Appendix A

Useful combinatorial formulas

Recall the following facts about factorials and binomial coefficients:

\[
\frac{n^n}{e^{n-1}} \leq n! \leq \frac{n^{n+1}}{e^{n-1}},
\]

\[
\frac{n^k}{k^k} \leq \binom{n}{k} \leq \frac{e^{k^k}}{k^k},
\]

\[
(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k},
\]

\[
\sum_{k=0}^{d} \binom{n}{k} \leq \left( \frac{en}{d} \right)^d,
\]

\[
n! \sim \sqrt{2\pi n} \left( \frac{n}{e} \right)^n,
\]

\[
\binom{2n}{n} = (1 + o(1)) \frac{4^n}{\sqrt{\pi n}},
\]

and

\[
\log \binom{n}{k} = (1 + o(1))nH(k/n),
\]

where \( H(p) := -p \log p - (1 - p) \log(1 - p) \). The third one is the binomial theorem. The fifth one is Stirling’s formula.
Appendix B

Measure-theoretic foundations

This appendix contains relevant background on measure-theoretic probability. We follow closely the highly recommended [Wil91]. Missing proofs (and a lot more details and examples) can be found there. Another excellent textbook on this topic is [Dur10].

B.1 Probability spaces

Let $S$ be a set. In general it turns out that we cannot assign a probability to every subset of $S$. Here we discuss “well-behaved” collections of subsets. First an algebra on $S$ is a collection of subsets stable under finitely many set operations.

**Definition B.1.1** (Algebra on $S$). A collection $\Sigma_0$ of subsets of $S$ is an algebra on $S$ if the following conditions hold:

(i) $S \in \Sigma_0$;

(ii) $F \in \Sigma_0$ implies $F^c \in \Sigma_0$;

(iii) $F, G \in \Sigma_0$ implies $F \cup G \in \Sigma_0$.

That, of course, implies that the empty set as well as all pairwise intersections are also in $\Sigma_0$. The collection $\Sigma_0$ is an actual algebra (i.e., a vector space with a bilinear product) with the symmetric difference as its “sum,” the intersection as its “product” and the underlying field being the field with two elements.
Example B.1.2. On $\mathbb{R}$, sets of the form

$$\bigcup_{i=1}^{k} (a_i, b_i]$$

where the union is disjoint with $k < +\infty$ and $-\infty \leq a_i \leq b_i \leq +\infty$ form an algebra.

Finite set operations are not enough for our purposes. For instance, we want to be able to take limits. A $\sigma$-algebra is stable under countably many set operations.

**Definition B.1.3 ($\sigma$-algebra on $S$).** A collection $\Sigma$ of subsets of $S$ is a $\sigma$-algebra on $S$ (or $\sigma$-field on $S$) if

(i) $S \in \Sigma$;
(ii) $F \in \Sigma$ implies $F^c \in \Sigma$;
(iii) $F_n \in \Sigma, \forall n$ implies $\bigcup_{n} F_n \in \Sigma$.

**Example B.1.4.** $2^S$ is a trivial example.

To give a nontrivial example, we need the following definition. We begin with a lemma.

**Lemma B.1.5 (Intersection of $\sigma$-algebras).** Let $\mathcal{F}_i, i \in I$, be $\sigma$-algebras on $S$ where $I$ is arbitrary. Then $\bigcap_i \mathcal{F}_i$ is a $\sigma$-algebra.

**Proof.** We prove only one of the conditions. The other ones are similar. Suppose $A \in \mathcal{F}_i$ for all $i$. Then $A^c$ is in $\mathcal{F}_i$ for all $i$ since each $\mathcal{F}_i$ is itself a $\sigma$-algebra.

**Definition B.1.6 ($\sigma$-algebra generated by $C$).** Let $\mathcal{C}$ be a collection of subsets of $S$. Then we let $\sigma(\mathcal{C})$ be the smallest $\sigma$-algebra containing $\mathcal{C}$, defined as the intersection of all such $\sigma$-algebras (including in particular $2^S$).

**Example B.1.7.** The smallest $\sigma$-algebra containing all open sets in $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is called the Borel $\sigma$-algebra. This is a non-trivial $\sigma$-algebra in the sense that it can be proved that there exist subsets of $\mathbb{R}$ that are not in $\mathcal{B}$, but that any “reasonable” set is in $\mathcal{B}$. In particular, it contains the algebra in Example B.1.2.

**Example B.1.8.** The $\sigma$-algebra generated by the algebra in Example B.1.2 is $\mathcal{B}(\mathbb{R})$. This follows from the fact that all open sets of $\mathbb{R}$ can be written as a countable union of open intervals. (Indeed, for $x \in O$ an open set, let $I_x$ be the largest open interval contained in $O$ and containing $x$. If $I_x \cap I_y \neq \emptyset$ then $I_x = I_y$ by maximality (i.e., take the union). Then $O = \bigcup_x I_x$ and there are only countably many disjoint ones because each one contains a rational.)
We now define measures.

**Definition B.1.9** (Additivity and $\sigma$-additivity). A non-negative set function on an algebra $\Sigma_0$

\[ \mu_0 : \Sigma_0 \to [0, +\infty], \]

is additive if

(i) $\mu_0(\emptyset) = 0$;

(ii) $F, G \in \Sigma_0$ with $F \cap G = \emptyset$ implies $\mu_0(F \cup G) = \mu_0(F) + \mu_0(G)$.

Moreover $\mu_0$ is said to be $\sigma$-additive if condition (ii) is true for any countable collection of disjoint sets whose union is in $\Sigma_0$, that is, if $F_n \in \Sigma_0$, $n \geq 0$, all pairwise disjoint with $\cup_n F_n \in \Sigma_0$, then $\mu_0(\cup_n F_n) = \sum_n \mu_0(F_n)$.

**Example B.1.10.** For the algebra in the Example B.1.2, the set function

\[ \lambda_0 \left( \bigcup_{i=1}^{k} [a_i, b_i] \right) = \sum_{i=1}^{k} (b_i - a_i) \]

is additive. (In fact, it is also $\sigma$-additive. We will show this later.)

**Definition B.1.11** (Measure space). Let $\Sigma$ be a $\sigma$-algebra on $S$. Then $(S, \Sigma)$ is a measurable space. A $\sigma$-additive function $\mu$ on $\Sigma$ is called a measure and $(S, \Sigma, \mu)$ is called a measure space.

**Definition B.1.12** (Probability space). If $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure space with $\mathbb{P}(\Omega) = 1$ then $\mathbb{P}$ is called a probability measure and $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space (or probability triple).

The sets in $\mathcal{F}$ are referred to as events.

To define a measure on $B(\mathbb{R})$ we need the following tools from abstract measure theory.

**Theorem B.1.13** (Carathéodory’s extension theorem). Let $\Sigma_0$ be an algebra on $S$ and let $\Sigma = \sigma(\Sigma_0)$. If $\mu_0$ is $\sigma$-additive on $\Sigma_0$ then there exists a measure $\mu$ on $\Sigma$ that agrees with $\mu_0$ on $\Sigma_0$.

If in addition $\mu_0$ is finite, the next lemma implies that the extension is unique.

