Chapter 3

Martingales and potentials

Martingales are a central tool in probability theory. In this chapter we illustrate their use on a number of applications in discrete probability. We also give an introduction to the related electrical network theory of Markov chains.

3.1 Background

We begin with a quick review of stopping times and martingales. Recall:

**Definition 3.1** (Filtered space). A filtered space is a tuple \((\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathbb{Z}_+}, \mathbb{P})\) where:

- \((\Omega, \mathcal{F}, \mathbb{P})\) is a probability space
- \((\mathcal{F}_t)_{t \in \mathbb{Z}_+}\) is a filtration, i.e.,
  \[ \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_\infty := \sigma(\bigcup \mathcal{F}_t) \subseteq \mathcal{F}, \]
  where each \(\mathcal{F}_t\) is a \(\sigma\)-field.

**Example 3.2** (I.i.d. random variables). Let \(X_0, X_1, \ldots\) be i.i.d. random variables. Then a filtration is given by

\[ \mathcal{F}_t = \sigma(X_0, \ldots, X_t), \forall t \geq 0. \]

Fix \((\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathbb{Z}_+}, \mathbb{P})\). A process \((W_t)_t\) is adapted if \(W_t \in \mathcal{F}_t\) for all \(t\). A process \(\{C_t\}_{t \geq 1}\) is predictable if \(C_t \in \mathcal{F}_{t-1}\) for all \(t\).
Example 3.3 (I.i.d. random variables (continued)). The process \((S_t)_t\), where \(S_t = \sum_{i \leq t} X_i\), is adapted. The process \((C_t)_t\), where \(C_t = 1\{S_{t-1} \leq k\}\), is predictable.

3.1.1 Stopping times

Definitions. Roughly, a stopping time is a random time whose value only depends on the process up to that time. More formally:

Definition 3.4 (Stopping time). A random variable \(\tau : \Omega \to \mathbb{Z}_+ := \{0, 1, \ldots, +\infty\}\) is called a stopping time if
\[
\{\tau \leq t\} \in \mathcal{F}_t, \forall t \in \mathbb{Z}_+,
\]
or, equivalently,
\[
\{\tau = t\} \in \mathcal{F}_t, \forall t \in \mathbb{Z}_+.
\]
(To see the equivalence, note that \(\{\tau = t\} = \{\tau \leq t\} \setminus \{\tau \leq t - 1\}\), and \(\{\tau \leq t\} = \bigcup_{i \leq t} \{\tau = i\}\).)

Example 3.5 (Hitting time). Let \((A_t)_{t \in \mathbb{Z}_+}\), with values in \((E, \mathcal{E})\), be adapted and \(B \in \mathcal{E}\). Then
\[
\tau = \inf\{t \geq 0 : A_t \in B\},
\]
is a stopping time known as a hitting time. In contrast, the last visit to a set is typically not a stopping time.

Let \(\tau\) be a stopping time. Denote by \(\mathcal{F}_\tau\) the set of all events \(F\) such that \(\forall t \in \mathbb{Z}_+ F \cap \{\tau = t\} \in \mathcal{F}_t\). Roughly speaking, the \(\sigma\)-field \(\mathcal{F}_\tau\) captures the information up to time \(\tau\). The following lemmas help clarify the definition of \(\mathcal{F}_\tau\).

Lemma 3.6. \(\mathcal{F}_\tau = \mathcal{F}_t\) if \(\tau \equiv t\), \(\mathcal{F}_\tau = \mathcal{F}_\infty\) if \(\tau \equiv +\infty\) and \(\mathcal{F}_\tau \subseteq \mathcal{F}_\infty\) for any \(\tau\).

Proof. In the first case, note that \(F \cap \{\tau = s\}\) is empty if \(s \neq t\) and is \(F\) if \(s = t\). So if \(F \in \mathcal{F}_\tau\) then \(F = F \cap \{\tau = s\} \in \mathcal{F}_s\) and if \(F \in \mathcal{F}_s\) then \(F = F \cap \{\tau = s\} \in \mathcal{F}_s\). Moreover \(\emptyset \in \mathcal{F}_s\) so we have proved both inclusions. This works also for \(s = +\infty\).

For the third claim note that
\[
F = \bigcup_{s \in \mathbb{Z}_+} F \cap \{\tau = s\} \in \mathcal{F}_\infty.
\]

Lemma 3.7. If \((X_t)\) is adapted and \(\tau\) is a stopping time then \(X_\tau \in \mathcal{F}_\tau\) (where we assume that \(X_\infty \in \mathcal{F}_\infty\), e.g., \(X_\infty := \lim \inf X_n\)).
Proof. For $B \in \mathcal{E}$,

$$\{X_\tau \in B\} \cap \{\tau = t\} = \{X_t \in B\} \cap \{\tau = t\} \in \mathcal{F}_t.$$ 

\[\square\]

**Lemma 3.8.** If $\sigma, \tau$ are stopping times then $\mathcal{F}_{\sigma \wedge \tau} \subseteq \mathcal{F}_\tau$.

**Proof.** Let $F \in \mathcal{F}_{\sigma \wedge \tau}$. Note that

$$F \cap \{\tau = t\} = \bigcup_{s \leq t} ([F \cap \{\sigma \wedge \tau = s\}] \cap \{\tau = t\}) \in \mathcal{F}_t.$$ 

Indeed, the expression in parenthesis is in $\mathcal{F}_s \subseteq \mathcal{F}_t$ and $\{\tau = t\} \in \mathcal{F}_t$. \[\square\]

Let $(X_t)$ be a Markov chain on a countable space $V$. The following two examples of stopping time will play an important role.

**Definition 3.9** (First visit and return). The first visit time and first return time to $x \in V$ are

$$\tau_x := \inf\{t \geq 0 : X_t = x\} \quad \text{and} \quad \tau_x^+ := \inf\{t \geq 1 : X_t = x\}.$$ 

Similarly, $\tau_B$ and $\tau_B^+$ are the first visit and first return to $B \subseteq V$.

**Definition 3.10** (Cover time). Assume $V$ is finite. The cover time of $(X_t)$ is the first time that all states have been visited, i.e.,

$$\tau_{\text{cov}} := \inf\{t \geq 0 : \{X_0, \ldots, X_t\} = V\}.$$ 

**Strong Markov property** Let $(X_t)$ be a Markov chain with transition matrix $P$ and initial distribution $\mu$. Let $\mathcal{F}_t = \sigma(X_0, \ldots, X_t)$. Recall that the Markov property says that, given the present, the future is independent of the past. The Markov property naturally extends to stopping times. Let $\tau$ be a stopping time with $\mathbb{P}[\tau < +\infty] > 0$. In its simplest form we have: $\mathbb{P}[X_{t+1} = y | \mathcal{F}_\tau] = \mathbb{P}_{X_\tau}[X_{t+1} = y] = P(X_\tau, y)$. More generally:

**Theorem 3.11** (Strong Markov property). Let $f_t : V^\infty \to \mathbb{R}$ be a sequence of measurable functions, uniformly bounded in $t$ and let $F_t(x) := \mathbb{E}_x[f_t((X_t)_{t \geq 0})]$, then

$$\mathbb{E}[f_\tau((X_{\tau+t})_{t \geq 0}) | \mathcal{F}_\tau] = F_\tau(X_\tau) \quad \text{on} \{\tau < +\infty\}.$$ 

101
Proof. Let $A \in \mathcal{F}_\tau$. Summing over the possible values of $\tau$ and using the Markov property

$$
\mathbb{E}[f_\tau((X_{\tau+t})_{t \geq 0}); A \cap \{\tau < +\infty\}] = \sum_{s \geq 0} \mathbb{E}[f_s((X_{s+t})_{t \geq 0}); A \cap \{\tau = s\}]
$$

$$
= \sum_{s \geq 0} \mathbb{E}[F_s(X_s); A \cap \{\tau = s\}]
$$

$$
= \mathbb{E}[F_\tau(X_\tau); A \cap \{\tau < +\infty\}].
$$

That concludes the proof.

The following typical application of the strong Markov property is useful.

**Theorem 3.12 (Reflection principle).** Let $X_1, X_2, \ldots$ be i.i.d. with a distribution symmetric about 0 and let $S_t = \sum_{i \leq t} X_i$. Then, for $b > 0$,

$$
\mathbb{P}\left[\sup_{i \leq t} S_i \geq b\right] \leq 2 \mathbb{P}[S_t \geq b].
$$

**Proof.** Let $\tau := \inf\{i \leq t : S_i \geq b\}$. By the strong Markov property, on $\{\tau < t\}$, $S_t - S_\tau$ is independent on $\mathcal{F}_\tau$ and is symmetric about 0. In particular, it has probability at least $1/2$ of being greater or equal to 0, which implies that $S_t$ is greater or equal to $b$. Hence

$$
\mathbb{P}[S_t \geq b] \geq \mathbb{P}[\tau = t] + \frac{1}{2} \mathbb{P}[\tau < t] \geq \frac{1}{2} \mathbb{P}[\tau \leq t].
$$

(The reader may want to try to prove this rigorously using the formal statement of the strong Markov property. Then read the proof of [Dur10, Theorem 6.3.5].)

In the case of simple random walk on $\mathbb{N}$, we get a stronger statement.

**Theorem 3.13 (Reflection principle: simple random walk).** Let $(S_t)$ be simple random walk on $\mathbb{Z}$. Then, $\forall a, b, t > 0$,

$$
\mathbb{P}_0[S_t = b + a] = \mathbb{P}_0\left[ S_t = b - a, \sup_{i \leq t} S_i \geq b \right].
$$

Summing over $a > 0$ and rearranging gives

$$
\mathbb{P}_0\left[ \sup_{i \leq t} S_i \geq b \right] = \mathbb{P}_0[S_t = b] + 2 \mathbb{P}_0[S_t > b].
$$

**Proof.** Reflect the sub-path after the first visit to $b$ across the line $y = b$.

102
We record another related result that will be useful later (see e.g. [Dur10, Theorem 4.3.2]).

**Theorem 3.14** (Ballot theorem). In an election with \( n \) voters, candidate \( A \) gets \( \alpha \) votes and candidate \( B \) gets \( \beta < \alpha \) votes. The probability that \( A \) leads \( B \) throughout the counting is \( \frac{\alpha - \beta}{n} \).

**Recurrence** Let \((X_t)\) be a Markov chain on a countable state space \( V \). The *time of the \( k \)-th return to \( y \)* is (letting \( \tau_0^y := 0 \))

\[
\tau^k_y := \inf\{ t > \tau^k_{y-1} : X_t = y \}. 
\]

In particular, \( \tau^1_y \equiv \tau^+_y \). Define \( \rho_{xy} := P_x[\tau^+_y < +\infty] \). Then by the strong Markov property

\[ P_x[\tau^+_y < +\infty] = \rho_{xy} \rho^{k-1}_{yy}. \]

Letting \( N_y := \sum_{t \geq 0} 1(X_t = y) \), by linearity \( E_x[N_y] = \frac{\rho_{xy}}{1 - \rho_{yy}} \). So either \( \rho_{yy} < 1 \) and \( E_{\tau_y^y}[N_y] \to +\infty \) or \( \rho_{yy} = 1 \) and \( \tau^k_y < +\infty \) a.s. for all \( k \). That leads us to the following definition.

**Definition 3.15** (Recurrence). A state \( x \) is *recurrent* if \( \rho_{xx} = 1 \). Otherwise it is *transient*. We refer to the recurrence or transience of a state as its *type*. A chain is recurrent or transient if all its states are. If \( x \) is recurrent and \( E_{\tau^+_x}[\tau^+_x] < +\infty \), we say that \( x \) is positive recurrent.

Recurrence is “contagious” in the following sense (see e.g. [Dur10, Theorem 6.4.3]).

**Lemma 3.16.** If \( x \) is recurrent and \( \rho_{xy} > 0 \) then \( y \) is recurrent and \( \rho_{yx} = \rho_{xy} = 1 \).

A subset \( C \subseteq V \) is *closed* if \( x \in C \) and \( \rho_{xy} > 0 \) implies \( y \in C \). A subset \( D \subseteq V \) is *irreducible* if \( x, y \in D \) implies \( \rho_{xy} > 0 \). Recall that we have the following decomposition theorem (see e.g. [Dur10, Theorem 6.4.5]).

**Theorem 3.17** (Decomposition theorem). Let \( R := \{ x : \rho_{xx} = 1 \} \) be the recurrent states of the chain. Then \( R \) can be written as a disjoint union \( \bigcup_j R_j \) where each \( R_j \) is closed and irreducible.

**Example 3.18** (Simple random walk on \( \mathbb{Z} \)). Consider simple random walk on \( \mathbb{Z} \). The chain is clearly irreducible so it suffices to check the type of the state 0. First note the periodicity of this chain. So we look at \( S_{2t} \). Then by Stirling’s formula

\[
P_0[S_{2t} = 0] = \binom{2t}{t} 2^{-2t} \sim 2^{-2t} \frac{(2t)^{2t}}{(t!)^2} \frac{\sqrt{2t}}{\sqrt{2\pi t}} \sim \frac{1}{\sqrt{\pi t}}.
\]

\[ 103 \]
Thus
\[ \mathbb{E}_0[N_0] = \sum_{t>0} \mathbb{P}_0[S_t = 0] = +\infty, \]
and the chain is recurrent.

Return times provide insight into stationary measures. Recall (see e.g. [Dur10, Theorems 6.5.2-5]):

**Theorem 3.19.** Let \( x \) be a recurrent state. Then the following defines a stationary measure
\[
\mu_x(y) := \mathbb{E}_x \left[ \sum_{0 \leq t < \tau_x^+} 1_{\{X_t = y\}} \right].
\]

**Theorem 3.20.** If \( (X_t) \) is irreducible and recurrent, then the stationary measure is unique up to a constant multiple.

**Theorem 3.21.** If there is a stationary distribution \( \pi \) then all states \( y \) that have \( \pi(y) > 0 \) are recurrent.

