert L Devaney. 2012. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press.

J.-F. Collet, *Discrete Stochastic Processes and Applications*,
Universitext, https://doi.org/10.1007/978-3-319-74018-8

15. Roger A Horn, and Charles R Johnson. 2012. *Matrix analysis*. Cambridge: Cambridge university press.
16. Gareth A Jones, and J Mary Jones. 2012. *Information and coding theory*. Berlin: Springer Science & Business Media.
17. Amy N Langville, and Carl D Meyer. 2004. Deeper inside PageRank. *Internet Mathematics* 1 (3): 335–380.
18. Amy N Langville, and Carl D Meyer. 2011. *Google's PageRank and beyond: The science of search engine rankings*. Princeton: Princeton University Press.
19. Gregory F Lawler, and Vlada Limic. 2010. *Random walk: a modern introduction*, vol. 123. Cambridge: Cambridge University Press.
20. Erich L Lehmann, and Joseph P Romano. 2006. *Testing statistical hypotheses*. Springer: Springer Science & Business Media.
21. Erich Leo Lehmann, and George Casella. 1998. *Theory of point estimation*, vol. 31. Berlin: Springer Science & Business Media.
22. David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge: Cambridge university press.
23. Marvin Marcus, and Henryk Minc. 1992. *A survey of matrix theory and matrix inequalities*, vol. 14. Courier Corporation.
24. Robert McEliece. 2002. *The theory of information and coding*. Cambridge: Cambridge University Press.
25. Carl D Meyer. 2000. *Matrix analysis and applied linear algebra*, vol. 2. Philadelphia: Siam.
26. James R Norris. 1998. *Markov chains*. Cambridge: Cambridge university press.
27. Fritz Oberhettinger, and Larry Badii. 2012. *Tables of Laplace transforms*. Berlin: Springer Science & Business Media.
28. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.
29. Oskar Perron. 1907. Zur theorie der matrices. *Mathematische Annalen* 64 (2): 248–263.
30. Edward M Reingold. 2000. Mathematical entertainments: Cliques, the Cauchy inequality, and information theory (department). *The Mathematical intelligencer* 22 (4): 14–15.
31. Ralph Tyrell Rockafellar. 2015. *Convex analysis*. Princeton: Princeton university press.
32. Walter Rudin. 1991. *Functional analysis*, International series in pure and applied mathematics. New York: McGraw-Hill Inc.
33. Joel L Schiff. 2013. *The Laplace Transform*, Undergraduate Texts in Mathematics. New York: Springer.
34. Richard P Stanley. 1999. *Enumerative combinatorics*, vol. 2, vol. 62 of cambridge studies in advanced mathematics. Cambridge: Cambridge University Press.

35. Richard S Varga. 2001. Gerschgorin disks, Brauer ovals of Cassini (a vindication), and Brualdi sets. *Information-Yamaguchi* 4 (2): 171–178.
36. Tomasz Zastawniak, and Zdzislaw Brzezniak. 2003. *Basic stochastic processes*. Berlin: Springer.

Index

J.-F. Collet, *Discrete Stochastic Processes and Applications*,
Universitext, https://doi.org/10.1007/978-3-319-74018-8