**Lemma B.1.14** (Uniqueness of extensions). Let $\mathcal{I}$ be a $\pi$-system on $S$, that is, a family of subsets closed under finite intersections, and let $\Sigma = \sigma(\mathcal{I})$. If $\mu_1, \mu_2$ are finite measures on $(S, \Sigma)$ that agree on $\mathcal{I}$, then they agree on $\Sigma$. 

```
Example B.1.15. The sets \((-\infty, x]\) for \(x \in \mathbb{R}\) form a \(\pi\)-system generating \(\mathcal{B}(\mathbb{R})\). That is, \(\mathcal{B}(\mathbb{R})\) is the smallest \(\sigma\)-algebra containing that \(\pi\)-system.

Finally we can define Lebesgue measure. We start with \((0, 1]\) and extend to \(\mathbb{R}\) in the obvious way. We need the following lemma.

Lemma B.1.16 (\(\sigma\)-additivity of \(\lambda_0\)). Let \(\lambda_0\) be the set function defined above in Example B.1.10, restricted to \((0, 1]\). Then \(\lambda_0\) is \(\sigma\)-additive.

Definition B.1.17 (Lebesgue measure on unit interval). The unique extension of \(\lambda_0\) (see Example B.1.10) to \((0, 1]\) is denoted \(\lambda\) and is called Lebesgue measure.

B.2 Random variables

Let \((S, \Sigma, \mu)\) be a measure space and let \(B = \mathcal{B}(\mathbb{R})\).

Definition B.2.1 (Measurable function). Suppose \(h : S \to \mathbb{R}\) and define
\[
h^{-1}(A) = \{s \in S : h(s) \in A\}.
\]

The function \(h\) is \(\Sigma\)-measurable if \(h^{-1}(B) \in \Sigma\) for all \(B \in \mathcal{B}\). We denote by \(m\Sigma\) (resp., \((m\Sigma)^+, b\Sigma\)) the \(\Sigma\)-measurable functions (resp., that are non-negative, bounded).

In the probabilistic case:

Definition B.2.2. A random variable is a measurable function on a probability space \((\Omega, \mathcal{F}, \mathbb{P})\).

The behavior of a random variable is characterized by its distribution function.

Definition B.2.3 (Distribution function). Let \(X\) be a random variable on a probability space \((\Omega, \mathcal{F}, \mathbb{P})\). The law of \(X\) is
\[
\mathcal{L}_X = \mathbb{P} \circ X^{-1},
\]
which is a probability measure on \((\mathbb{R}, \mathcal{B})\). By Lemma B.1.14, \(\mathcal{L}_X\) is determined by the distribution function (DF) of \(X\)
\[
F_X(x) = \mathbb{P}[X \leq x], \quad x \in \mathbb{R}.
\]
Example B.2.4. The distribution function of a constant random variable is a jump of size 1 at the value it takes almost surely. The distribution function of a random variable with law equal to Lebesgue measure on $(0, 1]$ is

$$F_X(x) = \begin{cases} x & x \in (0, 1], \\ 0 & x \leq 0, \\ 1 & x > 1. \end{cases}$$

We refer to such a random variable as a **uniform random variable** over $(0, 1]$. ▲

Distribution functions are characterized by a few simple properties.

**Proposition B.2.5.** Suppose $F = F_X$ is the distribution function of a random variable $X$ on $(\Omega, \mathcal{F}, \mathbb{P})$. Then the following hold:

(i) $F$ is non-decreasing;

(ii) $\lim_{x \to +\infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$;

(iii) $F$ is right-continuous.

**Proof.** The first property follows from the monotonicity of probability measure (which itself follows immediately from $\sigma$-additivity).

For the second property, note that the limit exists by the first property. The value of the limit follows from the following important lemma.

**Lemma B.2.6 (Monotone convergence properties of measures).** Let $(S, \Sigma, \mu)$ be a measure space.

(i) If $F_n \in \Sigma$, $n \geq 1$, with $F_n \uparrow F$, then $\mu(F_n) \uparrow \mu(F)$.

(ii) If $G_n \in \Sigma$, $n \geq 1$, with $G_n \downarrow G$ and $\mu(G_k) < +\infty$ for some $k$, then $\mu(G_n) \downarrow \mu(G)$.

**Proof.** Clearly $F = \bigcup_n F_n \in \Sigma$. For $n \geq 1$, write $H_n = F_n \setminus F_{n-1}$ (with $F_0 = \emptyset$). Then by disjointness

$$\mu(F_n) = \sum_{k \leq n} \mu(H_k) \uparrow \sum_{k < +\infty} \mu(H_k) = \mu(F).$$

The second statement is similar. □

Similarly, for the third property, by Lemma B.2.6 again

$$\mathbb{P}[X \leq x_n] \downarrow \mathbb{P}[X \leq x],$$

if $x_n \downarrow x$. □
It turns out that the properties above characterize distribution functions in the following sense.

**Theorem B.2.7** (Skorokhod representation). Let $F$ satisfy the three properties above in Proposition B.2.5. Then there is a random variable $X$ on

$$(\Omega, \mathcal{F}, \mathbb{P}) = ((0, 1], \mathcal{B}(0, 1], \lambda),$$

with distribution function $F$. The law of $X$ is called the Lebesgue-Stieltjes measure associated to $F$.

The result says that all real random variables can be generated from uniform random variables over $(0, 1]$.

**Proof.** Assume first that $F$ is continuous and strictly increasing. Define $X(\omega) = F^{-1}(\omega)$ for all $\omega \in \Omega$. Then, $\forall x \in \mathbb{R},$

$$\mathbb{P}[X \leq x] = \mathbb{P}[\{\omega : F^{-1}(\omega) \leq x\}] = \mathbb{P}[\{\omega : \omega \leq F(x)\}] = F(x).$$

In general, let

$$X(\omega) = \inf\{x : F(x) \geq \omega\}.$$

It suffices to prove that

$$X(\omega) \leq x \iff \omega \leq F(x).$$

The $\iff$ direction is clear by definition of $X$. On the other hand, by the right-continuity of $F$, we have that $\omega \leq F(X(\omega))$. Therefore, by monotonicity of $F$,

$$X(\omega) \leq x \implies \omega \leq F(X(\omega)) \leq F(x).$$

That proves the claim.

Turning measurability on its head, we get the following important definition.

**Definition B.2.8.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $Y_\gamma, \gamma \in \Gamma$, be a collection of maps from $\Omega$ to $\mathbb{R}$. We let

$$\sigma(Y_\gamma, \gamma \in \Gamma),$$

be the smallest $\sigma$-algebra on which the $Y_\gamma$’s are measurable.

In a sense, the above $\sigma$-algebra corresponds to “the partial information available when the $Y_\gamma$’s are observed.”
Example B.2.9. Suppose we flip two unbiased coins and let $X$ be the number of heads observed. Then, denoting heads by $H$ and tails by $T$,

$$
\sigma(X) = \sigma(\{\{HH\}, \{HT, TH\}, \{TT\}\}),
$$

which is coarser than the full $\sigma$-algebra $2^\Omega$.

Note that $h^{-1}$ preserves all set operations. For example, $h^{-1}(A \cup B) = h^{-1}(A) \cup h^{-1}(B)$. This gives the following important lemma.

Lemma B.2.10 (Sufficient condition for measurability). Suppose $C \subseteq B$ with $\sigma(C) = B$. Then $h^{-1} : C \to \Sigma$ implies $h \in m\Sigma$. That is, it suffices to check measurability on a collection generating $B$.

Proof. Let $\mathcal{E}$ be the sets such that $h^{-1}(B) \in \Sigma$. By the observation before the statement, $\mathcal{E}$ is a $\sigma$-algebra. But $C \subseteq \mathcal{E}$ which implies $\sigma(C) \subseteq \mathcal{E}$ by minimality.

As a consequence we get the following properties of measurable functions.

Proposition B.2.11 (Properties of measurable functions). Let $h, h_n, n \geq 1$, be in $m\Sigma$ and $f \in mB$.

(i) $f \circ h \in m\Sigma$.

(ii) If $S$ is a topological space and $h$ is continuous, then $h$ is $\mathcal{B}(S)$-measurable, where $\mathcal{B}(S)$ is generated by the open sets of $S$.

(iii) The function $g : S \to \mathbb{R}$ is in $m\Sigma$ if for all $c \in \mathbb{R}$, 

$$
\{g \leq c\} \in \Sigma.
$$

(iv) $\forall \alpha \in \mathbb{R}, h_1 + h_2, h_1 h_2, \alpha h \in m\Sigma$.

(v) $\inf h_n, \sup h_n, \lim \inf h_n, \lim \sup h_n$ are in $m\Sigma$.

(vi) The set 

$$
\{ s : \lim h_n(s) \text{ exists in } \mathbb{R}\},
$$

is measurable.

Proof. We sketch the proof of a few of them.

(ii) This follows from Lemma B.2.10 by taking $C$ as the open sets of $\mathbb{R}$.

(iii) Similarly, take $C$ to be the sets of the form $(-\infty, c]$. 

APPENDIX B. MEASURE-THEORETIC FOUNDATIONS

(iv) This follows from (iii). For example note that, writing the left-hand side as $h_1 > c - h_2$,

$$\{h_1 + h_2 > c\} = \cup_{q \in \mathbb{Q}} \{ \{h_1 > q\} \cap \{q > c - h_2\} \},$$

which is a countable union of measurable sets by assumption.

(v) Note that

$$\{\sup h_n \leq c\} = \cap_n \{h_n \leq c\}.$$

Further, note that $\lim \inf$ is the $\sup$ of an $\inf$. ■

B.3 Independence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