**Theorem 3.22.** If \( (X_t) \) is irreducible and has a stationary distribution \( \pi \), then
\[
\pi(x) = \frac{1}{\mathbb{E}_x \tau_x^+}.
\]

**A useful identity**  A slight generalization of the “cycle trick” in the proof of Theorem 3.19 gives a useful identity.

**Lemma 3.23** (Occupation measure identity). Consider an irreducible Markov chain \( (X_t)_t \) with transition matrix \( P \) and stationary distribution \( \pi \). Let \( x \) be a state and \( \sigma \) be a stopping time such that \( \mathbb{E}_x[\sigma] < +\infty \) and \( \mathbb{P}_x[X_\sigma = x] = 1 \). Denote by \( G_\sigma(x, y) \) the expected number of visits to \( y \) before \( \sigma \) when started at \( x \), known as the Green function. For any \( y \),
\[
G_\sigma(x, y) = \pi_y \mathbb{E}_x[\sigma].
\]

**Proof.** By the uniqueness of the stationary distribution, it suffices to show that
\[
\sum_y G_\sigma(x, y) P(y, z) = G_\sigma(x, z), \forall z,
\]
and use the fact that \( \sum_y G_\sigma(x, y) = \mathbb{E}_x[\sigma] \).
To check this, because $X_\sigma = X_0$, observe that

$$G(x, z) = E_x \left[ \sum_{0 \leq t < \sigma} 1_{X_t = z} \right]$$

$$= E_x \left[ \sum_{0 \leq t < \sigma} 1_{X_{t+1} = z} \right]$$

$$= \sum_{t \geq 0} \mathbb{P}_x [X_{t+1} = z, \sigma > t].$$

Since $\{\sigma > t\} \in \mathcal{F}_t$, applying the Markov property we get

$$G(x, z) = \sum_{t \geq 0} \sum_y \mathbb{P}_x [X_t = y, X_{t+1} = z, \sigma > t]$$

$$= \sum_{t \geq 0} \sum_y \mathbb{P}_x [X_{t+1} = z | X_t = y, \sigma > t] \mathbb{P}_x [X_t = y, \sigma > t]$$

$$= \sum_{t \geq 0} \sum_y P(y, z) \mathbb{P}_x [X_t = y, \sigma > t]$$

$$= \sum_y G(x, y) P(y, z).$$

That proves the claim. 

Here is a typical application of this lemma.

**Corollary 3.24.** *In the setting of Lemma 3.23, for all $x \neq y$,*

$$P_x [\tau_y < \tau_x^+] = \frac{1}{\tau_x (E_x[\tau_y] + E_y[\tau_x])}.$$  

**Proof.** Let $\sigma$ be the time of the first visit to $x$ after the first visit to $x$. Then $E_x[\sigma] = E_x[\tau_y] + E_y[\tau_x] < +\infty$, where we used that the network is finite and connected. The number of visits to $x$ before the first visit to $y$ is geometric with success probability $P_x [\tau_y < \tau_x^+]$. Moreover the number of visits to $x$ after the first visit to $y$ but before $\sigma$ is 0 by definition. Hence $G_x(x, y)$ is the mean of the geometric, namely $1/P_x [\tau_y < \tau_x^+]$. Applying the occupation measure identity gives the result.

Many more results of this form can be derived from the occupation measure identity. See e.g. [AF, Chapter 2].
3.1.2 Markov chains: exponential tail of hitting times and some cover time bounds

Tail of a hitting time
On a finite state space, the tail of a hitting time converges to 0 exponentially fast.

Lemma 3.25. Let \((X_t)\) be a finite, irreducible Markov chain with state space \(V\) and initial distribution \(\mu\). For \(A \subseteq V\), there is \(\beta_1 > 0\) and \(0 < \beta_2 < 1\) depending on \(A\) such that
\[
\Pr_{\mu}[\tau_A > t] \leq \beta_1 \beta_2^t.
\]
In particular, \(\mathbb{E}_{\mu}[\tau_A] < +\infty\) for any \(\mu, A\).

Proof. For any integer \(m\), for some distribution \(\theta\),
\[
\Pr_{\mu}[\tau_A > ms \mid \tau_A > (m - 1)s] = \Pr_{\theta}[\tau_A > s] \leq \max_x \Pr_x[\tau_A > s] = 1 - \alpha_s.
\]
Choose \(s\) large enough that, from any \(x\), there is a path to \(A\) of length at most \(s\) of positive probability. Such an \(s\) exists by irreducibility. In particular \(\alpha_s > 0\). By induction,
\[
\Pr_{\mu}[\tau_A > ms] \leq (1 - \alpha_s)^m \quad \text{or} \quad \Pr_{\mu}[\tau_A > t] \leq (1 - \alpha_s)^{\lfloor \frac{t}{s} \rfloor} \leq \beta_1 \beta_2^t \text{ for } \beta_1 > 0 \text{ and } 0 < \beta_2 < 1 \text{ depending on } \alpha_s.
\]
The result for the expectation follows from
\[
\mathbb{E}_{\mu}[\tau_A] = \sum_{k \geq 0} \Pr_{\mu}[\tau_A > k] \leq \sum_t \beta_1 \beta_2^t < +\infty.
\]
That concludes the proof. \(\blacksquare\)

Here is a more precise bound in terms of the maximum hitting time.

Lemma 3.26. Let \((X_t)\) be a finite, irreducible Markov chain with state space \(V\) and initial distribution \(\mu\). For \(A \subseteq V\), let \(\bar{t}_A := \max_x \mathbb{E}_x[\tau_A]\). Then
\[
\Pr_{\mu}[\tau_A > t] \leq \exp\left(-\frac{t}{[e \cdot t_A]}\right).
\]

Proof. For any integer \(m\), for some distribution \(\theta\),
\[
\Pr_{\mu}[\tau_A > ms \mid \tau_A > (m - 1)s] = \Pr_{\theta}[\tau_A > s] \leq \max_x \Pr_x[\tau_A > s] \leq \frac{\bar{t}_A}{s},
\]
by the Markov property and Markov’s inequality (Theorem 2.1). By induction,
\[
\Pr_{\mu}[\tau_A > ms] \leq \left(\frac{\bar{t}_A}{s}\right)^m \quad \text{or} \quad \Pr_{\mu}[\tau_A > t] \leq \left(\frac{\bar{t}_A}{s}\right)^{\lfloor \frac{t}{s} \rfloor} .
\]
By differentiating w.r.t. \(s\), it can be checked that a good choice is \(s = \lfloor e \cdot t_A \rfloor\). \(\blacksquare\)
Application to cover times  We give an application of the previous bound to cover times. Let $(X_t)$ be a finite, irreducible Markov chain on $V$ with $n := |V| > 1$. Recall that the cover time is $\tau_{\text{cov}} := \max_y \tau_y$. We bound the mean cover time in terms of $t_{\text{hit}} := \max_{x \neq y} E_x \tau_y$.

Claim 3.27.

$$\max_x E_x[\tau_{\text{cov}}] \leq (3 + \ln n) [e t_{\text{hit}}].$$

Proof.  By a union bound (Corollary 2.11) over all states to be visited and Lemma 3.26,

$$\max_x P_x[\tau_{\text{cov}} > t] \leq \min \left\{1, n \cdot \exp\left(-\left|\frac{t}{e t_{\text{hit}}}\right|\right)\right\}.$$ 

Summing over $t$ and appealing to the sum of a geometric series,

$$\max_x E_x[\tau_{\text{cov}}] \leq (\ln(n) + 1)[e t_{\text{hit}}] + \frac{1}{1 - e^{-1}} [e t_{\text{hit}}].$$

A clever argument gives a better constant as well as a lower bound.

Theorem 3.28 (Matthews’ cover time bounds). Let $t^A_{\text{hit}} := \min_{x,y \in A, x \neq y} E_x \tau_y$ and $h_n := \sum_{m=1}^{n} \frac{1}{m}$. Then

$$\max_x E_x[\tau_{\text{cov}}] \leq h_n \ t_{\text{hit}}, \quad (3.1)$$

and

$$\min_x E_x[\tau_{\text{cov}}] \geq \max_{A \subseteq V, |A| - 1} h^A_{|A|-1} \ t_{\text{hit}}. \quad (3.2)$$

Clearly, $\max_{x \neq y} t^A_{\text{hit}}$ is a lower bound on the worst cover time. Lower bound (3.2) says that a tighter bound is obtained by finding a large subset of vertices $A$ that are far away from each other.

Proof.  We prove the lower bound for $A = V$. The other cases are similar. Let $(J_1, \ldots, J_n)$ be a uniform random ordering of $V$, let $C_m := \max_{i \leq J_m} \tau_i$, and let $L_m$ be the last state visited among $J_1, \ldots, J_m$. Then

$$E_x[C_m - C_{m-1} \mid J_1, \ldots, J_m, \{X_t, t \leq C_{m-1}\}] \geq t^V_{\text{hit}} 1\{L_m = J_m\}.$$ 

By symmetry, $P[L_m = J_m] = \frac{1}{m}$. To see this, first pick the set of vertices corresponding to $\{J_1, \ldots, J_m\}$, wait for all of these vertices to be visited, then pick the ordering. Moreover observe that $E_x C_1 \geq (1 - \frac{1}{m}) t^V_{\text{hit}}$ where the factor of $(1 - \frac{1}{m})$ accounts for the probability that $J_1 \neq x$. Taking expectations above and summing over $m$ gives the result. ■
Remark 3.29. The bounds (3.1) and (3.2) are tight up to smaller order for the coupon collector problem, which can be stated in terms of the cover time of a lazy random walk on the complete graph.

3.1.3 Martingales

Definition We first recall the definition of a martingale.

Definition 3.30 (Martingale). An adapted process \( \{M_t\}_{t \geq 0} \) with \( \mathbb{E}|M_t| < +\infty \) for all \( t \) is a martingale if

\[
\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = M_t, \quad \forall t \geq 0
\]

If the equality is replaced with \( \leq \) or \( \geq \), we get a supermartingale or a submartingale respectively. We say that a martingale is bounded in \( L^p \) if \( \sup_n \mathbb{E}[|X_n|^p] < +\infty \).

Note that for a martingale, by the tower property (Lemma A.5), we have \( \mathbb{E}[M_n \mid \mathcal{F}_m] = M_m \) for all \( n > m \), and similarly for supermartingales and submartingales.

Jensen’s inequality immediately implies:

Lemma 3.31. If \( \{M_t\}_{t \geq 0} \) is a martingale and \( \phi \) is a convex function with \( \mathbb{E}|\phi(M_t)| < +\infty \) for all \( t \), then \( \{\phi(M_t)\}_{t \geq 0} \) is a submartingale. Moreover, if \( \{M_t\}_{t \geq 0} \) is a submartingale and \( \phi \) is an increasing convex function with \( \mathbb{E}|\phi(M_t)| < +\infty \) for all \( t \), then \( \{\phi(M_t)\}_{t \geq 0} \) is a submartingale.

We start with a straightforward example.

Example 3.32 (Sums of i.i.d. random variables with mean 0). Let \( X_0, X_1, \ldots \) be i.i.d. centered random variables, \( F_t = \sigma(X_0, \ldots, X_t) \) and \( S_t = \sum_{i \leq t} X_i \). Note that \( \mathbb{E}|S_t| < \infty \) by the triangle inequality and

\[
\mathbb{E}[S_{t+1} \mid F_t] = \mathbb{E}[S_{t+1} + X_{t+1} \mid F_t] = S_t + \mathbb{E}[X_{t+1}] = S_t,
\]

which proves that \( (S_t) \) is a martingale.

Martingales can also be a little more hidden.

Example 3.33 (Variance of a sum of i.i.d. random variables). Consider the same setup as the previous example with \( \sigma^2 := \text{Var}[X_1] < \infty \). Define \( M_t = S_t^2 - t\sigma^2 \).

Note that \( \mathbb{E}|M_t| \leq 2t\sigma^2 < +\infty \) and

\[
\mathbb{E}[M_t \mid F_{t-1}] = \mathbb{E}[(X_t + S_{t-1})^2 - t\sigma^2 \mid F_{t-1}]
= \mathbb{E}[X_t^2 + 2X_tS_{t-1} + S_{t-1}^2 - t\sigma^2 \mid F_{t-1}]
= \sigma^2 + 0 + S_{t-1}^2 - t\sigma^2 = M_{t-1},
\]

which proves that \( (M_t) \) is a martingale.
Or we can create martingales out of thin air.

**Example 3.34** (Doob martingale: accumulating data). Let \( X \) with \( \mathbb{E}|X| < +\infty \). Define \( M_t = \mathbb{E}[X | \mathcal{F}_t] \). Note that \( \mathbb{E}|M_t| \leq \mathbb{E}|X| < +\infty \), and

\[
\mathbb{E}[M_t | \mathcal{F}_{t-1}] = \mathbb{E}[X | \mathcal{F}_{t-1}] = M_{t-1},
\]

by the tower property (Lemma A.5). This is known as a *Doob martingale*.

**Convergence** The following is a key result about martingales (see e.g. [Wil91, Section 11.5]).

**Theorem 3.35** (Doob’s martingale convergence theorem). Let \((X_t)\) be a supermartingale bounded in \( L^1 \). Then \((X_t)\) converges a.s. to a finite limit \( X_\infty \). Moreover, letting \( X_\infty := \limsup_n X_n \), then \( X_\infty \in \mathcal{F}_\infty \) and \( \mathbb{E}|X_\infty| < +\infty \).

**Corollary 3.36** (Convergence of nonnegative martingales). If \((X_t)\) is a nonnegative martingale then \( X_t \) converges a.s.

*Proof.* \((X_t)\) is bounded in \( L^1 \) since

\[
\mathbb{E}|X_t| = \mathbb{E}|X_t| = \mathbb{E}|X_0|, \forall t.
\]

\[\blacklozenge\]

**Example 3.37** (Pólya’s urn). An urn contains 1 red ball and 1 green ball. At each time, we pick one ball and put it back with an extra ball of the same color. This process is known as *Pólya’s urn*. Let \( R_t \) (respectively \( G_t \)) be the number of red balls (respectively green balls) after the \( t \)th draw. Let

\[
\mathcal{F}_t := \sigma(R_0, G_0, R_1, G_1, \ldots, R_t, G_t).
\]

Define \( M_t \) to be the fraction of green balls after the \( t \)th draw. Then

\[
\mathbb{E}[M_t | \mathcal{F}_{t-1}] = \frac{R_{t-1}}{G_{t-1} + R_{t-1}} \frac{G_{t-1}}{G_{t-1} + R_{t-1} + 1} + \frac{G_{t-1}}{G_{t-1} + R_{t-1}} \frac{G_{t-1} + 1}{G_{t-1} + R_{t-1} + 1}
\]

\[
= \frac{G_{t-1}}{G_{t-1} + R_{t-1}} = \frac{G_{t-1}}{G_{t-1} + R_{t-1}} = M_{t-1}.
\]
Since $M_t \geq 0$ and is a martingale, we have $M_t \to M_\infty$ a.s. Observe further that
\[ \mathbb{P}[G_t = m + 1] = \binom{t}{m} \frac{m!(t-m)!}{(t+1)!} = \frac{1}{t+1}, \]
so that
\[ \mathbb{P}[M_t \leq x] = \frac{x(t+2) - 1}{t+1} \to x, \]
by a sandwiching argument. That is, $(M_t)$ converges in distribution to a uniform random variable on $[0, 1]$.

\section*{Some useful properties}

We collect here a few more properties of martingales.

\textbf{Theorem 3.38 (Doob’s submartingale inequality).} Let $(M_t)$ be a nonnegative submartingale. Then for $b > 0$
\[ \mathbb{P} \left[ \sup_{1 \leq s \leq t} M_s \geq b \right] \leq \frac{\mathbb{E}[M_t]}{b}. \]

Observe that Markov’s inequality (Theorem 2.1) implies only that
\[ \sup_{1 \leq s \leq t} \mathbb{P}[M_s \geq b] \leq \frac{\mathbb{E}[M_t]}{b}, \]
where we used that $\mathbb{E}[M_t] \geq \mathbb{E}[M_s]$ for all $1 \leq s \leq t$.

\textit{Proof.} Divide $F = \{ \sup_{1 \leq s \leq t} M_s \geq b \}$ according to the first time $M_i$ crosses $b$:
\[ F = F_0 \cup \cdots \cup F_t, \]
where
\[ F_s = \{ M_0 < b \} \cap \cdots \cap \{ M_{s-1} < b \} \cap \{ M_s \geq b \}. \]
Since $F_s \in \mathcal{F}_s$ and $\mathbb{E}[M_t | \mathcal{F}_s] \geq M_s$, we have
\[ b \mathbb{P}[F_s] \leq \mathbb{E}[M_s; F_s] \leq \mathbb{E}[M_t; F_s]. \]
Summing over $s$ gives the result.

A useful consequence of the previous inequality:

\textbf{Corollary 3.39 (Kolmogorov’s inequality).} Let $X_1, X_2, \ldots$ be independent random variables with $\mathbb{E}[X_i] = 0$ and $\text{Var}[X_i] < +\infty$. Define $S_t = \sum_{i \leq t} X_i$. Then for $\beta > 0$
\[ \mathbb{P} \left[ \max_{i \leq t} |S_i| \geq \beta \right] \leq \frac{\text{Var}[S_t]}{\beta^2}. \]
Proof. By Example 3.32, \((S_t)\) is a martingale. By Jensen’s inequality, \((S_t^2)\) is hence a submartingale. The result then follows from Doob’s submartingale inequality.

We will also need the following orthogonality property.

**Lemma 3.40** (Orthogonality of increments). Let \((M_t)\) be a martingale with \(M_t \in L^2\). Let \(s \leq t \leq u \leq v\). Then,

\[
\langle M_t - M_s, M_u - M_s \rangle = 0.
\]

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

**Proof.** Use \(M_u = \mathbb{E}[M_v | \mathcal{F}_u]\), \(M_t - M_s \in \mathcal{F}_u\) and apply the \(L^2\) characterization of conditional expectations.

Optional stopping Finally, we recall the important optional stopping theorem (see e.g. [Wil91, Section 10.10]).

**Definition 3.41.** Let \(\{M_t\}\) be an adapted process and \(\sigma\) be a stopping time. Then

\[
M_\sigma^\sigma(\omega) := M_{\sigma(\omega) \wedge t}(\omega),
\]

is \((M_t)\) stopped at \(\sigma\).

**Theorem 3.42.** Let \((M_t)\) be a supermartingale and \(\sigma\) be a stopping time. Then the stopped process \((M_\sigma^\sigma)\) is a supermartingale and in particular

\[
\mathbb{E}[M_{\sigma \wedge t}] \leq \mathbb{E}[M_0].
\]

The same result holds with equality if \((M_t)\) is a martingale.

**Theorem 3.43** (Doob’s optional stopping theorem). Let \((M_t)\) be a supermartingale and \(\sigma\) be a stopping time. Then \(M_\sigma\) is integrable and

\[
\mathbb{E}[M_\sigma] \leq \mathbb{E}[M_0],
\]

if one of the following holds:

1. \(\sigma\) is bounded
2. \((M_t)\) is uniformly bounded and \(\sigma\) is a.s. finite
3. \(\mathbb{E}[\sigma] < +\infty\) and \((M_t)\) has bounded increments (i.e., there \(c > 0\) such that \(|M_t - M_{t-1}| \leq c\) a.s. for all \(t\))
4. \((M_t)\) is nonnegative and \(\sigma\) is a.s. finite.

The first three imply equality above if \((M_t)\) is a martingale.
Gambler’s ruin  Although the optional stopping theorem as stated in Theorem 3.43 is occasionally useful, one often works directly with Theorem 3.42 and applies suitable limit theorems. The following martingale-based proof of Wald’s first identity provides an illustration (see also [Dur10, Theorem 4.1.5] for an alternative proof).

**Theorem 3.44** (Wald’s first identity). Let $X_1, X_2, \ldots \in L^1$ be i.i.d. with $\mathbb{E}[X_1] = \mu$ and let $\tau \in L^1$ be a stopping time. Let $S_t = \sum_{s=1}^t X_s$. Then

$$\mathbb{E}[S_\tau] = \mathbb{E}[X_1] \mathbb{E}[\tau].$$

**Proof.** We first prove the result for nonnegative $X_i$s. By Example 3.32, $S_t - t \mathbb{E}[X_1]$ is a martingale and Theorem 3.42 implies that $\mathbb{E}[S_{\tau \wedge t}] = \mathbb{E}[X_1] \mathbb{E}[\tau \wedge t]$. Note that we have $S_{\tau \wedge t} \uparrow S_\tau$ and $\tau \wedge t \uparrow \tau$. Thus by monotone convergence $\mathbb{E}[S_\tau] = \mathbb{E}[X_1] \mathbb{E}[\tau]$.

Consider now the general case. Again, $\mathbb{E}[S_{\tau \wedge t}] = \mathbb{E}[X_1] \mathbb{E}[\tau \wedge t]$ and $\mathbb{E}[\tau \wedge t] \uparrow \mathbb{E}[\tau]$. Applying the previous argument to $R_t = \sum_{s=1}^t |X_s|$ shows that $\mathbb{E}[R_\tau] = \mathbb{E}[X_1] \mathbb{E}[\tau] < +\infty$ by assumption. Since $S_{\tau \wedge t} \leq R_t$ for all $t$ by the triangle inequality, dominated convergence implies $\mathbb{E}[S_{\tau \wedge t}] \to \mathbb{E}[S_\tau]$ and we are done. □

We also recall Wald’s second identity (see e.g. [Dur10, Theorem 4.1.6]).

**Theorem 3.45** (Wald’s second identity). Let $X_1, X_2, \ldots \in L^2$ be i.i.d. with $\mathbb{E}[X_1] = 0$ and $\text{Var}[X_1] = \sigma^2$ and let $\tau \in L^1$ be a stopping time. Then

$$\mathbb{E}[S_\tau^2] = \sigma^2 \mathbb{E}[\tau].$$

We illustrate Wald’s identities on an important example.

**Example 3.46** (Gambler’s ruin: unbiased case). Let $(S_t)$ be simple random walk on $\mathbb{Z}$ started at 0 and let $\tau = \tau_a \wedge \tau_b$ where $a < 0 < b$.

**Claim 3.47.** We have:

1) $\tau < +\infty$ a.s.

2) $\mathbb{P}[\tau_a < \tau_b] = \frac{b}{b-a}$

3) $\mathbb{E}[\tau] = -ab$

4) $\tau_a < +\infty$ a.s. but $\mathbb{E}[\tau_a] = +\infty$.

**Proof.** We prove the claims in order.

1) We argue that in fact $\mathbb{E}[\tau] < \infty$. That follows immediately from the exponential tail of hitting times in Lemma 3.25 for the chain $(S_{\tau \wedge t})$ whose state space, $\{a, a + 1, \ldots, b\}$, is finite.

112
2) By Wald’s first identity, $\mathbb{E}[S_\tau] = 0$ or

$$a \mathbb{P}[S_\tau = a] + b \mathbb{P}[S_\tau = b] = 0,$$

that is,

$$\mathbb{P}[\tau_a < \tau_b] = \frac{b}{b-a} \quad \text{and} \quad \mathbb{P}[\tau_a < +\infty] \geq \mathbb{P}[\tau_a < \tau_b] \to 1,$$

where we took $b \to \infty$ in the first expression to obtain the second one.

3) Because $\sigma^2 = 1$, Wald’s second identity says that $\mathbb{E}[S_\tau^2] = \mathbb{E}[\tau]$. Furthermore, we have by 2)

$$\mathbb{E}[S_\tau^2] = \frac{b}{b-a} a^2 + \frac{-a}{b-a} b^2 = -ab.$$

Thus $\mathbb{E}[\tau] = -ab$.

4) The first claim was proved in 2). When $b \to +\infty$, $\tau = \tau_a \wedge \tau_b \uparrow \tau_a$ and monotone convergence applied to 3) gives that $\mathbb{E}[\tau_a] = +\infty$.

That concludes the proof.

Note that 4) shows that the $L^1$ condition on the stopping time in Wald’s second identity is necessary. Indeed we have shown $a^2 = \mathbb{E}[S_\tau^2] \neq \sigma^2 \mathbb{E}[\tau] = +\infty$. □

**Example 3.48** (Gambler’s ruin: biased case). The *biased random walk on $\mathbb{Z}$* with parameter $1/2 < p < 1$ is the process $(S_t)$ with $S_0 = 0$ and $S_t = \sum_{s \leq t} X_s$ where the $X_s$s are i.i.d. in $\{-1, +1\}$ with $\mathbb{P}[X_1 = 1] = p$. Let $\tau = \tau_a \wedge \tau_b$ where $a < 0 < b$. Let $q := 1 - p$ and $\phi(x) := (q/p)^x$.

**Claim 3.49.** We have:

1) $\tau < +\infty$ a.s.

2) $\mathbb{P}[\tau_a < \tau_b] = \frac{\phi(b)-\phi(0)}{\phi(b)-\phi(a)}$

3) $\mathbb{E}[\tau_b] = \frac{b}{2p-1}$

4) $\tau_a = +\infty$ with positive probability.

**Proof.** Let $\psi_t(x) := x - (p-q)t$. We use two martingales: $(\phi(S_t))$ and $(\psi_t(S_t))$. Observe that indeed

$$\mathbb{E}[\phi(S_t) \mid \mathcal{F}_{t-1}] = p(q/p)^{S_{t-1}+1} + q(q/p)^{S_{t-1}-1} = \phi(S_{t-1}),$$

and

$$\mathbb{E}[\psi_t(S_t) \mid \mathcal{F}_{t-1}] = p[S_{t-1} + 1 - (p-q)t] + q[S_{t-1} - 1 - (p-q)t]
= \psi_{t-1}(S_{t-1}).$$
1) This claim follows by the same argument as in the unbiased case.

2) Note that \((\phi(S_{\tau\wedge t}))\) is a bounded martingale. Therefore, by Theorem 3.42 and dominated convergence,

\[ \phi(0) = \mathbb{E}[\phi(S_t)] = \mathbb{P}[\tau_a < \tau_b] \phi(a) + \mathbb{P}[\tau_a > \tau_b] \phi(b), \]

or, rearranging, \(\mathbb{P}[\tau_a < \tau_b] = \frac{\phi(b) - \phi(0)}{\phi(b) - \phi(a)}\). Taking \(b \to +\infty\), by monotonicity

\[ \mathbb{P}[\tau_a < +\infty] = \frac{1}{\phi(a)} < 1, \tag{3.3} \]

so that \(\tau_a = +\infty\) with positive probability.

3) By Theorem 3.42 again,

\[ 0 = \mathbb{E}[S_{\tau_b \wedge t} - (p - q)(\tau_b \wedge t)]. \]

By monotone convergence, \(\mathbb{E}[\tau_b \wedge t] \uparrow \mathbb{E}[\tau_b]\). Furthermore, observe that

\[ -\inf_t S_t \geq 0 \text{ a.s. since } S_0 = 0 \text{ and, for } x \geq 0, \text{ by (3.3)} \]

\[ \mathbb{P}[-\inf_t S_t \geq x] = \mathbb{P}[\tau_x < +\infty] = \left(\frac{q}{p}\right)^x, \]

so that \(\mathbb{E}[-\inf_t S_t] = \sum_{x \geq 1} \mathbb{P}[-\inf_t S_t \geq x] < +\infty\). Hence, we can use dominated convergence with \(|S_{\tau_b \wedge t}| \leq \max\{b, -\inf_t S_t\}\) to deduce that

\[ \mathbb{E}[\tau_b] = \mathbb{E}[S_{\tau_b}] = \frac{b}{2p-1}. \]

4) That claim was proved in 2).

That concludes the proof.