**Definition B.3.1 (Independence).** Sub-$\sigma$-algebras $\mathcal{G}_1, \mathcal{G}_2, \ldots$ of $\mathcal{F}$ are independent if: for all $G_i \in \mathcal{G}_i$, $i \geq 1$, and distinct $i_1, \ldots, i_n$ we have

$$\mathbb{P}[G_{i_1} \cap \cdots \cap G_{i_n}] = \prod_{j=1}^{n} \mathbb{P}[G_{i_j}].$$

Specializing to events and random variables:

**Definition B.3.2 (Independent random variables).** Random variables $X_1, X_2, \ldots$ are independent if the $\sigma$-algebras $\sigma(X_1), \sigma(X_2), \ldots$ are independent.

**Definition B.3.3 (Independent events).** Events $E_1, E_2, \ldots$ are independent if the $\sigma$-algebras

$$\mathcal{E}_i = \{\emptyset, E_i, E_i^c, \Omega\}, \quad i \geq 1,$$

are independent.

Recall the more familiar definitions.

**Theorem B.3.4 (Independent random variables: familiar definition).** Random variables $X, Y$ are independent if and only if for all $x, y \in \mathbb{R}$

$$\mathbb{P}[X \leq x, Y \leq y] = \mathbb{P}[X \leq x] \mathbb{P}[Y \leq y].$$

**Theorem B.3.5 (Independent events: familiar definition).** Events $E_1, E_2$ are independent if and only if

$$\mathbb{P}[E_1 \cap E_2] = \mathbb{P}[E_1] \mathbb{P}[E_2].$$
The proofs of these characterizations follow immediately from the following lemma.

**Lemma B.3.6** (Independence and $\pi$-systems). Suppose that $\mathcal{G}$ and $\mathcal{H}$ are sub-$\sigma$-algebras and that $\mathcal{I}$ and $\mathcal{J}$ are $\pi$-systems such that

$$\sigma(\mathcal{I}) = \mathcal{G}, \quad \sigma(\mathcal{J}) = \mathcal{H}.$$ 

Then $\mathcal{G}$ and $\mathcal{H}$ are independent if and only if $\mathcal{I}$ and $\mathcal{J}$ are as well, that is,

$$P[I \cap J] = P[I] P[J], \quad \forall I \in \mathcal{I}, J \in \mathcal{J}.$$ 

**Proof.** Suppose $\mathcal{I}$ and $\mathcal{J}$ are independent. For fixed $I \in \mathcal{I}$, the measures $P[I \cap H]$ and $P[I] P[H]$ are equal for $H \in \mathcal{J}$ and have total mass $P[I] < +\infty$. By the Uniqueness of Extensions Lemma (Lemma B.1.14) the above measures agree on $\sigma(\mathcal{J}) = \mathcal{H}$.

Repeat the argument. Fix $H \in \mathcal{H}$. Then the measures $P[G \cap H]$ and $P[G] P[H]$ agree on $\mathcal{I}$ and have total mass $P[H] < +\infty$. Therefore they must agree on $\sigma(\mathcal{I}) = \mathcal{G}$.

We give a standard construction of an infinite sequence of independent random variables with prescribed distributions.

Let $(\Omega, \mathcal{F}, P) = ((0, 1], \mathcal{B}(0, 1], \lambda)$ and for $\omega \in \Omega$ consider the binary expansion

$$\omega = 0.\omega_1\omega_2\ldots.$$ 

(For dyadic rationals, use the all-1 ending and note that the dyadic rationals have measure 0 by countability.) This construction produces a sequence of independent so-called *Bernoulli trials*. That is, under $\lambda$, each bit is Bernoulli(1/2) and any finite collection is independent.

To get two independent uniform random variables, consider the following construction:

$$U_1 = 0.\omega_1\omega_3\omega_5\ldots$$

$$U_2 = 0.\omega_2\omega_4\omega_6\ldots$$

Let $\mathcal{A}_1$ (resp. $\mathcal{A}_2$) be the $\pi$-system consisting of all finite intersections of events of the form $\{\omega_i \in H_i\}$ for odd $i$ (resp. even $i$). By Lemma B.3.6, the $\sigma$-fields $\sigma(\mathcal{A}_1)$ and $\sigma(\mathcal{A}_2)$ are independent.

More generally, let

$$V_1 = 0.\omega_1\omega_3\omega_5\ldots$$

$$V_2 = 0.\omega_2\omega_4\omega_6\ldots$$

$$V_3 = 0.\omega_4\omega_8\omega_{13}\ldots$$

$$\vdots$$
that is, fill up the array diagonally. By the argument above, the $V_i$’s are independent and Bernoulli($1/2$).

Finally let $\mu_n$, $n \geq 1$, be a sequence of probability measures with distribution functions $F_n$, $n \geq 1$. For each $n$, define

$$X_n(\omega) = \inf\{x : F_n(x) \geq V_n(\omega)\}$$

By the (proof of the) Skorokhod Representation (Theorem B.2.7), $X_n$ has distribution function $F_n$.

**Definition B.3.7** (I.i.d. random variables). A sequence of independent random variables $(X_n)$ as above is independent and identically distributed (i.i.d.) if $F_n = F$ for some $n$.

Alternatively, we have the following more general result.

**Theorem B.3.8** (Kolmogorov’s extension theorem). Suppose we are given probability measures $\mu_n$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ that are consistent, that is,

$$\mu_{n+1}((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R}) = \mu_n((a_1, b_1] \times \cdots \times (a_n, b_n]).$$

Then there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}^N, \mathcal{R}^N)$ with

$$\mathbb{P}[\omega : \omega_i \in (a_i, b_i], 1 \leq i \leq n] = \mu_n((a_1, b_1] \times \cdots \times (a_n, b_n]).$$

Here $\mathcal{R}^N$ is the product $\sigma$-algebra, that is, the $\sigma$-algebra generated by finite-dimensional rectangles.

Next, we discuss a first non-trivial result about independent sequences.

**Definition B.3.9** (Tail $\sigma$-algebra). Let $X_1, X_2, \ldots$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define

$$\mathcal{T} = \bigcap_{n \geq 1} \mathcal{T}_n,$$

where

$$\mathcal{T}_n = \sigma(X_{n+1}, X_{n+2}, \ldots).$$

As an intersection of $\sigma$-algebras, $\mathcal{T}$ is a $\sigma$-algebra. It is called the tail $\sigma$-algebra of the sequence $(X_n)$.

Intuitively, an event is in the tail if changing a finite number of values does not affect its occurrence.
Example B.3.10. If $S_n = \sum_{k \leq n} X_k$, then
\[
\{ \lim_{n} S_n \text{ exists} \} \in \mathcal{T},
\]
\[
\{ \limsup_{n} n^{-1} S_n > 0 \} \in \mathcal{T},
\]
but
\[
\{ \limsup_{n} S_n > 0 \} \notin \mathcal{T}.
\]

Theorem B.3.11 (Kolmogorov’s 0-1 law). Let $(X_n)$ be a sequence of independent random variables with tail $\sigma$-algebra $\mathcal{T}$. Then $\mathcal{T}$ is $\mathbb{P}$-trivial, that is, for all $A \in \mathcal{T}$ we have either $\mathbb{P}[A] = 0$ or 1.

Proof. Let $\mathcal{X}_n = \sigma(X_1, \ldots, X_n)$. Note that $\mathcal{X}_n$ and $\mathcal{T}_n$ are independent. Moreover, since $\mathcal{T} \subseteq \mathcal{T}_n$ we have that $\mathcal{X}_n$ is independent of $\mathcal{T}_n$. Now let
\[
\mathcal{X}_\infty = \sigma(X_n, n \geq 1).
\]
Note that
\[
\mathcal{K}_\infty = \bigcup_{n \geq 1} \mathcal{X}_n,
\]
is a $\pi$-system generating $\mathcal{X}_\infty$. Therefore, by Lemma B.3.6, $\mathcal{X}_\infty$ is independent of $\mathcal{T}$. But $\mathcal{T} \subseteq \mathcal{X}_\infty$ and therefore $\mathcal{T}$ is independent of itself! Hence if $A \in \mathcal{T}$,
\[
\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A]^2,
\]
which can occur only if $\mathbb{P}[A] \in \{0, 1\}$. □

B.4 Expectation

Let $(S, \Sigma, \mu)$ be a measure space. We denote by $1_A$ the indicator of a set $A$, that is,
\[
1_A(s) = \begin{cases} 
1, & \text{if } s \in A \\
0, & \text{otherwise}
\end{cases}
\]

Definition B.4.1 (Simple functions). A simple function is a function of the form
\[
f = \sum_{k=1}^{m} a_k 1_{A_k},
\]
where \(a_k \in [0, +\infty]\) and \(A_k \in \Sigma\) for all \(k\). We denote the set of all such functions by \(SF^+\). We define the integral of \(f\) by

\[
\mu(f) := \sum_{k=1}^{m} a_k \mu(A_k) \leq +\infty.
\]

We also write \(\mu f = \mu(f)\).

The following is left as a (somewhat tedious but) immediate exercise.

**Proposition B.4.2.** Let \(f, g \in SF^+\).

(i) If \(\mu(f \neq g) = 0\), then \(\mu f = \mu g\). [Hint: Rewrite \(f\) and \(g\) over the same disjoint sets.]

(ii) For all \(c \geq 0\), \(f + g, cf \in SF^+\) and

\[
\mu(f + g) = \mu f + \mu g, \quad \mu(cf) = c\mu f.
\]

[Hint: This one is obvious by definition.]

(iii) If \(f \leq g\) then \(\mu f \leq \mu g\). [Hint: Show that \(g - f \in SF^+\) and use linearity.]

The main definition and theorem of integration theory follows.

**Definition B.4.3** (Nonnegative functions). Let \(f \in (m \Sigma)^+\). Then the integral of \(f\) is defined by

\[
\mu(f) = \sup\{\mu(h) : h \in SF^+, h \leq f\}.
\]

Again we also write \(\mu f = \mu(f)\).

**Theorem B.4.4** (Monotone convergence theorem). If \(f_n, f \in (m \Sigma)^+, n \geq 1\), with \(f_n \uparrow f\), then

\[