Note that, in 3), in order to apply Wald’s first identity we would have to prove that \(\tau_b \in L^1\).

**3.1.4 Percolation on trees: critical regime**

Consider bond percolation on the infinite \(d\)-regular tree \(T_d\) with density \(p = \frac{1}{d-1}\). Let \(X_n := |\partial_n \cap C_0|\), where \(\partial_n\) are the \(n\)-th level vertices and \(C_0\) is the open cluster of the root. The first moment method does not work in this case because

\[ \mathbb{E}X_n = d(d-1)^{n-1}p^n = \frac{d}{d-1} \not\to 0. \]

**Theorem 3.50.** \(|C_0| < +\infty \text{ a.s.}\)
Proof. Let \( b := d - 1 \) be the branching ratio. Because the root has a different branching ratio, we consider the descendants of its children. Let \( Z_n \) be the number of vertices in the open cluster of the first child of the root \( n \) levels below it and let \( \mathcal{F}_n = \sigma(Z_0, \ldots, Z_n) \). Then \( Z_0 = 1 \) and
\[
\mathbb{E}[Z_n \mid \mathcal{F}_{n-1}] = bpZ_{n-1} = Z_{n-1}.
\]
So \( (Z_n) \) is a nonnegative, integer-valued martingale and it converges to an a.s. finite limit. But, clearly, for any integer \( k > 0 \) and \( N \geq 0 \)
\[
P[Z_n = k, \forall n \geq N] = 0,
\]
so \( Z_{\infty} \equiv 0 \).

We give a more precise result that will be useful later. Consider the descendant subtree, \( T_1 \), of the first child, 1, of the root. Let \( \widetilde{C}_1 \) be the open cluster of 1 in \( T_1 \). Assume \( d \geq 3 \).

**Theorem 3.51.** \( P \left[ |\widetilde{C}_1| > k \right] \leq \frac{4\sqrt{2}}{\sqrt{k}}, \) for \( k \) large enough

Proof. Note first that \( \mathbb{E}[|\widetilde{C}_1|] = +\infty \) by summing over the levels. So we cannot use the first moment method directly to give a bound on the tail. Instead, we use Markov’s inequality (Theorem 2.1) on a stopped process. We use an exploration process with 3 types of vertices:

- \( A_t \): active vertices
- \( E_t \): explored vertices
- \( N_t \): neutral vertices

We start with \( A_0 := \{1\} \), \( E_0 := \emptyset \), and \( N_0 \) contains all other vertices in \( T_1 \). At time \( t \), if \( A_{t-1} = \emptyset \) we let \( (A_t, E_t, N_t) \) be \( (A_{t-1}, E_{t-1}, N_{t-1}) \). Otherwise, we pick a random element, \( a_t \), from \( A_{t-1} \) and: we set:

- \( A_t := A_{t-1} \cup \{ x \in N_{t-1} : \{ x, a_t \} \text{ is open} \} \setminus \{ a_t \} \)
- \( E_t := E_{t-1} \cup \{ a_t \} \)
- \( N_t := N_{t-1} \setminus \{ x \in N_{t-1} : \{ x, a_t \} \text{ is open} \} \)

Let \( M_t := |A_t| \). Revealing the edges as they are explored and letting \( (\mathcal{F}_t) \) be the corresponding filtration, we have \( \mathbb{E}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1} + bp - 1 = M_{t-1} + \) on
\{M_{t-1} > 0\}$ so $(M_t)$ is a nonnegative martingale. Let $\sigma^2 := b p (1 - p) \geq \frac{1}{2}$, $
abla := \inf \{t \geq 0 : M_t = 0\}$, and $Y_t := M_{t \wedge \tau}^2 - \sigma^2 (t \wedge \tau)$. Then, on $\{M_{t-1} > 0\}$,

\[
\mathbb{E}[Y_t | \mathcal{F}_{t-1}] = \mathbb{E}[\{(M_t - M_{t-1})^2 - \sigma^2 t | \mathcal{F}_{t-1}\}

= \mathbb{E}[M_{t-1}^2 + 2M_{t-1}(M_t - M_{t-1}) + (M_t - M_{t-1})^2 - \sigma^2 t | \mathcal{F}_{t-1}]

= M_{t-1}^2 + 2M_{t-1} \cdot 0 + \sigma^2 - \sigma^2 t = Y_{t-1},
\]

so $(Y_t)$ is also a martingale.

For $h > 0$, let

\[
\tau_h' := \inf \{t \geq 0 : M_t = 0 \text{ or } M_t \geq h\}.
\]

Note that $\tau_h' \leq \tau = |\tilde{C}| < +\infty$ a.s. We use 

\[
\mathbb{P}[\tau > k] = \mathbb{P}[M_t > 0, \forall t \in [k]] \leq \mathbb{P}[\tau_h' > k] + \mathbb{P}[M_{\tau_h'} \geq h].
\]

By Markov’s inequality (Theorem 2.1),

\[
\mathbb{P}[M_{\tau_h'} \geq h] \leq \frac{\mathbb{E}[M_{\tau_h'}]}{h},
\]

and

\[
\mathbb{P}[\tau_h' > k] \leq \frac{\mathbb{E}[\tau_h']}{k}.
\]

To compute $\mathbb{E}[M_{\tau_h'}]$, we use Theorem 3.42 to obtain

\[
1 = \mathbb{E}[M_{\tau_h' \wedge s}] = \mathbb{E}[M_{\tau_h'}],
\]

as $s \to +\infty$, where we used that $|M_{\tau_h' \wedge s}| \leq h + b$ and bounded convergence.

To compute $\mathbb{E}[\tau_h']$, we use Theorem 3.42 again

\[
1 = \mathbb{E}[M_{\tau_h' \wedge s}^2 - \sigma^2 (\tau_h' \wedge s)] = \mathbb{E}[M_{\tau_h'}^2] - \sigma^2 \mathbb{E}[\tau_h' \wedge s] \to \mathbb{E}[M_{\tau_h'}^2] - \sigma^2 \mathbb{E}[\tau_h'],
\]

as $s \to +\infty$ by bounded convergence again and monotone convergence respectively. Because

\[
\mathbb{E}[M_{\tau_h'}^2 | M_{\tau_h'} \geq h] \leq (h + b)^2,
\]

we have

\[
\mathbb{E}[\tau_h'] \leq \frac{1}{\sigma^2} \left\{ \frac{1}{h} \mathbb{E}[M_{\tau_h'}^2 | M_{\tau_h'} \geq h] \right\} \leq \frac{(h + b)^2}{\sigma^2 h} \leq \frac{2(h + b)^2}{h}.
\]

Take $h := \sqrt{\frac{k}{8}}$. For $k$ large enough, $h \geq b$ and

\[
\mathbb{P}[\tau > k] \leq \mathbb{P}[\tau_h' > k] + \mathbb{P}[M_{\tau_h'} \geq h] \leq \frac{8h}{k} + \frac{1}{h} = 2\sqrt{\frac{8}{k}}.
\]
3.2 Concentration for martingales and applications

The Chernoff-Cramér method extends naturally to martingales. This observation leads to powerful new concentration inequalities that hold far beyond the case of sums of independent variables. In particular, it will allow us to prove one version of the concentration phenomenon, which can be stated informally as:

If $X_1, \ldots, X_n$ are independent (or “weakly dependent”) random variables, then the random variable $f(X_1, \ldots, X_n)$ is “close” to its mean $\mathbb{E}f(X_1, \ldots, X_n)$ provided that the function $f(x_1, \ldots, x_n)$ is not too “sensitive” to any of the coordinates $x_i$.

3.2.1 Azuma-Hoeffding inequality

The main result of this section is the following generalization of Hoeffding’s inequality (Theorem 2.40).

**Theorem 3.52** (Maximal Azuma-Hoeffding inequality). Let $(Z_t)_{t \in \mathbb{Z}_+}$ be a martingale with respect to the filtration $(F_t)_{t \in \mathbb{Z}_+}$. Assume that there are predictable processes $(A_t)$ and $(B_t)$ (i.e., $A_t, B_t \in F_{t-1}$) and constants $0 < c_t < +\infty$ such that: for all $t \geq 1$, almost surely,

$$A_t \leq Z_t - Z_{t-1} \leq B_t \quad \text{and} \quad B_t - A_t \leq c_t.$$

Then for all $\beta > 0$

$$\mathbb{P}\left[ \sup_{0 \leq i \leq t} (Z_t - Z_0) \geq \beta \right] \leq \exp\left(-\frac{2\beta^2}{\sum_{i \leq t} c_i^2}\right).$$

Applying this inequality to $(-Z_t)$ gives a tail bound in the other direction.

**Proof of Theorem 3.52.** As in the Chernoff-Cramér method, we start by applying Markov’s inequality. Here we use the maximal version for submartingales, Doob’s submartingale inequality (Theorem 3.38). First notice that $e^{sx}$ is increasing and convex for $s > 0$, so that by Lemma 3.31 the process $(e^{s(Z_t - Z_0)})_t$ is a submartin-

---

*Requires: Section 2.4.3.
†Quoting [vH].
gale. Hence, for $s > 0$, by Theorem 3.38

$$
P \left[ \sup_{0 \leq i \leq t} (Z_i - Z_0) \geq \beta \right] = \mathbb{P} \left[ \sup_{0 \leq i \leq t} e^{s(Z_i - Z_0)} \geq e^{s\beta} \right]
\leq \frac{e^{s\beta}}{e^{s\beta}}
= \mathbb{E} \left[ e^{s \sum_{r=1}^{t} (Z_r - Z_{r-1})} \right].
$$

(3.4)

Unlike the Chernoff-Cramér case, however, the terms in the exponent are not independent. Instead, to exploit the martingale property, we condition on the filtration

$$
\mathbb{E} \left[ e^{s \sum_{r=1}^{t} (Z_r - Z_{r-1})} \mid F_{t-1} \right] = \mathbb{E} \left[ e^{s \sum_{r=1}^{t-1} (Z_r - Z_{r-1})} \mathbb{E} \left[ e^{s(Z_t - Z_{t-1})} \mid F_{t-1} \right] \right].
$$

The martingale property and the assumption in the statement implies that, conditioned on $F_{t-1}$, the random variable $Z_t - Z_{t-1}$ is centered and lies in an interval of length $c_t$. Hence by Hoeffding’s lemma (Lemma 2.42), it holds almost surely that

$$
\mathbb{E} \left[ e^{s(Z_t - Z_{t-1})} \mid F_{t-1} \right] \leq \exp \left( \frac{s^2 c_t^2}{4} \right) = \exp \left( \frac{c_t^2 s^2}{8} \right).
$$

(3.5)

Arguing by induction, we obtain

$$
\mathbb{E} \left[ e^{s(Z_t - Z_0)} \right] \leq \exp \left( \frac{s^2 \sum_{r \leq t} c_r^2}{8} \right).
$$

Put differently, we have proved that $Z_t - Z_0$ is sub-Gaussian with variance factor $\frac{1}{4} \sum_{r \leq t} c_r^2$. By (2.39) (or, equivalently, by choosing $s = \beta / \frac{1}{4} \sum_{r \leq t} c_r^2$ in (3.4)) we get the result.

In Theorem 3.52 the martingale difference sequence $(X_t)$, where $X_t := Z_t - Z_{t-1}$, is not only “pairwise uncorrelated” by Lemma 3.40, i.e.,

$$
\mathbb{E}[X_s X_r] = 0, \quad \forall r \neq s,
$$

but by the same argument it is in fact “mutually uncorrelated,”

$$
\mathbb{E}[X_{j_1} \cdots X_{j_k}] = 0, \quad \forall k \geq 1, \forall j_1 < \cdots < j_k.
$$

This much stronger property helps explain why $\sum_{r \leq t} X_r$ is highly concentrated. This point is the subject of Exercise 3.1, which guides the reader through a slightly different proof of the Azuma-Hoeffding inequality. Compare with Exercises 2.4 and 2.5.
3.2.2 Method of bounded differences

The power of the Azuma-Hoeffding inequality is that it produces tail inequalities for quantities other than sums of independent variables. The setting is the following. Let $X_1, \ldots, X_n$ be independent random variables where $X_i$ is $\mathcal{X}_i$-valued for all $i$ and let $X = (X_1, \ldots, X_n)$. Assume that $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ is a measurable function. Our goal is to characterize the concentration properties of $f(X)$ around its expectation in terms of its “discrete derivatives”

$$D_i f(x) := \sup_{y \in \mathcal{X}_i} f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) - \inf_{y' \in \mathcal{X}_i} f(x_1, \ldots, x_{i-1}, y', x_{i+1}, \ldots, x_n),$$

where $x = (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. We think of $D_i f(x)$ as a measure of the “sensitivity” of $f$ to its $i$-th coordinate.

**High-level idea** We begin with two easier bounds that we will improve below. To analyze the behavior of $f(X)$, the idea is to consider the Doob martingale (see Example 3.34)

$$Z_i = \mathbb{E}[f(X) \mid \mathcal{F}_i],$$

where $\mathcal{F}_i = \sigma(X_1, \ldots, X_i)$, which is well-defined provided $\mathbb{E}|f(X)| < +\infty$. Note that

$$Z_n = \mathbb{E}[f(X) \mid \mathcal{F}_n] = f(X),$$

and

$$Z_0 = \mathbb{E}[f(X)],$$

so that we can write

$$f(X) - \mathbb{E}[f(X)] = \sum_{i=1}^n (Z_i - Z_{i-1}).$$

A clever observation relates the martingale differences to the discrete derivatives through the use of an independent copy of $X$. Let $X' = (X'_1, \ldots, X'_n)$ be an independent copy of $X$ and let

$$X^{(i)} = (X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n).$$
Then

\[ Z_i - Z_{i-1} = \mathbb{E}[f(X) \mid \mathcal{F}_i] - \mathbb{E}[f(X) \mid \mathcal{F}_{i-1}] \]
\[ = \mathbb{E}[f(X) \mid \mathcal{F}_i] - \mathbb{E}[f(X^{(i)}) \mid \mathcal{F}_{i-1}] \]
\[ = \mathbb{E}[f(X) \mid \mathcal{F}_i] - \mathbb{E}[f(X^{(i)}) \mid \mathcal{F}_i] \]
\[ = \mathbb{E}[f(X) - f(X^{(i)}) \mid \mathcal{F}_i]. \]

Note that we crucially used the independence of the \( X_k \)s in the second and third lines. But then, by Jensen’s inequality,

\[ |Z_i - Z_{i-1}| \leq \|D_i f\|_\infty. \quad (3.7) \]

By the orthogonality of increments of martingales in \( L^2 \) (Lemma 3.40), we immediately obtain

\[ \text{Var}[f(X)] = \mathbb{E}[(Z_n - Z_0)^2] = \sum_{i=1}^n \mathbb{E}[(Z_i - Z_{i-1})^2] \leq \sum_{i=1}^n \|D_i f\|_\infty^2. \]

Moreover, by the Azuma-Hoeffding inequality (Theorem 3.52) and the fact that \( Z_i - Z_{i-1} \in [-\|D_i f\|_\infty, \|D_i f\|_\infty] \),

\[ P[f(X) - \mathbb{E}[f(X)] \geq \beta] \leq \exp \left( -\frac{\beta^2}{2 \sum_{i=1}^n \|D_i f\|_\infty^2} \right). \]

A more careful analysis, which we detail below, leads to better bounds.

We emphasize that, although it may not be immediately obvious, independence plays a crucial role in the bound (3.7), as the next example shows.

**Example 3.53 (A counterexample).** Let \( f(x_1, \ldots, x_n) = x_1 + \cdots + x_n \) where \( x_i \in \{-1, 1\} \) for all \( i \). Then,

\[ \|D_1 f\|_\infty = \sup_{x_2, \ldots, x_n} [(1 + x_2 + \cdots + x_n) - (-1 + x_2 + \cdots + x_n)] = 2, \]

and similarly \( \|D_i f\|_\infty = 2 \) for \( i = 2, \ldots, n \). Let \( X_1 \) be a uniform random variable on \( \{-1, 1\} \). First consider the case where we set \( X_2, \ldots, X_n \) all equal to \( X_1 \). Then

\[ \mathbb{E}[f(X_1, \ldots, X_n)] = 0, \]

and

\[ \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1] = nX_1, \]

120
so that

$$|\mathbb{E}[f(X_1, \ldots, X_n) | X_1] - \mathbb{E}[f(X_1, \ldots, X_n)]| = n > 2.$$ 

In particular, the corresponding Doob martingale does not have increments bounded by 2.

For a less extreme example which has support over all of \(\{-1, 1\}^n\), let

$$U_i = \begin{cases} 1, & \text{w.p. } 1 - \varepsilon \\ -1, & \text{w.p. } \varepsilon, \end{cases}$$

for some \(\varepsilon > 0\) independently for all \(i = 1, \ldots, n - 1\). Let again \(X_1\) be a uniform random variable on \(\{-1, 1\}\) and, for \(i = 2, \ldots, n\), define the random variable \(X_i = U_{i-1}X_{i-1}\), that is, \(X_i\) is the same as \(X_{i-1}\) with probability \(\varepsilon\) and otherwise is flipped. Then,

$$\mathbb{E}[f(X_1, \ldots, X_n)] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E} \left[ X_1 \left( 1 + \sum_{i=1}^{n-1} U_j \right) \right] = \mathbb{E}[X_1] \mathbb{E} \left[ 1 + \sum_{i=1}^{n-1} U_j \right] = 0,$$

by the independence of \(X_1\) and the \(U_i\)'s. Similarly

$$\mathbb{E}[f(X_1, \ldots, X_n) | X_1] = X_1 \mathbb{E} \left[ 1 + \sum_{i=1}^{n-1} U_j \right] = X_1 \left( \sum_{i=1}^{n} (1 - 2\varepsilon)^i \right),$$

so that

$$|\mathbb{E}[f(X_1, \ldots, X_n) | X_1] - \mathbb{E}[f(X_1, \ldots, X_n)]| = \left( \sum_{i=1}^{n} (1 - 2\varepsilon)^i \right) > 2,$$

for \(\varepsilon\) small enough and \(n \geq 3\). In particular, the corresponding Doob martingale does not have increments bounded by 2.

Bounds on \(\|D_i f\|_\infty\) are often expressed in terms of a Lipschitz condition under an appropriate metric. The Hamming distance is defined as

$$\rho(x, x') := \sum_{i=1}^{n} 1_{\{x_i \neq x'_i\}},$$

for \(\varepsilon\) small enough and \(n \geq 3\). In particular, the corresponding Doob martingale does not have increments bounded by 2.

Bounds on \(\|D_i f\|_\infty\) are often expressed in terms of a Lipschitz condition under an appropriate metric. The Hamming distance is defined as

$$\rho(x, x') := \sum_{i=1}^{n} 1_{\{x_i \neq x'_i\}},$$
for \( x, x' \in X_1 \times \cdots \times X_n \). Let \( 0 < c < +\infty \). A function \( f : X_1 \times \cdots \times X_n \to \mathbb{R} \) is \( c \)-Lipschitz (with respect to the Hamming distance) if for all \( i = 1, \ldots, n \), all \( (x_1, \ldots, x_n) \in X_1 \times \cdots \times X_n \) and all \( y, y' \in X_i \)

\[
|f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, y', x_{i+1}, \ldots, x_n)| \leq c.
\]

**Lemma 3.54.** If \( f \) is \( c \)-Lipschitz, then

\[
\|D_i f\|_\infty \leq c, \quad \forall i.
\]

**Variance bounds.** We give improved bounds on the variance. Our first bound decomposes the variance of \( f(X) \) over the contributions of its individual entries.

**Theorem 3.55 (Tensorization of the variance).** Let \( X_1, \ldots, X_n \) be independent random variables where \( X_i \) is \( X_i \)-valued for all \( i \) and let \( X = (X_1, \ldots, X_n) \). Assume that \( f : X_1 \times \cdots \times X_n \to \mathbb{R} \) is a measurable function with \( \mathbb{E}[f(X)^2] < +\infty \). Define \( \mathcal{F}_i = \sigma(X_1, \ldots, X_i) \), \( \mathcal{G}_i = \sigma(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \) and \( Z_i = \mathbb{E}[f(X) | \mathcal{F}_i] \). Then we have

\[
\text{Var}[f(X)] \leq \sum_{i=1}^n \mathbb{E}[\text{Var}[f(X) | \mathcal{G}_i]].
\]

(Recall the formula: \( \text{Var}[Y] = \mathbb{E}[\text{Var}[Y | \mathcal{H}]] + \text{Var}[\mathbb{E}[Y | \mathcal{H}]] \).)

**Proof of Theorem 3.55.** The key lemma is the following.

**Lemma 3.56.**

\[
\mathbb{E}[\mathbb{E}[f(X) | \mathcal{G}_i] | \mathcal{F}_i] = \mathbb{E}[f(X) | \mathcal{F}_{i-1}]
\]

**Proof.** By the tower property (Lemma A.5),

\[
\mathbb{E}[f(X) | \mathcal{F}_{i-1}] = \mathbb{E}[\mathbb{E}[f(X) | \mathcal{G}_i] | \mathcal{F}_{i-1}].
\]

Moreover, \( \sigma(X_i) \) is independent of \( \sigma(\mathcal{G}_i, \mathcal{F}_{i-1}) \) so by the role of independence (Lemma A.4), we have

\[
\mathbb{E}[\mathbb{E}[f(X) | \mathcal{G}_i] | \mathcal{F}_{i-1}] = \mathbb{E}[\mathbb{E}[f(X) | \mathcal{G}_i] | \mathcal{F}_{i-1}, X_i] = \mathbb{E}[\mathbb{E}[f(X) | \mathcal{G}_i] | \mathcal{F}_i].
\]

Combining the last two displays gives the result.
Again, we take advantage of the orthogonality of increments (Lemma 3.40) to write

\[ \text{Var}[f(X)] = \sum_{i=1}^{n} \mathbb{E} [ (Z_i - Z_{i-1})^2 ] . \]

By the lemma above,

\[ (Z_i - Z_{i-1})^2 = (\mathbb{E} [f(X) | \mathcal{F}_i] - \mathbb{E} [f(X) | \mathcal{F}_{i-1}])^2 \]

\[ = (\mathbb{E} [f(X) | \mathcal{F}_i] - \mathbb{E} [\mathbb{E} [f(X) | \mathcal{G}_i] | \mathcal{F}_i])^2 \]

\[ \leq \mathbb{E} \left[ (f(X) - \mathbb{E} [f(X) | \mathcal{G}_i])^2 \right] , \]

where we used Jensen’s inequality on the last line. Taking expectations

\[ \text{Var}[f(X)] = \sum_{i=1}^{n} \mathbb{E} [ (Z_i - Z_{i-1})^2 ] \]

\[ \leq \sum_{i=1}^{n} \mathbb{E} \left[ (f(X) - \mathbb{E} [f(X) | \mathcal{G}_i])^2 \right] \]

\[ = \sum_{i=1}^{n} \mathbb{E} \left[ (f(X) - \mathbb{E} [f(X) | \mathcal{G}_i])^2 \right] \]

\[ = \sum_{i=1}^{n} \mathbb{E} \left[ (f(X) - \mathbb{E} [f(X) | \mathcal{G}_i])^2 \right] \]

\[ = \sum_{i=1}^{n} \mathbb{E} [\text{Var}[f(X) | \mathcal{G}_i]] . \]

That concludes the proof.

We derive two useful consequences of the tensorization property of the variance. The first one is the *Efron-Stein inequality.*

**Theorem 3.57** (Efron-Stein inequality). Let \( X_1, \ldots , X_n \) be independent random variables where \( X_i \) is \( \mathcal{X}_i \)-valued for all \( i \) and let \( X = (X_1, \ldots , X_n) \). Assume that \( f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R} \) is a measurable function with \( \mathbb{E}[f(X)^2] < +\infty \). Let \( X' = (X'_1, \ldots , X'_n) \) be an independent copy of \( X \) and

\[ X^{(i)} = (X_1, \ldots , X_{i-1}, X'_i, X_{i+1}, \ldots , X_n) . \]

Then,

\[ \text{Var}[f(X)] \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E} [(f(X) - f(X^{(i)}))^2] . \]
Proof. Observe that if \( Y' \) is an independent copy of \( Y \in \mathcal{L}^2 \), then \( \text{Var}[Y] = \frac{1}{2} \mathbb{E}[(Y - Y')^2] \), which can be seen by adding and subtracting the mean, expanding and using independence. Hence,

\[
\text{Var}[f(X) \mid G_i] = \frac{1}{2} \mathbb{E}[(f(X) - f(X^{(i)}))^2],
\]

where we used the independence of the \( X_i \)s and \( X'_i \)s.

Our second consequence of Theorem 3.55 is a Poincaré-type inequality which relates the variance of a function to its expected “square gradient.”

**Theorem 3.58** (Bounded differences inequality). Let \( X_1, \ldots, X_n \) be independent random variables where \( X_i \) is \( \mathcal{X}_i \)-valued for all \( i \) and let \( X = (X_1, \ldots, X_n) \). Assume that \( f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R} \) is a measurable function with \( \mathbb{E}[f(X)^2] < +\infty \). Then

\[
\text{Var}[f(X)] \leq \frac{1}{4} \sum_{i=1}^n \mathbb{E}[D_i f(X)^2].
\]

**Proof.** By Lemma 2.41 (which we previously used to prove Hoeffding’s lemma),

\[
\text{Var}[f(X) \mid G_i] \leq \frac{1}{4} D_i f(X)^2.
\]

**Remark 3.59.** For comparison, a version of the classical Poincaré inequality in one dimension asserts the following: let \( f : [0, 1] \to \mathbb{R} \) be continuously differentiable with \( \int_0^1 f(x)^2 + f'(x)^2 \, dx < +\infty \) and \( \int_0^1 f(x) \, dx = 0 \), then

\[
\int_0^1 f(x)^2 \, dx \leq \int_0^1 f'(x)^2 \, dx.
\]

Indeed, \( f(x) - f(0) = \int_0^x f'(x) \, dx \) so that, by Cauchy-Schwarz, \( (f(x) - f(0))^2 \leq \int_0^x f'(x)^2 \, dx \) and the result follows by integration after noting that \( \int_0^1 (f(x) - a)^2 \, dx \) is minimized at \( a = \int_0^1 f(x) \, dx = 0 \). Intuitively, for a function with 0 mean to have a large norm, it must have a large absolute derivate somewhere.

**Example 3.60** (Longest common subsequence). Let \( X_1, \ldots, X_{2n} \) be independent uniform random variables in \( \{-1, +1\} \). Let \( Z \) be the length of the longest common subsequence in \( (X_1, \ldots, X_n) \) and \( (X_{n+1}, \ldots, X_{2n}) \), that is,

\[
Z = \max \{ k : \exists 1 \leq i_1 < i_2 < \cdots < i_k \leq n \text{ and } n + 1 \leq j_1 < j_2 < \cdots < j_k \leq 2n \text{ such that } X_{i_1} = X_{j_1}, X_{i_2} = X_{j_2}, \ldots, X_{i_k} = X_{j_k} \}.
\]
Then, writing \( Z = f(X_1, \ldots, X_{2n}) \), it follows that \( \| D_i f \|_\infty \leq 1 \). Indeed, fix \( x = (x_1, \ldots, x_{2n}) \) and let \( x^{i,+} \) (respectively \( x^{i,-} \)) be \( x \) where the \( i \)-th component is replaced with \( +1 \) (respectively \( -1 \)). Assume w.l.o.g. that \( f(x^{i,-}) \leq f(x^{i,+}) \). Then \( | f(x^{i,+}) - f(x^{i,-}) | \leq 1 \) because removing the \( i \)-th component (and its match) from a longest common subsequence when \( x_i = +1 \) (if present) decreases the length by 1. Since this is true for any \( x \), we have \( \| D_i f \|_\infty \leq 1 \). Finally, by Theorem 3.58,

\[
\text{Var}[Z] \leq \frac{1}{4} \sum_{i=1}^{2n} \| D_i f \|_\infty^2 \leq \frac{n}{2}.
\]

**McDiarmid’s inequality**  The following powerful consequence of the Azuma-Hoeffding inequality (Theorem 3.52) is commonly referred to as the *method of bounded differences*.