\mu f_n \uparrow \mu f.
\]

Many theorems in integration follow from the monotone convergence theorem. In that context, the following approximation is useful.

**Definition B.4.5** (Staircase function). For \(f \in (m \Sigma)^+\) and \(r \geq 1\), the \(r\)-th staircase function \(\alpha^{(r)}\) is

\[
\alpha^{(r)}(x) = \begin{cases} 
0, & \text{if } x = 0, \\
(i - 1)2^{-r}, & \text{if } (i - 1)2^{-r} < x \leq i2^{-r} \leq r, \\
r, & \text{if } x > r,
\end{cases}
\]

We let \(f^{(r)} = \alpha^{(r)}(f)\). Note that \(f^{(r)} \in SF^+\) and \(f^{(r)} \uparrow f\) as \(r \to +\infty\).
Using the previous definition, we get for example the following properties.

**Proposition B.4.6.** Let \( f, g \in (m\Sigma)^+ \).

(i) If \( \mu(f \neq g) = 0 \), then \( \mu(f) = \mu(g) \).

(ii) For all \( c \geq 0 \), \( f + g, cf \in (m\Sigma)^+ \) and

\[
\mu(f + g) = \mu f + \mu g, \quad \mu(cf) = c\mu f.
\]

(iii) If \( f \leq g \) then \( \mu f \leq \mu g \).

For a function \( f \), let \( f^+ \) and \( f^- \) be the positive and negative parts of \( f \), that is,

\[
f^+(s) = f(s) \lor 0, \quad f^-(s) = (-f(s)) \lor 0.
\]

Note that \( |f| = f^+ + f^- \). Finally we define

\[
\mu(f) := \mu(f^+) - \mu(f^-),
\]

provided \( \mu(f^+) + \mu(f^-) < +\infty \), in which case we write \( f \in L^1(S, \Sigma, \mu) \). Proposition B.4.6 can be generalized naturally to this definition. Moreover we have the following.

**Theorem B.4.7** (Dominated convergence theorem). If \( f_n, f \in m\Sigma, n \geq 1 \), with \( f_n(s) \to f(s) \) for all \( s \in S \), and there is a nonnegative function \( g \in L^1(S, \Sigma, \mu) \) such that \( |f_n| \leq g \), then

\[
\mu(|f_n - f|) \to 0,
\]

and in particular

\[
\mu f_n \to \mu f,
\]

as \( n \to \infty \).

More generally, for \( 0 < p < +\infty \), the space \( L^p(S, \Sigma, \mu) \) contains all functions \( f : S \to \mathbb{R} \) such that \( ||f||_p < +\infty \), where

\[
||f||_p := \mu(|f|^p)^{1/p},
\]

up to equality almost everywhere. We state the following results without proof.

**Theorem B.4.8** (Hölder’s inequality). Let \( 1 < p, q < +\infty \) such that \( p^{-1} + q^{-1} = 1 \). Then, for any \( f \in L^p(S, \Sigma, \mu) \) and \( g \in L^q(S, \Sigma, \mu) \), it holds that \( fg \in L^1(S, \Sigma, \mu) \) and further

\[
||fg||_1 \leq ||f||_p ||g||_q.
\]

The case \( p = q = 2 \) is known as the Cauchy-Schwarz inequality (or Schwarz inequality).
Theorem B.4.9 (Minkowski’s inequality). Let $1 < p < +\infty$. Then, for any $f, g \in L^p(S, \Sigma, \mu)$, it holds that $f + g \in L^p(S, \Sigma, \mu)$ and further

$$||f + g||_p \leq ||f||_p + ||g||_p.$$ 

Theorem B.4.10 ($L^p$ completeness). Let $1 \leq p < +\infty$. If $(f_n)_{n \in \mathbb{N}}$ in $L^p(S, \Sigma, \mu)$ is Cauchy, that is,

$$\sup_{n,m \geq k} ||f_n - f_m||_p \to 0,$$

as $k \to +\infty$, then there exists $f \in L^p(S, \Sigma, \mu)$ such that

$$||f_n - f||_p \to 0,$$

as $n \to +\infty$.

We can now define the expectation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition B.4.11 (Expectation). If $X \geq 0$ is a random variable then we define the expectation of $X$, denoted by $\mathbb{E}[X]$, as the integral of $X$ over $\mathbb{P}$. More generally (i.e., not assuming non-negativity), if

$$\mathbb{E}|X| = \mathbb{E}[X^+] + \mathbb{E}[X^-] < +\infty,$$

we let

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-].$$

We denote the set of all such integrable random variables (up to equality almost surely) by $L^1(\Omega, \mathcal{F}, \mathbb{P})$.

The properties of the integral for nonnegative functions (see Proposition B.4.6) extend to the expectation.

Proposition B.4.12. Let $X, X_1, X_2$ be random variables in $L^1(\Omega, \mathcal{F}, \mathbb{P})$.

(LIN) If $a_1, a_2 \in \mathbb{R}$, then $\mathbb{E}[a_1 X_1 + a_2 X_2] = a_1 \mathbb{E}[X_1] + a_2 \mathbb{E}[X_2]$.

(POS) If $X \geq 0$, then $\mathbb{E}[X] \geq 0$.

One useful implication of (POS) is that $|X| - X \geq 0$ so that $\mathbb{E}[|X|] \leq \mathbb{E}[X]$ and, by applying the same argument to $-X$, we have further $|\mathbb{E}[|X|]| \leq \mathbb{E}[|X|]$.

The monotone convergence theorem (Theorem B.4.4) implies the following results. We first need a definition.

Definition B.4.13 (Convergence almost sure). We say that $X_n \to X$ almost surely (a.s.) if

$$\mathbb{P}[X_n \to X] = 1.$$
Proposition B.4.14. Let $X, Y, X_n, n \geq 1$, be random variables in $L^1(\Omega, \mathcal{F}, \mathbb{P})$.

(MON) If $0 \leq X_n \uparrow X$, then $\mathbb{E}[X_n] \uparrow \mathbb{E}[X] \leq +\infty$.

(FATOU) If $X_n \geq 0$, then $\mathbb{E}[\lim \inf_n X_n] \leq \lim \inf_n \mathbb{E}[X_n]$.

(DOM) If $|X_n| \leq Y$, $n \geq 1$, with $\mathbb{E}[Y] < +\infty$ and $X_n \rightarrow X$ a.s., then

$$\mathbb{E}|X_n - X| \rightarrow 0,$$

and, hence,

$$\mathbb{E}[X_n] \rightarrow \mathbb{E}[X].$$

(Indeed,

$$|\mathbb{E}[X_n] - \mathbb{E}[X]| = |\mathbb{E}[X_n - X]| \leq \mathbb{E}|X_n - X|.\)$$

(SCHEFFE) If $X_n \rightarrow X$ a.s. and $\mathbb{E}|X_n| \rightarrow \mathbb{E}|X|$ then

$$\mathbb{E}|X_n - X| \rightarrow 0.$$

(BDD) If $X_n \rightarrow X$ a.s. and $|X_n| \leq K < +\infty$ for all $n$ then

$$\mathbb{E}|X_n - X| \rightarrow 0.$$

Proof. We only prove (FATOU). To use (MON) we write the lim inf as an increasing limit. Letting $Z_k = \inf_{n \geq k} X_n$, we have

$$\lim \inf_n X_n = \uparrow \lim_k Z_k,$$

so that by (MON)

$$\mathbb{E}[\lim \inf_n X_n] = \uparrow \lim_k \mathbb{E}[Z_k].$$

For $n \geq k$ we have $X_n \geq Z_k$ so that $\mathbb{E}[X_n] \geq \mathbb{E}[Z_k]$ hence

$$\mathbb{E}[Z_k] \leq \inf_n \mathbb{E}[X_n].$$

Finally, we get

$$\mathbb{E}[\lim \inf_n X_n] \leq \uparrow \lim_k \inf_n \mathbb{E}[X_n].$$

The following inequality is often useful. We give an example below.
Theorem B.4.15 (Jensen’s inequality). Let \( h : G \to \mathbb{R} \) be a convex function on an open interval \( G \) such that \( P[X \in G] = 1 \) and \( X, h(X) \in L^1(\Omega, \mathcal{F}, P) \) then

\[
E[h(X)] \geq h(E[X]).
\]

The \( L^p \) norm defined earlier applies to random variables as well. That is, for \( p \geq 1 \), we let \( \|X\|_p = E[|X|^p]^{1/p} \) and denote by \( L^p(\Omega, \mathcal{F}, P) \) the collection of random variables \( X \) (up to almost sure equality) such that \( \|X\|_p < +\infty \). Jensen’s inequality (Theorem B.4.15) implies the following relationship.

Lemma B.4.16 (Monotonicity of norms). For \( 1 \leq p \leq r < +\infty \), we have \( \|X\|_p \leq \|X\|_r \).

Proof. For \( n \geq 0 \), let

\[
X_n = (|X| \land n)^p.
\]

Take \( h(x) = x^{r/p} \) which is convex on \((0, +\infty)\). Then, by Jensen’s inequality,

\[
(\mathbb{E}[X_n])^{r/p} \leq \mathbb{E}[(X_n)^{r/p}] = \mathbb{E}[(|X| \land n)^r] \leq \mathbb{E}[|X|^r].
\]

Take \( n \to \infty \) and use (MON).

This latter inequality is useful among other things to argue about the convergence of expectations. We say that \( X_n \) converges to \( X_\infty \) in \( L^p \) if \( \|X_n - X_\infty\|_p \to 0 \). By the previous lemma, convergence on \( L^r \) implies convergence in \( L^p \) for \( r \geq p \geq 1 \). Further we have:

Lemma B.4.17 (Convergence of expectations). Assume \( X_n, X_\infty \in L^1 \). Then

\[
\|X_n - X_\infty\|_1 \to 0,
\]

implies

\[
\mathbb{E}[X_n] \to \mathbb{E}[X_\infty].
\]

Proof. Note that

\[
|\mathbb{E}[X_n] - \mathbb{E}[X_\infty]| \leq \mathbb{E}|X_n - X_\infty| \to 0.
\]

So, a fortiori, convergence in \( L^p, p \geq 1 \), implies convergence of expectations.

Square integrable random variables have a nice geometry by virtue of forming a Hilbert space.
Definition B.4.18 (Square integrable variables). Recall that $L^2(\Omega, \mathcal{F}, \mathbb{P})$ denotes the set of all square integrable random variables (up to equality almost surely), that is, those $X$ with $\mathbb{E}[X^2] < +\infty$. For $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, define the inner product

$$\langle X, Y \rangle := \mathbb{E}[XY].$$

Then the $L^2$ norm is $\|X\|_2 = \sqrt{\langle X, X \rangle}.$