**Theorem 3.61** (McDiarmid’s inequality). Let \( X_1, \ldots, X_n \) be independent random variables where \( X_i \) is \( \mathcal{X}_i \)-valued for all \( i \), and let \( X = (X_1, \ldots, X_n) \). Assume \( f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R} \) is a measurable function such that \( \| D_i f \|_\infty < +\infty \) for all \( i \). Then for all \( \beta > 0 \)

\[
\mathbb{P}[f(X) - \mathbb{E}f(X) \geq \beta] \leq \exp \left( -\frac{2 \beta^2}{\sum_{i=1}^{n} \| D_i f \|_\infty^2} \right).
\]

Once again, applying the inequality to \(-f\) gives a tail bound in the other direction.

**Proof of Theorem 3.61.** As before, we let

\[
Z_i = \mathbb{E}[f(X) \mid \mathcal{F}_i],
\]

where \( \mathcal{F}_i = \sigma(X_1, \ldots, X_i) \). We also define \( \mathcal{G}_i = \sigma(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \). Then, it holds that \( A_i \leq Z_i - Z_{i-1} \leq B_i \) where

\[
B_i = \mathbb{E} \left[ \sup_{y \in \mathcal{X}_i} f(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) - f(X) \mid \mathcal{F}_{i-1} \right],
\]

and

\[
A_i = \mathbb{E} \left[ \inf_{y \in \mathcal{X}_i} f(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) - f(X) \mid \mathcal{F}_{i-1} \right].
\]
Indeed, since $\sigma(X_i)$ is independent of $\mathcal{F}_{i-1}$ and $\mathcal{G}_i$, by the role of independence (Lemma A.4)

\[
Z_i = \mathbb{E} \left[ f(X) \mid \mathcal{F}_i \right] 
\leq \mathbb{E} \left[ \sup_{y \in \mathcal{X}_i} f(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) \mid \mathcal{F}_i \right]
= \mathbb{E} \left[ \sup_{y \in \mathcal{X}_i} f(X_1, \ldots, X_{i-1}, y, X_{i+1}, X_n) \mid \mathcal{F}_{i-1}, X_1 \right]
= \mathbb{E} \left[ \sup_{y \in \mathcal{X}_i} f(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) \mid \mathcal{F}_{i-1} \right],
\]

and similarly for the other direction. Moreover, by definition, $B_i - A_i \leq \|D_i f\|_\infty := c_i$. The Azuma-Hoeffding inequality (Theorem 3.52) then gives the result.

**Examples** The moral of McDiarmid’s inequality is that functions of independent variables that are smooth, in the sense that they do not depend too much on any one of their variables, are concentrated around their mean. Here are some straightforward applications.

**Example 3.62** (Balls and bins: empty bins). Suppose we throw $m$ balls into $n$ bins independently, uniformly at random. The number of empty bins, $Z_{n,m}$, is centered at

\[
\mathbb{E} Z_{n,m} = n \left( 1 - \frac{1}{n} \right)^m.
\]

Writing $Z_{n,m}$ as the sum of indicators $\sum_{i=1}^n 1_{B_i}$, where $B_i$ is the event that bin $i$ is empty, is a natural first attempt at proving concentration around the mean. However there is a problem—the $B_i$s are not independent. Indeed, because there is a fixed number of bins, the event $B_i$ intuitively makes the other such events less likely. Instead let $X_j$ be the index of the bin in which ball $j$ lands. The $X_j$s are independent by construction and, moreover, $Z_{n,m} = f(X_1, \ldots, X_m)$ where $f$ is 1-Lipschitz. Indeed, moving a single ball changes the number of empty bins by at most 1 (if at all). Hence by the method of bounded differences

\[
\mathbb{P} \left[ \left| Z_{n,m} - n \left( 1 - \frac{1}{n} \right)^m \right| \geq b \sqrt{m} \right] \leq 2e^{-2b^2}.
\]

**Example 3.63** (Pattern matching). Let $X = (X_1, X_2, \ldots, X_n)$ be i.i.d. random variables taking values uniformly at random in a finite set $S$ of size $s = |S|$. Let
Let $a = (a_1, \ldots, a_k)$ be a fixed substring of elements of $S$. We are interested in the number of occurrences of $a$ as a (consecutive) substring in $X$, which we denote by $N_n$. Denote by $E_i$ the event that the substring of $X$ starting at $i$ is $a$. Summing over the starting positions and using the linearity of expectation, the mean of $N_n$ is

$$E[N_n] = E\left[ \sum_{i=1}^{n-k+1} 1_{E_i} \right] = (n - k + 1) \left( \frac{1}{s} \right)^k.$$

However the $1_{E_i}$s are not independent. So we cannot use a Chernoff bound for Poisson trials. Instead we use the fact that $N_n = f(X)$ where $f$ is $k$-Lipschitz, as each $X_i$ appears in at most $k$ substrings of length $k$. By the method of bounded differences, for all $b > 0$,

$$P[|N_n - E[N_n]| \geq bk\sqrt{n}] \leq 2e^{-2b^2}.$$

The last two examples are perhaps not surprising in that they involve “sums of weakly independent” indicator variables. One might reasonably expect a sub-Gaussian-type inequality in that case. The application in the next section is more striking.

One more example:

**Example 3.64** (Concentration of measure on the hypercube). For $A \subseteq \{0, 1\}^n$ a subset of the hypercube and $r > 0$, we let

$$A_r = \left\{ x \in \{0, 1\}^n : \inf_{a \in A} \|x - a\|_1 \leq r \right\},$$

be the points at $\ell_1$ distance $r$ from $A$. Fix $\varepsilon \in (0, 1/2)$ and assume that $|A| \geq \varepsilon 2^n$. Let $\lambda_\varepsilon$ be such that $e^{-2\lambda_\varepsilon^2} = \varepsilon$. The following application of the method of bounded differences indicates that much of the uniform measure on the high-dimensional hypercube lies in a close neighborhood of any such “small” set $A$. This is an example of the concentration of measure phenomenon.

**Claim 3.65.**

$$r > 2\lambda_\varepsilon \sqrt{n} \implies |A_r| \geq (1 - \varepsilon) 2^n.$$

**Proof.** Let $X = (X_1, \ldots, X_n)$ be uniformly distributed in $\{0, 1\}^n$. Note that the coordinates are in fact independent. The function $f(x) = \inf_{a \in A} \|x - a\|_1$ is 1-Lipschitz. Indeed changing one coordinate of $x$ can only increase the $\ell_1$ distance to the closest point to $x$ by 1. Hence McDiarmid’s inequality (Theorem 3.61) gives

$$P[|E[f(X)] - f(X)| \geq \beta] \leq \exp \left( \frac{-2\beta^2}{n} \right).$$

127
Choosing $\beta = \mathbb{E} f(X)$ and noting that $f(x) \leq 0$ if and only if $x \in A$ gives

$$\mathbb{P}[A] \leq \exp \left( -\frac{2(\mathbb{E} f(X))^2}{n} \right),$$

or, rearranging and using our assumption on $A$,

$$\mathbb{E} f(X) \leq \sqrt{\frac{1}{2} n \log \frac{1}{\mathbb{P}[A]}} \leq \sqrt{\frac{1}{2} n \log \frac{1}{\varepsilon}} = \lambda \sqrt{n}.$$

By a second application of the method of bounded differences with $\beta = \lambda \sqrt{n}$,

$$\mathbb{P} \left[ f(X) \geq 2\lambda \sqrt{n} \right] \leq \mathbb{P} \left[ f(X) - \mathbb{E} f(X) \geq b \right] \leq \exp \left( -\frac{2\beta^2}{n} \right) = \varepsilon.$$

The result follows by observing that, with $r > 2\lambda \sqrt{n}$,

$$\frac{|A_r|}{2^n} \geq \mathbb{P} \left[ f(X) < 2\lambda \sqrt{n} \right] \geq 1 - \varepsilon.$$

Claim 3.65 is striking for two reasons: 1) the radius $2\lambda \sqrt{n}$ is much smaller than $n$, the diameter of $\{0, 1\}^n$; and 2) it applies to any $A$. The smallest $r$ such that $|A_r| \geq (1 - \varepsilon)2^n$ in general depends on $A$. Here are two extremes.

For $\gamma > 0$, let

$$B(\gamma) := \left\{ x \in \{0, 1\}^n : \|x\|_1 \leq \frac{n}{2} - \frac{n}{4} \right\}.$$

Note that, letting for $Y_n \sim B(n, \frac{1}{2})$,

$$\frac{1}{2^n} |B(\gamma)| = \sum_{\ell=0}^{\frac{n}{2} \gamma \sqrt{\ell}} \binom{n}{\ell} 2^{-n} = \mathbb{P} \left[ Y_n \leq \frac{n}{2} - \frac{n}{4} \right]. \quad (3.8)$$

By the Berry-Esséen theorem (e.g., [Dur10, Theorem 3.4.9]), there is a $C > 0$ such that, after rearranging the final quantity in (3.8),

$$\left| \mathbb{P} \left[ \frac{Y_n - n/2}{\sqrt{n}/4} \leq -\gamma \right] - \mathbb{P}[Z \leq -\gamma] \right| \leq \frac{C}{\sqrt{n}},$$

where $Z \sim N(0, 1)$. Let $\varepsilon < \varepsilon' < 1/2$ and let $\gamma_{\varepsilon'}$ be such that $\mathbb{P}[Z \leq -\gamma_{\varepsilon'}] = \varepsilon'$. Then setting $A := B(\gamma_{\varepsilon'})$, for $n$ large enough, we have $|A| \geq \varepsilon' 2^n$ by (3.8). On
the other hand, setting \( r := \gamma \varepsilon \sqrt{n}/4 \), we have \( A_r \subseteq B(0) \), so that \( |A_r| \leq \frac{1}{2} 2^n < (1 - \varepsilon) 2^n \). We have shown that \( r = \Omega(\sqrt{n}) \) is in general required for Claim 3.65 to hold.

For an example at the other extreme, assume for simplicity that \( N := \varepsilon 2^n \) is an integer. Let \( A \subseteq \{0, 1\}^n \) be constructed as follows: starting from the empty set, add points in \( \{0, 1\}^n \) to \( A \) independently, uniformly at random until \( |A| = N \). Set \( r := 2 \). By the first moment method (Corollary 2.11), the probability that \( A_r \) does not cover all of \( \{0, 1\}^n \) is at most

\[
\mathbb{P}[|\{0, 1\}^n \setminus A_r| > 0] \leq \sum_{x \in \{0, 1\}^n} \mathbb{P}[x \notin A_r] \leq 2^n \left( 1 - \frac{n}{2^n} \right)^{\varepsilon 2^n} \leq 2^n e^{-\varepsilon(\frac{n}{2^n})},
\]

where, in the second inequality, we considered only the first \( N \) picks in the construction of \( A \). In particular, as \( n \to +\infty \), \( \mathbb{P}[|\{0, 1\}^n \setminus A_r| > 0] < 1 \). So for \( n \) large enough there is a set \( A \) such that \( A_r = \{0, 1\}^n \) where \( r = 2 \).

**Remark 3.66.** In fact, it can be shown that sets of the form \( \{x : \|x\|_1 \leq s\} \) have the smallest “expansion” among subsets of \( \{0, 1\}^n \) of the same size, a result known as Harper’s vertex isoperimetric theorem. See, e.g., [BLM13, Theorem 7.6 and Exercises 7.11-7.13].

### 3.2.3 Erdős-Rényi: exposure martingales and application to the chromatic number

**Exposure martingales** In the context of Erdős-Rényi graphs, a common way to apply the Azuma-Hoeffding inequality (Theorem 3.52) is to introduce a so-called exposure martingale. Let \( G \sim G_{n,p} \) and let \( F \) be any function on graphs such that \( \mathbb{E}_{n,p}[F(G)] < +\infty \) for all \( n, p \). Choose an arbitrary ordering of the vertices and, for \( i = 1, \ldots, n \), denote by \( H_i \) the subgraph of \( G \) induced by the first \( i \) vertices. Then the filtration \( \mathcal{H}_i = \sigma(H_1, \ldots, H_i), i = 1, \ldots, n \), corresponds to exposing the vertices of \( G \) one at a time. The Doob martingale

\[
Z_i = \mathbb{E}_{n,p}[F(G) | \mathcal{H}_i], \quad i = 1, \ldots, n,
\]

is known as a *vertex exposure martingale*. An alternative way to define the filtration is to consider instead the random variables \( X_i = (1_{\{i,j\} \in E} : 1 \leq j \leq i) \) for \( i = 2, \ldots, n \). In words, \( X_i \) is a vector whose entries indicate the status (present or absent) of all potential edges incident to \( i \) and a vertex preceding it. Hence, \( \mathcal{H}_i = \sigma(X_2, \ldots, X_i) \) for \( i = 2, \ldots, n \) (and note that \( \mathcal{H}_1 \) is trivial as it corresponds to a graph with a single vertex and no edge). This representation has an important property: the \( X_i \)'s are *independent* as they pertain to disjoint subsets of edges. We are then in the setting of the method of bounded differences. Re-writing \( F(G) = \).
the vertex exposure martingale coincides with the martingale (3.6) used in that context.

As an example, consider the chromatic number $\chi(G)$, i.e., the smallest number of colors needed in a proper coloring of $G$. Define $f_\chi(X_1, \ldots, X_n) := \chi(G)$. We use the following combinatorial observation to bound $\|D_if_\chi\|_\infty$.

**Lemma 3.67.** Altering the status (absent or present) of edges incident to a fixed vertex $v$ changes the chromatic number by at most 1.

**Proof.** Altering the status of edges incident to $v$ increases the chromatic number by at most 1, since in the worst case one can simply use an extra color for $v$. On the other hand, if the chromatic number were to decrease by more than 1 after altering the status of edges incident to $v$, reversing the change and using the previous observation would produce a contradiction. □

A fortiori, since $X_i$ depends on a subset of the edges incident to node $i$, Lemma 3.67 implies that $f_\chi$ is 1-Lipschitz. Hence, for all $0 < p < 1$ and $n$, by an immediate application of the McDiarmid’s inequality (Theorem 3.61):

**Claim 3.68.**

$$\Pr_{n,p} \left[ |\chi(G) - \mathbb{E}_{n,p}[\chi(G)]| \geq b\sqrt{n-1} \right] \leq 2e^{-2b^2}.$$  

**Edge exposure** can be defined in a manner similar to vertex exposure: reveal the edges one at a time in an arbitrary order. By Lemma 3.67, the corresponding function is 1-Lipschitz. Observe however that, for the chromatic number, edge exposure results in a much weaker bound as the $\Theta(n^2)$ random variables produce only a linear in $n$ deviation for the same tail probability. (The reader may want to ponder the apparent paradox: using a larger number of independent variables seemingly leads to weaker concentration in this case.)

**Remark 3.69.** Note that Claim 3.68 tells us nothing about the expectation of $\chi(G)$. It turns out that, up to logarithmic factors, $\mathbb{E}_{n,p}[\chi(G)]$ is of order $np_n$ when $p_n \sim n^{-\alpha}$ for some $0 < \alpha < 1$. We will not prove this result here. See the “Bibliographic remarks” at the end of this chapter for more on the chromatic number of Erdős-Rényi graphs.

$\chi(G)$ is **concentrated on few values**. Much stronger concentration results can be obtained: when $p_n = n^{-\alpha}$ with $0 < \alpha < 1$, $\chi(G)$ is in fact concentrated on two values! We give a partial result along those lines which illustrates a less straightforward choice of martingale in the Azuma-Hoeffding inequality (Theorem 3.52).
Claim 3.70. Let $p_n = n^{-\alpha}$ with $\alpha > \frac{5}{6}$ and let $G_n \sim G_{n,p_n}$. Then for any $\varepsilon > 0$ there is $\varphi_n := \varphi_n(\alpha, \varepsilon)$ such that

$$\mathbb{P}_{n,p_n} \left[ \varphi_n \leq \chi(G_n) \leq \varphi_n + 3 \right] \geq 1 - \varepsilon,$$

for all $n$ large enough.

Proof. We consider the following martingale. Let $\varphi_n$ be the smallest integer such that

$$\mathbb{P}_{n,p_n} \left[ \chi(G_n) \leq \varphi_n \right] > \frac{\varepsilon}{3}. \quad (3.9)$$

Let $F_n(G_n)$ be the minimal size of a set of vertices, $U$, in $G_n$ such that $G_n \setminus U$ is $\varphi_n$-colorable. Let $(Z_i)$ be the corresponding vertex exposure martingale. The proof proceeds in two steps: we show that 1) all but $O(\sqrt{n})$ vertices can be $\varphi_n$-colored and 2) the remaining vertices can be colored using 3 additional colors. See Figure 3.2.3.

We claim that $(Z_i)$ has bounded increments with bound 1.

Lemma 3.71. Changing the edges adjacent to a single vertex can change $F_n$ by at most 1.
Proof. Changing the edges adjacent to \( v \) can increase \( F_n \) by at most 1. Indeed, if \( F_n \) increases, it must be that \( v \notin U \) and we can add \( v \) to \( U \). On the other hand, if \( F_n \) were to decrease by more than 1, reversing the change and using the previous observation would give a contradiction. ■

Choose \( b \) such that \( e^{-b^2/2} = \frac{\varepsilon}{3} \). Then, applying the Azuma-Hoeffding inequality to \((-Z_i)\),

\[
\Pr_{n,p_n} \left[ F_n(G_n) - \mathbb{E}_{n,p_n}[F_n(G_n)] \leq -b \varepsilon \sqrt{n-1} \right] \leq \frac{\varepsilon}{3}
\]

which, since \( \Pr_{n,p_n}[F_n(G_n) = 0] = \Pr_{n,p_n}[\chi(G_n) \leq \varphi_n] > \frac{\varepsilon}{3} \), implies that

\[
\mathbb{E}_{n,p_n}[F_n(G_n)] \leq b \varepsilon \sqrt{n-1}.
\]

Applying the Azuma-Hoeffding inequality to \((Z_i)\) gives

\[
\Pr_{n,p_n} \left[ F_n(G_n) \geq 2b \varepsilon \sqrt{n-1} \right] \leq \Pr_{n,p_n} \left[ F_n(G_n) - \mathbb{E}_{n,p_n}[F_n(G_n)] \geq b \varepsilon \sqrt{n-1} \right] \leq \frac{\varepsilon}{3}.
\]

(3.10)

(3.11)

So with probability at least \( 1 - \frac{\varepsilon}{3} \), we can color all vertices but \( 2b \varepsilon \sqrt{n-1} \) using \( \varphi_n \) colors. Let \( U \) be the remaining uncolored vertices.

We claim that, with high probability, we can color the vertices in \( U \) using at most 3 extra colors.

**Lemma 3.72.** Fix \( c > 0 \), \( \alpha > \frac{5}{6} \) and \( \varepsilon > 0 \). Let \( G_n \sim \mathbb{G}_{n,p_n} \) with \( p_n = n^{-\alpha} \). For all \( n \) large enough,

\[
\Pr_{n,p_n} \left[ \text{every subset of } c\sqrt{n} \text{ vertices of } G_n \text{ can be 3-colored} \right] \geq 1 - \frac{\varepsilon}{3}.
\]

(3.12)

**Proof.** We use the first moment method (Theorem 2.10). To bound the probability that a subset of vertices is not 3-colorable, we consider a minimal such subset and notice that all of its vertices must have degree at least 3. Indeed, suppose \( W \) is not 3-colorable but that all of its subsets are (we call such a subset minimal, non 3-colorable), and suppose that \( w \in W \) has degree less than 3. Then \( W \setminus \{w\} \) is 3-colorable. But, since \( w \) has fewer than 3 neighbors, it can also be properly colored without adding a new color—a contradiction. In particular, the subgraph of \( G_n \) induced by \( W \) must have at least \( \frac{3}{2} |W| \) edges.

Let \( Y_n \) be the number of minimal, non 3-colorable subsets of vertices of \( G_n \) of size at most \( c\sqrt{n} \). By the argument above, the probability that a subset of vertices...
of $G_n$ of size $\ell$ is minimal, non 3-colorable is at most $\binom{n}{\ell} p_n^{3\ell}$ by a union bound over subsets of edges of size $\frac{3\ell}{2}$. Then, by the first moment method,

$$\mathbb{P}_{n,p_n}[Y_n > 0] \leq \mathbb{E}_{n,p_n} Y_n$$

$$\leq \sum_{\ell=4}^{\infty} \binom{n}{\ell} \left( \frac{\ell}{2} \right) p_n^{3\ell}$$

$$\leq \sum_{\ell=4}^{\infty} \left( \frac{e}{\ell} \right)^{\ell} \left( \frac{e\ell}{3} \right)^{3\ell/2} n^{3\alpha/2}$$

$$\leq \sum_{\ell=4}^{\infty} \left( \frac{e\frac{5}{4} n^{1-\frac{3\alpha}{2}} \ell^{\frac{1}{2}}}{3^{\frac{3}{2}}} \right)^{\ell}$$

$$\leq \sum_{\ell=4}^{\infty} \left( c' n^{\frac{5}{4} - \frac{3\alpha}{2}} \right)^{\ell}$$

$$\leq O \left( n^{\frac{5}{4} - \frac{3\alpha}{2}} \right)^{4}$$

$$\rightarrow 0,$$

as $n \to +\infty$, for some $c' > 0$, where we used that $\frac{5}{4} - \frac{3\alpha}{2} < \frac{5}{4} - \frac{5}{4} = 0$ when $\alpha > \frac{5}{6}$.

By the choice of $\varphi_n$ in (3.9),

$$\mathbb{P}_{n,p_n}[\chi(G_n) < \varphi_n] \leq \frac{\varepsilon}{3}.$$

By (3.11) and (3.12),

$$\mathbb{P}_{n,p_n}[\chi(G_n) > \varphi_n + 3] \leq \frac{2\varepsilon}{3}.$$

So, overall,

$$\mathbb{P}_{n,p_n}[\varphi_n \leq \chi(G_n) \leq \varphi_n + 3] \geq 1 - \varepsilon.$$

### 3.2.4 Preferential attachment: degree sequence

Let $(G_t)_{t \geq 1} \sim \text{PA}_m$ be a preferential attachment graph process with parameter $m \geq 1$. A key feature of preferential attachment graphs is a power-law degree sequence: the fraction of vertices with degree $d$ behaves like $\propto d^{-\alpha}$ for some $\alpha > 0$, i.e., it has a fat tail. We prove this in the case of scale-free trees, $m = 1$. 133
Power law degree sequence Let $D_i(t)$ be the degree of the $i$-th vertex, $v_i$, in $G_t$, and denote by

$$N_d(t) := \sum_{i=0}^{t} 1\{D_i(t)=d\},$$

the number of vertices of degree $d$ in $G_t$. Define

$$f_d := \frac{4}{d(d+1)(d+2)}, \quad d \geq 1.$$ (3.13)

Claim 3.73.

$$\frac{1}{t}N_d(t) \to_p f_d, \quad \forall d \geq 1.$$

Proof. Claim 3.73 follows from the following lemmas. Fix $\delta > 0$.

**Lemma 3.74** (Convergence of the mean).

$$\frac{1}{t}\mathbb{E}N_d(t) \to f_d, \quad \forall d \geq 1.$$

**Lemma 3.75** (Concentration around the mean).

$$\mathbb{P}\left[\left|\frac{1}{t}N_d(t) - \frac{1}{t}\mathbb{E}N_d(t)\right| \geq \sqrt{\frac{2\log \delta^{-1}}{t}}\right] \leq 2\delta, \quad \forall d \geq 1, \forall t.$$

An alternative representation of the process We start with the proof of Lemma 3.75, which follows from an application of the method of bounded differences.

**Proof of Lemma 3.75.** In our description of the preferential attachment process, the random choices made at each time depend in a seemingly complicated way on previous choices. In order to establish concentration of the process around its mean, we introduce a clever, alternative construction of the $m=1$ case which has the advantage that it involves independent choices.

We start with a single vertex $v_0$. At time 1, we add a single vertex $v_1$ and an edge $e_1$ connecting $v_0$ and $v_1$. For bookkeeping we orient edges away from the vertex of lower time index. For all $s \geq 2$, let $X_s$ be an independent, uniformly chosen edge extremity among the edges in $G_{s-1}$, i.e., pick a uniform element in

$$\mathcal{X}_s := \{(1, \text{tail}), (1, \text{head}), \ldots, (s-1, \text{tail}), (s-1, \text{head})\}.$$

To form $G_s$, attach a new edge $e_s$ to the vertex of $G_{s-1}$ corresponding to $X_s$. A vertex of degree $d'$ in $G_{s-1}$ is selected with probability $\frac{d'}{2^{s-1}}$, as it should. Note that $X_s$ can be picked in advance independently of the sequence $(G_{s'})_{s' < s}$.
Figure 3.2: Graph obtained when $x_2 = (1, \text{head})$, $x_3 = (2, \text{tail})$ and $x_4 = (3, \text{head})$.

For instance, if $x_2 = (1, \text{head})$, $x_3 = (2, \text{tail})$ and $x_4 = (3, \text{head})$, the graph obtained at time 4 is depicted in Figure 3.2.

We claim that $N_d(t) = h(X_2, \ldots, X_t)$ seen as a function of $X_2, \ldots, X_t$ is 2-Lipschitz. Indeed let $(x_2, \ldots, x_t)$ be a realization of $(X_2, \ldots, X_t)$ and let $y \in X_s$ with $y \neq x_s$. Replacing $x_s = (i, \text{end})$ with $y = (j, \text{end}')$ where $i, j \in \{1, \ldots, s - 1\}$ and $\text{end}, \text{end}' \in \{\text{tail, head}\}$ has the effect of redirecting the head of edge $e_s$ from the end of $e_i$ to the end' of $e_j$. This redirection also brings along with it the heads of all other edges associated with the choice $(s, \text{head})$. But, crucially, those changes only affect the degrees of the vertices corresponding to $(i, \text{end})$ and $(j, \text{end}')$ in the original graph. Hence the number of vertices with degree $d$ changes by at most 2.

For instance, returning to the example of Figure 3.2. If we replace $x_3 = (2, \text{tail})$ with $y = (1, \text{tail})$, one obtains the graph in Figure 3.3. Note that only the degrees of vertices $v_1$ and $v_2$ are affected by this change.

By the method of bounded differences, for all $\beta > 0$,

$$\mathbb{P}[|N_d(t) - \mathbb{E}N_d(t)| \geq \beta] \leq 2 \exp \left(-\frac{2\beta^2}{(2)^2(t-1)}\right),$$
Substituting $x_3 = (2, \text{tail})$ with $y = (1, \text{tail})$ in the example of Figure 3.2 has the effect of replacing the red edges with the green edges. Note that only the degrees of vertices $v_1$ and $v_2$ are affected by this change.

which, choosing $\beta = \sqrt{2t \log \delta^{-1}}$, we can re-write as

$$
\mathbb{P}
\left[
\left| \frac{1}{t} \Delta N_d(t) - \frac{1}{t} \mathbb{E} N_d(t) \right| \geq \sqrt{\frac{2 \log \delta^{-1}}{t}}
\right] \leq 2\delta.
$$

Dynamics of the mean  Once again the method of bounded differences tells us nothing about the mean, which must be analyzed by other means. The proof of Lemma 3.74 does not rely on the Azuma-Hoeffding inequality but is given for completeness (and may be skipped).