Theorem B.4.19 (Cauchy-Schwarz inequality). If $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, then $XY \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and

$$\mathbb{E}|XY| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]},$$

or put differently

$$|\langle X, Y \rangle| \leq \|X\|_2 \|Y\|_2.$$

Theorem B.4.20 (Parallelogram law). If $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, then

$$\|X + Y\|_2^2 + \|X - Y\|_2^2 = 2\|X\|_2^2 + 2\|Y\|_2^2.$$

B.5 Fubini’s theorem

We now define product measures and state (without proof) Fubini’s Theorem.

Definition B.5.1 (Product $\sigma$-algebra). Let $(S_1, \Sigma_1)$ and $(S_2, \Sigma_2)$ be measure spaces. Let $S = S_1 \times S_2$ be the Cartesian product of $S_1$ and $S_2$. For $i = 1, 2$, let $\pi_i : S \to S_i$ be the projection on the $i$-th coordinate, that is,

$$\pi_i(s_1, s_2) = s_i.$$

The product $\sigma$-algebra $\Sigma = \Sigma_1 \times \Sigma_2$ is defined as

$$\Sigma = \sigma(\pi_1, \pi_2).$$

In words, it is the smallest $\sigma$-algebra that makes coordinate maps measurable. It is generated by sets of the form

$$\pi_1^{-1}(B_1) = B_1 \times S_2, \quad \pi_2^{-1}(B_2) = S_1 \times B_2, \quad B_1 \in \Sigma_1, B_2 \in \Sigma_2.$$

Theorem B.5.2 (Fubini’s Theorem). For $F \in \Sigma$, let $f = 1_F$ and define

$$\mu(F) := \int_{S_1} I_1^f(s_1)\mu_1(ds_1) = \int_{S_2} I_2^f(s_2)\mu_2(ds_2),$$

where

$$I_1^f(s_1) := \int_{S_2} f(s_1, s_2)\mu_2(ds_2) \in b\Sigma_1,$$

$$I_2^f(s_2) := \int_{S_1} f(s_1, s_2)\mu_1(ds_1) \in b\Sigma_2.$$
APPENDIX B. MEASURE-THEORETIC FOUNDATIONS

(The equality and inclusions above are part of the statement.) The set function $\mu$ is a measure on $(S, \Sigma)$ called the product measure of $\mu_1$ and $\mu_2$ and we write $\mu = \mu_1 \times \mu_2$ and

$$(S, \Sigma, \mu) = (S_1, \Sigma_1, \mu_1) \times (S_2, \Sigma_2, \mu_2).$$

Moreover $\mu$ is the unique measure on $(S, \Sigma)$ for which

$$\mu(A_1 \times A_2) = \mu(A_1)\mu(A_2), \quad A_i \in \Sigma_i.$$

If $f \in (m\Sigma)^+$ then

$$\mu(f) = \int_{S_1} I^f_1(s_1)\mu_1(ds_1) = \int_{S_2} I^f_2(s_2)\mu_2(ds_2),$$

where $I^f_1, I^f_2$ are defined as before (i.e., as the sup over bounded functions from below). The same is valid if $f \in m\Sigma$ and $\mu(|f|) < +\infty$.

Some applications of Fubini’s Theorem (Theorem B.5.2) follow. We first recall the following useful formula.

**Theorem B.5.3** (Change-of-variables formula). Let $X$ be a random variable with law $\mathcal{L}$. If $f : \mathbb{R} \to \mathbb{R}$ is such that either $f \geq 0$ or $\mathbb{E}|f(X)| < +\infty$ then

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}} f(y)\mathcal{L}(dy).$$

**Proof.** We use the standard machinery.

1. For $f = 1_B$ with $B \in \mathcal{B}$,

$$\mathbb{E}[1_B(X)] = \mathcal{L}(B) = \int_{\mathbb{R}} 1_B(y)\mathcal{L}(dy).$$

2. If $f = \sum_{k=1}^m a_k 1_{A_k}$ is a simple function, then by (LIN)

$$\mathbb{E}[f(X)] = \sum_{k=1}^m a_k \mathbb{E}[1_{A_k}(X)] = \sum_{k=1}^m a_k \int_{\mathbb{R}} 1_{A_k}(y)\mathcal{L}(dy) = \int_{\mathbb{R}} f(y)\mathcal{L}(dy).$$

3. Let $f \geq 0$ and approximate $f$ by a sequence $\{f_n\}$ of increasing simple functions. By (MON)

$$\mathbb{E}[f(X)] = \lim_n \mathbb{E}[f_n(X)] = \lim_n \int_{\mathbb{R}} f_n(y)\mathcal{L}(dy) = \int_{\mathbb{R}} f(y)\mathcal{L}(dy).$$
4. Finally, assume that $f$ is such that $\mathbb{E}|f(X)| < +\infty$. Then by (LIN)
\begin{align*}
\mathbb{E}[f(X)] &= \mathbb{E}[f^+(X)] - \mathbb{E}[f^-(X)] \\
&= \int_{\mathbb{R}} f^+(y) \mathcal{L}(dy) - \int_{\mathbb{R}} f^-(y) \mathcal{L}(dy) \\
&= \int_{\mathbb{R}} f(y) \mathcal{L}(dy).
\end{align*}

**Theorem B.5.4.** Let $X$ and $Y$ be independent random variables with respective laws $\mu$ and $\nu$. Let $f$ and $g$ be measurable functions such that either $f, g \geq 0$ or $\mathbb{E}|f(X)|, \mathbb{E}|g(Y)| < +\infty$. Then
\[ \mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]. \]

**Proof.** From the change-of-variables formula (Theorem B.5.3) and Fubini’s Theorem (Theorem B.5.2), we get
\begin{align*}
\mathbb{E}[f(X)g(Y)] &= \int_{\mathbb{R}^2} f(x)g(y)(\mu \times \nu)(dx \times dy) \\
&= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} f(x)g(y)\mu(dx) \right) \nu(dy) \\
&= \int_{\mathbb{R}} (g(y)\mathbb{E}[f(X)]) \nu(dy) \\
&= \mathbb{E}[f(X)]\mathbb{E}[g(Y)].
\end{align*}

**Definition B.5.5 (Density).** Let $X$ be a random variable with law $\mu$. We say that $X$ has density $f_x$ if for all $B \in \mathcal{B}(\mathbb{R})$
\[ \mu(B) = \mathbb{P}[X \in B] = \int_B f_X(x) \lambda(dx). \]

**Theorem B.5.6 (Convolution).** Let $X$ and $Y$ be independent random variables with distribution functions $F$ and $G$ respectively. Then the distribution function, $H$, of $X + Y$ is
\[ H(z) = \int F(z - y) dG(y). \]
This is called the convolution of $F$ and $G$. Moreover, if $X$ and $Y$ have densities $f$ and $g$ respectively, then $X + Y$ has density
\[ h(z) = \int f(z - y)g(y) dy. \]
APPENDIX B. MEASURE-THEORETIC FOUNDATIONS

Proof. From Fubini’s Theorem (Theorem B.5.3), denoting the laws of \( X \) and \( Y \) by \( \mu \) and \( \nu \) respectively,
\[
\Pr[X + Y \leq z] = \int \int 1_{(x+y \leq z)} \mu(dx)\nu(dy)
\]
\[
= \int F(z-y)\nu(dy)
\]
\[
= \int F(z-y)dG(y)
\]
\[
= \int \left( \int_{-\infty}^{z} f(x-y)dx \right) dG(y)
\]
\[
= \int_{-\infty}^{z} \left( \int f(x-y)dG(y) \right) dx
\]
\[
= \int_{-\infty}^{z} \left( \int f(x-y)g(y)dy \right) dx.
\]

See Exercise 2.1 for a proof of the following standard formula.

Theorem B.5.7 (Moments of nonegative random variables). For any nonnegative random variable \( X \) and positive integer \( k \),
\[
\mathbb{E}[X^k] = \int_{0}^{+\infty} kx^{k-1}\Pr[X > x] \, dx. \tag{B.5.1}
\]

B.6 Conditional expectation

Before defining the conditional expectation, we recall some elementary concepts. For two events \( A, B \), the conditional probability of \( A \) given \( B \) is defined as
\[
\Pr[A \mid B] = \frac{\Pr[A \cap B]}{\Pr[B]},
\]
where we assume \( \Pr[B] > 0 \).

Now let \( X \) and \( Z \) be random variables taking values \( x_1, \ldots, x_m \) and \( z_1, \ldots, z_n \) respectively. The conditional expectation of \( X \) given \( Z = z_j \) is defined as
\[
y_j = \mathbb{E}[X \mid Z = z_j] = \sum_i x_i \Pr[X = x_i \mid Z = z_j],
\]
where we assume \( \Pr[Z = z_j] > 0 \) for all \( j \). As motivation for the general definition, we make the following observations.
We can think of the conditional expectation as a random variable \( Y = E[X \mid Z] \) defined as follows

\[
Y(\omega) = y_j \text{ on } G_j = \{ \omega : Z(\omega) = z_j \}.
\]

Then \( Y \) is \( \mathcal{G} \)-measurable where \( \mathcal{G} = \sigma(Z) \).

On sets in \( \mathcal{G} \), the expectation of \( Y \) agrees with the expectation of \( X \). Indeed, note first that

\[
E[Y; G_j] = y_j \mathbb{P}[G_j] = \sum_i x_i \mathbb{P}[X = x_i \mid Z = z_j] \mathbb{P}[Z = z_j] = \sum_i x_i \mathbb{P}[X = x_i, Z = z_j] = \mathbb{E}[X; G_j].
\]

This is also true for all \( G \in \mathcal{G} \) by summation over \( j \).

We are ready to state the general definition of the conditional expectation. Its existence and uniqueness follow from the next theorem.

**Theorem B.6.1** (Conditional expectation). Let \( X \in L^1(\Omega, \mathcal{F}, \mathbb{P}) \) and \( \mathcal{G} \subseteq \mathcal{F} \) a sub-\( \sigma \)-algebra. Then:

(i) (Existence) There exists a random variable \( Y \in L^1(\Omega, \mathcal{G}, \mathbb{P}) \) such that

\[
E[Y; G] = E[X; G], \quad \forall G \in \mathcal{G}. \tag{B.6.1}
\]

Such a \( Y \) is called a version of the conditional expectation of \( X \) given \( \mathcal{G} \) and is denoted by \( E[X \mid \mathcal{G}] \).

(ii) (Uniqueness) It is unique in the sense that, if \( Y' \) and \( Y'' \) are two versions of the conditional expectation, then \( Y = Y' \) almost surely.

When \( \mathcal{G} = \sigma(Z) \), we sometimes use the notation \( E[X \mid Z] := E[X \mid \mathcal{G}] \). A similar convention applies to collections of random variables, for example, \( E[X \mid Z_1, Z_2] := E[X \mid \sigma(Z_1, Z_2)] \) and so on.

We first prove uniqueness. Existence is proved below after some more concepts are introduced.
APPENDIX B. MEASURE-THEORETIC FOUNDATIONS

Proof of Theorem B.6.1 (ii). By way of contradiction, let \( Y, Y' \) be two versions of \( E[X \mid G] \) such that without loss of generality \( P[Y > Y'] > 0 \). By monotonicity, there is \( n \geq 1 \) with \( G = \{ Y > Y' + n^{-1} \} \in \mathcal{G} \) such that \( P[G] > 0 \). Then, by definition,
\[
0 = E[Y - Y'; G] > n^{-1}P[G] > 0,
\]
which gives a contradiction. \( \blacksquare \)

To prove existence, we use the \( L^2 \) method. In \( L^2(\Omega, \mathcal{F}, \mathbb{P}) \), the conditional expectation reduces to an orthogonal projection.

**Theorem B.6.2** (Conditional expectation: \( L^2 \) case). Let \( X \in L^2(\Omega, \mathcal{F}, \mathbb{P}) \) and \( \mathcal{G} \subseteq \mathcal{F} \) a sub-\( \sigma \)-algebra. Then there exists an (almost surely) unique \( Y \in L^2(\Omega, \mathcal{G}, \mathbb{P}) \) such that
\[
\|X - Y\|_2 = \Delta := \inf \{\|X - W\|_2 : W \in L^2(\Omega, \mathcal{G}, \mathbb{P})\},
\]
and, moreover, \( \langle Z, X - Y \rangle = 0, \forall Z \in L^2(\Omega, \mathcal{G}, \mathbb{P}) \). In particular, it satisfies (B.6.1). Such a \( Y \) is called the orthogonal projection of \( X \) on \( L^2(\Omega, \mathcal{G}, \mathbb{P}) \).

**Proof.** Take \( (Y_n) \) such that \( \|X - Y_n\|_2 \to \Delta \). We use the fact that \( L^2(\Omega, \mathcal{G}, \mathbb{P}) \) is complete (Theorem B.4.10) and first seek to prove that \( (Y_n) \) is Cauchy. Using the parallelogram law (Theorem B.4.20), note that
\[
\|X - Y_r\|_2^2 + \|X - Y_s\|_2^2 = 2\left\|X - \frac{1}{2}(Y_r + Y_s)\right\|_2^2 + 2\left\|\frac{1}{2}(Y_r - Y_s)\right\|_2^2.
\]
The first term on the right-hand side is \( \geq 2\Delta^2 \) by definition of \( \Delta \), so taking limits \( r, s \to +\infty \) we have what we need, that is, that \( (Y_n) \) is indeed Cauchy.

Let \( Y \) be the limit of \( (Y_n) \) in \( L^2(\Omega, \mathcal{G}, \mathbb{P}) \). Note that by the triangle inequality
\[
\Delta \leq \|X - Y\|_2 \leq \|X - Y_n\|_2 + \|Y_n - Y\|_2 \to \Delta,
\]
as \( n \to +\infty \). As a result, for any \( Z \in L^2(\Omega, \mathcal{G}, \mathbb{P}) \) and \( t \in \mathbb{R} \),
\[
\|X - Y - tZ\|_2^2 \geq \Delta^2 = \|X - Y\|_2^2,
\]
so that, expanding and rearranging, we have
\[
-2t\langle Z, X - Y \rangle + t^2\|Z\|_2^2 \geq 0,
\]
which is only possible for every \( t \in \mathbb{R} \) if the first term is 0.

Uniqueness follows from the parallelogram law and the definition of \( \Delta \). \( \blacksquare \)
We return to the proof of existence of the conditional expectation. We use the standard machinery.

Proof of Theorem B.6.1 (i). The previous theorem implies that conditional expectations exist for indicators and simple functions. Now take \( X \in L^1(\Omega, F, P) \) and write \( X = X^+ - X^- \), so we can assume \( X \) is in fact nonnegative without loss of generality. Using the staircase function

\[
X^{(r)} = \begin{cases} 
0, & \text{if } X = 0 \\
(i-1)2^{-r}, & \text{if } (i-1)2^{-r} < X \leq i2^{-r} \leq r \\
r, & \text{if } X > r,
\end{cases}
\]

we have \( 0 \leq X^{(r)} \uparrow X \). Let \( Y^{(r)} = E[X^{(r)} \mid G] \). Using an argument similar to the proof of uniqueness, it follows that \( U \geq 0 \) implies \( E[U \mid G] \geq 0 \) for a simple function \( U \). Using linearity (which is immediate from the definition), we then have \( Y^{(r)} \uparrow Y := \limsup Y^{(r)} \) which is measurable in \( G \). By (MON),

\[
E[Y; G] = E[X; G], \quad \forall G \in G.
\]

That concludes the proof.

Before deriving some properties, we give a few examples.

**Example B.6.3.** If \( X \in L^1(\Omega, G, P) \) then \( E[X \mid G] = X \) almost surely trivially.

**Example B.6.4.** If \( G = \{\emptyset, \Omega\} \), then \( E[X \mid G] = E[X] \).

**Example B.6.5.** Let \( A, B \in \mathcal{F} \) with \( 0 < P[B] < 1 \). If \( G = \{\emptyset, B, B^c, \Omega\} \) and \( X = 1_A \), then

\[
P[A \mid G] = \begin{cases} 
P[A \cap B], & \text{on } \omega \in B, \\
\frac{P[A \cap B]}{P[B]}, & \text{on } \omega \in B^c.
\end{cases}
\]

Intuition about the conditional expectation sometimes breaks down.

**Example B.6.6.** On \( (\Omega, F, P) = ((0, 1], B(0, 1], \lambda) \), let \( G \) be the \( \sigma \)-algebra of all countable and co-countable (i.e., whose complement in \( (0, 1] \) is countable) subsets of \( (0, 1] \). Then \( P[G] \in \{0, 1\} \) for all \( G \in G \) and

\[
E[X; G] = E[E[X]; G] = E[X]P[G],
\]

so that \( E[X \mid G] = E[X] \). Yet, \( G \) contains all singletons and we seemingly have “full information,” which would lead to the wrong guess \( E[X \mid G] = X \).
We show that the conditional expectation behaves similarly to the ordinary expectation. Below all \(X\) and \(X_i\) are in \(L^1(\Omega, \mathcal{F}, \mathbb{P})\) and \(\mathcal{G}\) is a sub \(\sigma\)-algebra of \(\mathcal{F}\).

**Lemma B.6.7** (cLIN). If \(a_1, a_2 \in \mathbb{R}\), then \(E[a_1X_1 + a_2X_2 | \mathcal{G}] = a_1E[X_1 | \mathcal{G}] + a_2E[X_2 | \mathcal{G}]\) a.s.

**Proof.** Use the linearity of expectation and the fact that a linear combination of random variables in \(\mathcal{G}\) is also in \(\mathcal{G}\).

**Lemma B.6.8** (cPOS). If \(X \geq 0\) then \(E[X | \mathcal{G}] \geq 0\) a.s.

**Proof.** Let \(Y = E[X | \mathcal{G}]\) and assume for contradiction that \(\mathbb{P}[Y < 0] > 0\). There is \(n \geq 1\) such that \(\mathbb{P}[Y < -n^{-1}] > 0\). But that implies, for \(G = \{Y < -n^{-1}\}\),

\[
E[X; G] = E[Y; G] < -n^{-1} \mathbb{P}[G] < 0,
\]

a contradiction.

**Lemma B.6.9** (cMON). If \(0 \leq X_n \uparrow X\) then \(E[X_n | \mathcal{G}] \uparrow E[X | \mathcal{G}]\) a.s.

**Proof.** Let \(Y_n = E[X_n | \mathcal{G}]\). By (cLIN) and (cPOS), \(0 \leq Y_n \uparrow\). Then letting \(Y = \lim \sup Y_n\), by (MON),

\[
E[X; G] = E[Y; G],
\]

for all \(G \in \mathcal{G}\).

**Lemma B.6.10** (cFATOU). If \(X_n \geq 0\) then \(E[\lim \inf X_n | \mathcal{G}] \leq \lim \inf E[X_n | \mathcal{G}]\) a.s.

**Proof.** Note that, for \(n \geq m\),

\[
X_n \geq Z_m := \inf_{k \geq m} X_k \uparrow \in \mathcal{G},
\]

so that \(\inf_{n \geq m} E[X_n | \mathcal{G}] \geq E[Z_m | \mathcal{G}]\). Applying (cMON)

\[
E[\lim Z_m | \mathcal{G}] = \lim E[Z_m | \mathcal{G}] \leq \lim \inf_{n \geq m} E[X_n | \mathcal{G}].
\]

**Lemma B.6.11** (cDOM). If \(X_n \leq V \in L^1(\Omega, \mathcal{F}, \mathbb{P})\) and \(X_n \to X\) a.s., then

\[
E[X_n | \mathcal{G}] \to E[X | \mathcal{G}]\) a.s.
APPENDIX B. MEASURE-THEORETIC FOUNDATIONS

Proof. Applying (cFATOU) to $W_n := 2V - |X_n - X| \geq 0$,

$$
\mathbb{E}[2V \mid \mathcal{G}] = \mathbb{E}\left[\lim\inf_{n \to \infty} W_n \mid \mathcal{G}\right] \\
\leq \lim\inf_{n \to \infty} \mathbb{E}[W_n \mid \mathcal{G}] \\
= \mathbb{E}[2V \mid \mathcal{G}] - \lim\inf_{n \to \infty} \mathbb{E}[|X_n - X| \mid \mathcal{G}],
$$

so we must have

$$
\lim\inf_{n \to \infty} \mathbb{E}[|X_n - X| \mid \mathcal{G}] = 0.
$$

Now use that $|\mathbb{E}[X_n - X \mid \mathcal{G}]| \leq \mathbb{E}[|X_n - X| \mid \mathcal{G}]$ (which follows from (cPOS)). □

Lemma B.6.12 (cJENSEN). If $f$ is convex and $\mathbb{E}[|f(X)|] < +\infty$ then

$$
f(\mathbb{E}[X \mid \mathcal{G}]) \leq \mathbb{E}[f(X) \mid \mathcal{G}].
$$

In addition, we highlight (without proof) the following important properties of the conditional expectation.

Lemma B.6.13 (Taking out what is known). If $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $Z \in \mathfrak{m}\mathcal{G}$ is bounded or if $X$ is bounded and $Z \in L^1(\Omega, \mathcal{G}, \mathbb{P})$, then $\mathbb{E}[ZX \mid \mathcal{G}] = Z \mathbb{E}[X \mid \mathcal{G}]$. This is also true if $X, Z \geq 0$, $\mathbb{E}[X] < +\infty$ and $\mathbb{E}[ZX] < +\infty$, or $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$.

Lemma B.6.14 (Role of independence). If $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ is independent of $\mathcal{H}$ then $\mathbb{E}[X \mid \mathcal{H}] = \mathbb{E}[X]$. In fact, if $\mathcal{H}$ is independent of $\sigma(\sigma(X), \mathcal{G})$, then $\mathbb{E}[X \mid \sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X \mid \mathcal{G}]$.

Lemma B.6.15 (Conditioning on an independent random variable). Suppose $X, Y$ are independent. Let $\phi$ be a function with $\mathbb{E}[|\phi(X, Y)|] < +\infty$ and let $g(x) = \mathbb{E}(\phi(x, Y))$. Then,

$$
\mathbb{E}(\phi(X, Y) \mid X) = g(X).
$$

Lemma B.6.16 (Tower property). If $\mathcal{H} \subseteq \mathcal{G}$ is a $\sigma$-algebra and $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$

$$
\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}\left[\mathbb{E}[X \mid \mathcal{H}] \mid \mathcal{G}\right] = \mathbb{E}[X \mid \mathcal{H}].
$$

That is, the “smallest $\sigma$-algebra wins.”

An important special case of the latter, also known as the law of total probability or the law of total expectation, is $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$.

One last useful property:

Lemma B.6.17. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $Y_1 = Y_2$ a.s. on $B \in \mathcal{F}
then $\mathbb{E}[Y_1 \mid \mathcal{F}] = \mathbb{E}[Y_2 \mid \mathcal{F}]$ a.s. on $B$. 


B.7 Filtered spaces

Finally we define stochastic processes. Let $E$ be a set and let $\mathcal{E}$ be a $\sigma$-algebra defined over $E$.

**Definition B.7.1.** A stochastic process (or process) is a collection \( \{X_t\}_{t \in \mathcal{T}} \) of \((E, \mathcal{E})\)-valued random variables on a probability space \((\Omega, \mathcal{F}, \mathbb{P})\), where $\mathcal{T}$ is an arbitrary index set.

Here is a typical example.

**Example B.7.2.** When $\mathcal{T} = \mathbb{Z}_+$ (or $\mathcal{T} = \mathbb{N}$ or $\mathcal{T} = \mathbb{Z}$) we have a discrete-time process, in which case we often write the process as a sequence $(X_t)_{t \geq 0}$. For instance:

- $X_0, X_1, X_2, \ldots$ i.i.d. random variables;
- $(S_t)_{t \geq 0}$ where $S_t = \sum_{i \leq t} X_i$ with $X_i$ as above.

We let $\mathcal{F}_t = \sigma(X_0, X_1, \ldots, X_t)$, which can be thought of as “the information known up to time $t$.” For a fixed $\omega \in \Omega$, $(X_t(\omega) : t \in \mathcal{T})$ is called a sample path.

**Definition B.7.3.** A random walk on $\mathbb{R}^d$ is a process of the form:

\[
S_t = S_0 + \sum_{i=1}^t X_i, \quad t \geq 1
\]

where the $X_i$s are i.i.d. in $\mathbb{R}^d$, independent of $S_0$. The case $X_i$ uniform in $\{-1, +1\}$ is called a simple random walk on $\mathbb{Z}$.

Filtered spaces provide a formal framework for time-indexed processes. We restrict ourselves to discrete time. (We will not discuss continuous-time processes in this book.)

**Definition B.7.4.** A filtered space is a tuple $\langle \Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathbb{Z}_+}, \mathbb{P} \rangle$ where:

- $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space;
- $(\mathcal{F}_t)_{t \in \mathbb{Z}_+}$ is a filtration, that is,

\[
\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_\infty := \sigma(\cup_t \mathcal{F}_t) \subseteq \mathcal{F}.
\]

where each $\mathcal{F}_i$ is a $\sigma$-algebra.
Definition B.7.5. Fix \((\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathbb{Z}^+}, \mathbb{P})\). A process \((W_t)_{t \geq 0}\) is adapted if \(W_t \in \mathcal{F}_t\) for all \(t\).

Intuitively, in the previous definition, the value of \(W_t\) is “known at time \(t\).”

Definition B.7.6. A process \((C_t)_{t \geq 1}\) is predictable if \(C_t \in \mathcal{F}_{t-1}\) for all \(t \geq 1\).

Example B.7.7. Continuing Example B.7.2. The collection \((\mathcal{F}_t)_{t \geq 0}\) forms a filtration. The process \((S_t)_{t \geq 0}\) is adapted. On the other hand, the process \(C_t = 1\{S_{t-1} \leq k\}\) is predictable.
Bibliography


532


[Dev98] Luc Devroye. Branching processes and their applications in the analysis of tree structures and tree algorithms. In Michel Habib,


535


540


541


Index

\(\varepsilon\)-net, see epsilon-net

Azuma-Hoeffding inequality, see tail bounds

balancing vectors, 34
ballot theorem, 125
balls and bins, 153
bandits
  definition, 168
  optimism in the face of uncertainty, 169
  upper confidence bound, 169
Bernstein’s inequality, see tail bounds
Berry-Esséen theorem, 156
binary classification
  definitions, 101–102
  empirical risk minimizer, 102
  no free lunch, 102
binary search tree, 440
binomial coefficients
  binomial theorem, 500
  bounds, 23, 500
  definition, x
bond percolation
  definition, 18
  FKG, 263
Bonferroni inequalities, 115
Boole’s inequality, see union bound
Boolean functions
  random \(k\)-SAT problem, 37
  bounded differences inequality, 151
branching processes
  binomial offspring, 436
  duality principle, 430
  exploration process, 429
  extinction, 415, 416
  Galton-Watson process, 414
  Galton-Watson tree, 414, 420
  geometric offspring, 493
  history, 430, 494
  infinite line of descent, 494
  linear functionals, 425
  multitype, see multitype branching processes
  Poisson offspring, 418, 431, 433
  positive regular case, 423
  random-walk representation, 427, 429, 431

canonical paths, 404
Cauchy-Schwarz inequality, 513, 517
Cavender-Farris-Neyman (CFN) model, 449
Cayley’s formula, see trees
chaining method, 90
Chebyshev polynomials, 364
Chebyshev’s association inequality, 318
Chebyshev’s inequality, see tail bounds
Cheeger’s inequality, 384, 489
Chen-Stein method
  dissociated case, 303, 316
  main result, 299
  negatively related variables, 322

549
positively related variables, 322
Stein coupling, 300, 304
Stein equation, 299
Chernoff-Cramér bound, see tail bounds
Chernoff-Cramér method, 67, 116, 145
community recovery, 343
compressed sensing, 96
conzentration inequalities, see tail bounds, 119
conzentration of measure, 155
conditional expectation
definition, 520–523
eamples, 523–524
law of total probability, 525
properties, 524–525
tower property, 525
congestion ratio, 404
convergence in probability, 32
convex duality
Lagragian, 201
weak duality, 201
coupling
coualescence time, 280
coualescing, 280
coupling inequality, 239, 244
definition, 233
Erdös-Rényi graph model, 464, 481
independent, 233
Markovian, 244, 280
maximal coupling, 240
monotone, 233, 234, 249, 252, 256
path coupling, 291, 297
splitting, 281
stochastic domination, 249
Strassen’s theorem, 252
coupon collector, 31, 284
Courant-Fischer theorem, 329, 334, 338, 339, 382
covariance, 30
covering number, 86, 111
cumulant-generating function, 64, 66
Curie-Weiss model, 399
cutset, 58
Davis-Kahan theorem, 341, 347
dependency graph, 46
dimension reduction, 92
Dirichlet
ey, 382, 404
orm, 382
principle, 204, 227
problem, 181
drift condition, 188
Dudley’s inequality, 91, 110
edge expansion constant, see networks
Efron-Stein inequality, 150
electrical networks
definitions, 190–192
Dirichlet energy, see Dirichlet
effective conductance, 198, 204
effective resistance, 198, 212, 227
flow, see flows
Kirchhoff’s cycle law, 191
Kirchhoff’s node law, 191
Kirchhoff’s resistance formula, 216
Nash-Williams inequality, 204, 227
Ohm’s law, 191, 217
parallel law, 194
Rayleigh’s principle, 207, 217
series law, 194
Thomson’s principle, 200, 206
voltage, 190
empirical measure, 110, 112
epsilon-net, 86, 89, 98–100
Erdös-Rényi graph model
chromatic number, 157
clique number, 316, 319
cluster, 466
connectivity, 52

550
definition, 18
degree sequence, 244
evolution, 464
exploration process, 466
FKG, 263
giant component, 465, 479
isolated vertices, 51
largest connected component, 465
maximum degree, 78
positive associations, 261
random walk, 488
subgraph containment, 267
threshold function, 47–55

Erdős-Rényi random graph model
clique number, 48
exhaustive sequence, 227
expander graphs
\((d, \alpha)\)-expander family, 392
Pinaker’s model, 392

factorials
bounds, 22, 500
definition, x
Stirling’s formula, 500

Fenchel-Legendre dual, 66
first moment method, 36–41, 46, 48, 49, 52–54, 56, 58, 143, 160
first moment principle, 33, 36, 103
first-step analysis, 184
FKG
condition, 262, 319
inequality, 263, 319, 323
measure, 262, 320

flows
current, 191
definition, 6
energy, 200, 207, 210
flow-conservation constraints, 191, 208
max-flow min-cut theorem, 7, 253
strength, 191
to \(\infty\), 207, 209

gambler’s ruin, 137–139, 193, 199
Gibbs random fields, 19
Glauber dynamics
definition, 20
fast mixing, 295, 399

graph Laplacian
connectivity, 332
definition, 331
degree, 333
eigenvalues, 332
Fiedler vector, 332
network, 337
normalized, 338
quadratic form, 331

graphs
3–1 tree, see trees
\(b\)-ary tree \(\mathbb{T}_b^c\), 5, 288
\(n\)-clique, 381
adjacency matrix, see matrices
bridge, 216
Cayley’s formula, see trees
chromatic number, 8, 157
clique, 3, 48
clique number, 48, 316
coloring, 8
cutset, 58, 204
cycle \(C_n\), 5, 283
definitions, 2–9
degree, 78
diameter, 367
directed, 8
expander, see expander graphs
flow, see flows
graph distance, 4
hypercube \(\mathbb{Z}_q^n\), 5, 284
incidence matrix, see matrices
independent set, 8, 34
infinite, 5
infinite binary tree, 449
Laplacian, see graph Laplacian
matching, 8
matrix representation, see matrices
multigraph, 2
network, see networks
oriented incidence matrix, see matrices
perfect matching, 8
spanning arborescence, 220
torus $\mathbb{T}^d$, 5
tree, see trees
Turán graphs, 35
Green function, see Markov chains
Hölder’s inequality, 513
Hamming distance, 176
harmonic functions, see Markov chains
Harper’s vertex isoperimetric theorem, 156
Harris’ inequality, 318, 323
Harris’ theorem, see percolation
hitting time, see stopping time
Hoeffding’s inequality, see tail bounds
Hoeffding’s lemma, see tail bounds, 145
Holley’s inequality, 264, 319, 323
increasing event, see posets
indicator trick, 36
inherited property, 419
Ising model
boundary conditions, 258
complete graph, 399
definition, 19
FKG, 263
Glauber dynamics, 284, 295, 398
magnetization, 399
random cluster, 449
trees, 449, 497
isoperimetric inequality, 379
Janson’s inequality, 268, 323
Jensen’s inequality, 516
Johnson-Lindenstrauss
distributional lemma, 93
lemma, 92–95
Kesten’s theorem, see percolation
knapsack problem, 79
Kolmogorov’s maximal inequality, 140, 320
Kullback-Leibler divergence, 68
Laplacian
graphs, see graph Laplacian
Markov chains, 181, 202, 321
networks, see graph Laplacian
large deviations, 69, 119
law of total probability, see conditional expectation
laws of large numbers, 32–33
Lipschitz
condition, 176–178
process, 87
Lyapounov function, see Markov chains
Markov chain Monte Carlo, 369
Markov chain tree theorem, 229
Markov chains
average occupation time, 184
birth-death, 192, 226, 374
bottleneck ratio, see networks
Chapman-Kolmogorov, 10
commute time, 212
commute time identity, 212, 228, 288
construction, 9
cover time, 123
decomposition theorem, 126
definitions, 9–17
Doeblin’s condition, 281, 320
escape probability, 195, 203
examples, 9–16
exit law, 184
exit probability, 184
first return time, 123
first visit time, 123, 179
Green function, 128, 184, 195
harmonic functions, 179
hitting times, 311
irreducible set, 126
lower bound, 283
Lyapounov function, 188, 292
Markov property, 10
martingales, 133
Matthews’ cover time bounds, 131, 228
mean exit time, 184
Metropolis algorithm, 15
mixing time, see mixing times
positive recurrence, 126
potential theory, 183
recurrence, 197, 207–212
recurrent state, 126
relaxation time, 356
reversibility, 179, 364
splitting, see coupling
stationary measure, 127
stochastic domination, 258, 265
stochastic monotonicity, 257
strong Markov property, 123
uniform geometric ergodicity, 282
Varopoulos-Carne bound, 363, 412
Markov’s inequality, see tail bounds, 64
martingales, 454
Azuma-Hoeffding inequality, see tail bounds
convergence theorem, 141, 415
definition, 132
Doob martingale, 134, 146, 157
Doob’s submartingale inequality, 140, 145
edge exposure martingale, 158
exposure martingale, 157
hitting time, 228
Markov chain, see Markov chains
martingale difference, 146
optional stopping theorem, 135
orthogonality of increments, 142, 147
stopped process, 135
submartingale, 132
supermartingale, 132
vertex exposure martingale, 157
matrices
2-norm, 88
adjacency, 2, 330
block, 326
diagonal, 326
graph Laplacian, see graph Laplacian
incidence, 3
oriented incidence, 9
orthogonal, 326
spectral norm, 88, 177, 339
spectral radius, 408, 423
spectrum, 408
stochastic matrix, 9
symmetric, 326
maximal Azuma-Hoeffding inequality, see tail bounds
maximum principle, 14, 24, 181, 227
method of bounded differences, see tail bounds
method of moments, 118, 119
method of random paths, 214, 227
Metropolis algorithm, see Markov chains
minimum bisection problem, 345
Minkowski’s inequality, 514
mixing times
b-ary tree, 288, 389
cutoff, 287, 363, 409
cycle, 320, 358, 390
definition, 17
diameter, 368
diameter bound, 368
distinguishing statistic, 287, 320
hypercube, 361, 391, 411
lower bound, 282, 285, 295, 320, 366, 368
random walk on cycle, 283
random walk on hypercube, 284
separation distance, 356
upper bound, 295
moment-generating functions
\( \chi^2, 73 \)
definition, 28
Gaussian, 65
Poisson, 67
Rademacher, 65
moments, 28
exponential moment, see moment-generating functions
multi-type branching processes
definitions, 421–423
Kesten-Stigum bound, 453
mean matrix, 422
nonsingular case, 422
negative associations, 217
networks
cut, 380
definition, 8
edge boundary, 380
edge expansion, 380, 381
vertex boundary, 392
no free lunch, see binary classification
notation, ix–xi
operator
compact, 374
norm, 375
spectral radius, 376
optimal transport, 321
optional stopping theorem, see martingales
Pólya’s theorem, see random walk
Pólya’s urn, 141
packing number, 86
Pajor’s lemma, 113
parity functions, 361
Parseval’s identity, 350
pattern matching, 154
peeling method, see slicing method
percolation
contour lemma, 42
critical exponents, 439
critical value, 40, 55, 270, 421, 436
dual lattice, 41, 271
Galton-Watson tree, 420
Harris’ theorem, 270, 323
Kesten’s theorem, 271, 323
on \( L^2 \), 40, 270
on \( L^d \), 252
on a graph, 234
on infinite trees, 55, 143
percolation function, 40, 55, 252, 270, 436
RSW lemma, 271, 320, 323
permutations
Erdős-Szekeres Theorem, 40
longest increasing subsequence, 38
random, 38
Perron-Frobenius theory
Perron vector, 424
theorem, 12, 424, 454
Poincaré inequality, 151, 383, 385, 404
Poisson approximation, 243, 267
Poisson equation, 185
Poisson trials, 69

posets
decreasing event, 262
definition, 251
increasing event, 252, 318

positive associations
definition, 261
strong, 319

positively correlated events, 262
probabilistic method, 33–36, 93, 102, 112
probability generating function, 416
probability spaces
definitions, 501–504
distribution function, 504
expectation, 511–517
filtered spaces, 121, 526
Fubini’s theorem, 517
independence, 508–511
process, 526
random variables, 504–508

pseudo-regret, 169

random graphs
Erdős-Rényi, see Erdős-Rényi graph
model
preferential attachment, 19, 161
stochastic blockmodel, 343
random projection, 93
random target lemma, 182
random variables
$\chi^2$, 73, 94
Bernoulli, 68, 233, 242, 248, 305
binomial, 68, 285, 318
Gaussian, 30, 70, 73, 93
geometric, 32, 282
Poisson, 67, 239, 242, 248, 250, 418, 494
Rademacher, 65, 70
uncorrelated, 32, 116, 142
uniform, 31

random walk
b-ary tree, 288
asymmetric random walk on $\mathbb{Z}$, 378
biased random walk on $\mathbb{Z}$, 138, 234
cycle, 283, 358
hypercube, 284, 361
lazy, 16, 236, 282
loop erasure, 218
on a graph, 10–15
on a network, 20
Pólya’s theorem, 213
reflection principle, 124
simple random walk on $\mathbb{Z}$, 11, 126, 137, 181, 209, 363, 364, 526
simple random walk on $\mathbb{Z}^d$, 236
simple random walk on a graph, 20
tree, 237
Wald’s identities, 136–139
Rayleigh quotient, 328, 334, 382
reconstruction problem
definition, 449
MAP estimator, 451
solvability, 450
reflection principle, see random walk
relaxation times
cycle, 359
hypercube, 362
restricted isometry property, 96, 100–101
rough embedding, 208, 209
rough equivalence, 209, 228
rough isometry, 228
RSW lemma, see percolation
Sauer’s lemma, 108, 109, 112
second moment method, 45, 46, 48, 50, 52, 54, 57, 58, 60
set balancing, 66
shattering, 107
simple random walk on a graph, see random walk
type, see recurrence
uniform spanning trees, see trees
union bound, 36, 115
Varopoulos-Carne bound, see Markov chains
VC dimension, 107, 111
Wald’s identities, see random walk
Wasserstein distance, 321
Weyl’s inequality, 340