Proof of Lemma 3.74. The idea of the proof is to derive a recursion for $f_d$ by considering the evolution of $\mathbb{E} N_d(t)$ and taking a limit as $t \to +\infty$. By the description of the preferential attachment process, the following recursion holds for $t \geq d$

$$
\mathbb{E} N_d(t + 1) - \mathbb{E} N_d(t) = \underbrace{\frac{d-1}{2t} \mathbb{E} N_{d-1}(t)}_{(a)} - \underbrace{\frac{d}{2t} \mathbb{E} N_d(t)}_{(b)} + \underbrace{1_{\{d=1\}}}_{(c)},
$$

(3.14)
and \( \mathbb{E}N_d(d - 1) = 0 \). Indeed: (a) for \( d \geq 2 \), \( N_d(t) \) increases by 1 if a vertex of degree \( d - 1 \) is picked, an event of probability \( \frac{d - 1}{2t} \cdot N_{d-1}(t) \) because the sum of degrees at time \( t \) is twice the number of edges, i.e., \( 2t \); (b) for \( d \geq 1 \), \( N_d(t) \) decreases by 1 if a vertex of degree \( d \) is picked, an event of probability \( \frac{d}{2t} \cdot N_d(t) \); and (c) the last term comes from the fact that the new vertex always has degree 1.

We re-write (3.14) as

\[
\mathbb{E}N_d(t + 1) = \mathbb{E}N_d(t) + \frac{d - 1}{2t} \mathbb{E}N_{d-1}(t) - \frac{d}{2t} \mathbb{E}N_d(t) + 1_{\{d=1\}}
\]

\[
= \left(1 - \frac{d/2}{t}\right) \mathbb{E}N_d(t) + \left\{ \frac{d - 1}{2} \left[ \frac{1}{t} \mathbb{E}N_{d-1}(t) \right] + 1_{\{d=1\}} \right\}
\]

\[
= \left(1 - \frac{d^2/2}{t}\right) \mathbb{E}N_d(t) + g_d(t),
\]

where \( g_d(t) \) is defined as the expression in curly brackets on the second line. We show by induction on \( d \) that \( \frac{1}{t} \mathbb{E}N_d(t) \to f_d \). Because of the form of the recursion, the following lemma is what we need to proceed.

**Lemma 3.76.** Let \( f \) be a function of \( t \in \mathbb{N} \) satisfying the following recursion

\[
f(t + 1) = \left(1 - \frac{\alpha}{t}\right) f(t) + g(t), \quad \forall t \geq t_0
\]

with \( g(t) \to g \in (-\infty, +\infty) \) as \( t \to +\infty \), and where \( \alpha > 0, t_0 \geq 2\alpha, f(t_0) \geq 0 \) are constants. Then

\[
\frac{1}{t} f(t) \to \frac{g}{1 + \alpha},
\]

as \( t \to +\infty \).

The proof of this lemma is given after the proof of Claim 3.73. We first conclude the proof of Lemma 3.74. First let \( d = 1 \). In that case, \( g_1(t) = g_1 := 1, \alpha := 1/2, \) and \( t_0 := 1 \). By Lemma 3.76,

\[
\frac{1}{t} \mathbb{E}N_1(t) \to \frac{1}{1 + 1/2} = \frac{2}{3} = f_1.
\]

Assuming by induction that \( \frac{1}{t} \mathbb{E}N_{d'}(t) \to f_{d'} \) for all \( d' < d \) we get

\[
g_d(t) \to g_d := \frac{d - 1}{2} f_{d-1},
\]

as \( t \to +\infty \). Using Lemma 3.76 with \( \alpha := d/2 \) and \( t_0 := d \), we obtain

\[
\frac{1}{t} \mathbb{E}N_d(t) \to \frac{1}{1 + d/2} \left[ \frac{d - 1}{2} f_{d-1} \right] = \frac{d - 1}{d + 2} \cdot \frac{4}{(d - 1)d(d + 1)} = f_d,
\]

where we used (3.13). That concludes the proof of Lemma 3.74. \( \blacksquare \)
To prove Claim 3.73, we combine Lemmas 3.74 and 3.75. Fix any \( \delta, \varepsilon > 0 \). Choose \( t' \) large enough that for all \( t \geq t' \)
\[
\max \left\{ \frac{1}{t} E N_d(t) - f_d, \sqrt{\frac{2 \log \delta^{-1}}{t}} \right\} \leq \varepsilon.
\]
Then
\[
P \left[ \left| \frac{1}{t} N_d(t) - f_d \right| \geq 2\varepsilon \right] \leq 2\delta,
\]
for all \( t \geq t' \). That proves convergence in probability.

**Proof of the technical lemma**  It remains to prove Lemma 3.76.

**Proof of Lemma 3.76.** By induction on \( t \), we have
\[
f(t + 1) = \left( 1 - \frac{\alpha}{t} \right) f(t) + g(t)
\]
\[
= \left( 1 - \frac{\alpha}{t} \right) \left[ \left( 1 - \frac{\alpha}{t - 1} \right) f(t - 1) + g(t - 1) \right] + g(t)
\]
\[
= \left( 1 - \frac{\alpha}{t} \right) g(t - 1) + g(t) + \left( 1 - \frac{\alpha}{t} \right) \left( 1 - \frac{\alpha}{t - 1} \right) f(t - 1)
\]
\[
= \ldots
\]
\[
= \sum_{i=1}^{t-t_0} g(t - i) \prod_{j=0}^{i-1} \left( 1 - \frac{\alpha}{t - j} \right) + f(t_0) \prod_{j=0}^{t-t_0} \left( 1 - \frac{\alpha}{t - j} \right),
\]
or
\[
f(t + 1) = \sum_{s=t_0}^{t} g(s) \prod_{r=s+1}^{t} \left( 1 - \frac{\alpha}{r} \right) + f(t_0) \prod_{r=t_0}^{t} \left( 1 - \frac{\alpha}{r} \right). \quad (3.16)
\]
To guess the answer note that, for large \( s \), \( g(s) \) is roughly constant and that the product in the first term behaves like
\[
\exp \left( - \sum_{r=s+1}^{t} \frac{\alpha}{r} \right) \approx \exp (-\alpha(\log t - \log s)) \approx \frac{s^\alpha}{t^\alpha}.
\]
So approximating the sum by an integral we get that \( f(t + 1) \approx \frac{g t}{\alpha + 1} \).

Formally, we use that there is a constant \( \gamma = 0.577 \ldots \) such that (see e.g. [LL10, Lemma 12.1.3])
\[
\sum_{\ell=1}^{m} \frac{1}{\ell} = \log m + \gamma + \Theta(m^{-1}),
\]
138
and that by a Taylor expansion, for $|z| \leq 1/2$,

$$\log (1 - z) = -z + \Theta(z^2).$$

Fix $\eta > 0$ small and take $t$ large enough that $\eta t > 2\alpha$ and $|g(s) - g| < \eta$ for all $s \geq \eta t$. Then, for $s + 1 \geq t_0$,

$$\sum_{r=s+1}^{t} \log \left(1 - \frac{\alpha}{r}\right) = - \sum_{r=s+1}^{t} \left\{ \frac{\alpha}{r} + \Theta(r^{-2}) \right\}$$

$$= -\alpha \log t - \log s + \Theta(s^{-1}),$$

so, taking exponentials,

$$\prod_{r=s+1}^{t} \left(1 - \frac{\alpha}{r}\right) = \frac{s^\alpha}{t^\alpha} \left(1 + \Theta(s^{-1})\right).$$

Hence

$$\frac{1}{t} f(t_0) \prod_{r=t_0}^{t} \left(1 - \frac{\alpha}{r}\right) = \frac{t_0^\alpha}{t^\alpha+1} \left(1 + \Theta(t_0^{-1})\right) \to 0.$$ 

Moreover

$$\frac{1}{t} \sum_{s=\eta t}^{t} g(s) \prod_{r=s+1}^{t} \left(1 - \frac{\alpha}{r}\right) \leq \frac{1}{t} \sum_{s=\eta t}^{t} \left(g + \eta \right) \frac{s^\alpha}{t^\alpha} \left(1 + \Theta(s^{-1})\right)$$

$$\leq O(\eta) + (1 + \Theta(t^{-1})) \frac{g}{t^{\alpha+1}} \sum_{s=\eta t}^{t} s^\alpha$$

$$\leq O(\eta) + (1 + \Theta(t^{-1})) \frac{g}{t^{\alpha+1}} \frac{(t + 1)^{\alpha+1}}{\alpha + 1}$$

$$\to O(\eta) + \frac{g}{\alpha + 1},$$

and, similarly,

$$\frac{1}{t} \sum_{s=t_0}^{\eta t} g(s) \prod_{r=s+1}^{t} \left(1 - \frac{\alpha}{r}\right) \leq \frac{1}{t} \sum_{s=t_0}^{\eta t} \left(g + \eta \right) \frac{s^\alpha}{t^\alpha} \left(1 + \Theta(s^{-1})\right)$$

$$\leq \frac{\eta t}{t} (g + \delta) \frac{(\eta t)^\alpha}{t^\alpha} \left(1 + \Theta(t_0^{-1})\right)$$

$$\to O(\eta^{\alpha+1}).$$
Plugging these inequalities back into (3.16), we get

$$\limsup_t \frac{1}{t} f(t + 1) \leq \frac{g}{1 + \alpha} + O(\eta).$$

A similar inequality holds in the other direction. Taking $\eta \to 0$ concludes the proof.

**Remark 3.77.** A more quantitative result (uniform in $t$ and $d$) can be derived. See, e.g., [vdH14, Sections 8.5, 8.6]. See the same reference for the case $m > 1$.

### 3.2.5 Data science: stochastic bandits and the slicing method

In this section, we consider an application of the maximal Azuma-Hoeffding inequality (Theorem 3.52) to bandit problems. Quoting [BCB12]:

A multi-armed bandit problem (or, simply, a bandit problem) is a sequential allocation problem defined by a set of actions. At each time step, a unit resource is allocated to an action and some observable payoff is obtained. The goal is to maximize the total payoff obtained in a sequence of allocations. [...] Bandit problems are basic instances of sequential decision making with limited information and naturally address the fundamental tradeoff between exploration and exploitation in sequential experiments. Indeed, the player must balance the exploitation of actions that did well in the past and the exploration of actions that might give higher payoffs in the future. Although the original motivation of Thompson [Tho33] for studying bandit problems came from clinical trials (when different treatments are available for a certain disease and one must decide which treatment to use on the next patient), modern technologies have created many opportunities for new applications, and bandit problems now play an important role in several industrial domains. [...] Ad placement is the problem of deciding which advertisement to display on the web page delivered to the next visitor of a website. Similarly, website optimization deals with the problem of sequentially choosing design elements (font, images, layout) for the web page. Here the payoff is associated with visitor’s actions, e.g., clickthroughs or other desired behaviors.

In the simplest version of the stochastic bandit problem, there are two unknown distributions $\nu_1, \nu_2$ over $[0, 1]$ with respective means $\mu_1 \neq \mu_2$. At each time $t = 1, \ldots, n$, we request an independent sample from $\nu_{I_t}$, where we are free to choose $I_t \in \{1, 2\}$ based on past choices and observed rewards $\{(I_s, Z_s)\}_{s<t}$. This will
be referred to as pulling arm $I_t$. We then observe the reward $Z_t \sim \nu_{I_t}$. Letting $\mu^* := \mu_1 \lor \mu_2$, our goal is to minimize

$$R_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^n \mu_{I_t}\right],$$

(3.17)

which is known as the pseudo-regret. That is, we seek to make choices $(I_t)_{t=1}^n$ that minimize the difference between the best achievable cumulative mean reward and the expected cumulative mean reward from our decisions. Note that the expectation in (3.17) is taken over the choices $(I_t)_{t=1}^n$, which themselves depend on the random rewards $(Z_s)_{s=1}^n$. As indicated above, because $\nu_1$ and $\nu_2$ are unknown, there is a fundamental friction between exploiting the arm that has done best in the past and exploring further the other arm, which might perform better in the future.

One general approach that has proved effective in this type of problem is known as optimism in the face of uncertainty. Quoting again [BCB12]:

Assume that the forecaster has accumulated some data on the environment and must decide how to act next. First, a set of “plausible” environments which are “consistent” with the data (typically, through concentration inequalities) is constructed. Then, the most “favorable” environment is identified in this set. Based on that, the heuristic prescribes that the decision which is optimal in this most favorable and plausible environment should be made. [...] this principle gives simple and yet almost optimal algorithms for the stochastic multi-armed bandit problem.

A concrete implementation of this principle is the Upper Confidence Bound (UCB) algorithm.

We will need some notation. For $i = 1, 2$, let $T_i(t)$ be the number of times arm $i$ was pulled up to time $t$

$$T_i(t) = \sum_{s \leq t} 1\{I_s = i\},$$

and let $X_{i,s}$, $s = 1, \ldots, n$, be i.i.d. samples from $\nu_i$. Assume that the reward at time $t$ is

$$Z_t =\begin{cases} X_{1,T_1(t-1)+1}, & \text{if } I_t = 1 \\ X_{2,T_2(t-1)+1}, & \text{o.w.} \end{cases}$$

In other words, $X_{i,s}$ is the $s$-th observed reward from arm $i$. Let $\hat{\mu}_{i,s}$ be the sample average of rewards after pulling $s$ times on arm $i$

$$\hat{\mu}_{i,s} = \frac{1}{s} \sum_{r \leq s} X_{i,r}.$$
Since the $X_{i,s}$'s are independent and $[0, 1]$-valued by assumption, by Hoeffding’s inequality (Theorem 2.40), for any $\beta > 0$

$$\mathbb{P}[\hat{\mu}_{i,s} - \mu_i \geq \beta] \lor \mathbb{P}[\mu_i - \hat{\mu}_{i,s} \geq \beta] \leq \exp\left(-2s\beta^2\right).$$

The right-hand side can be made $\leq \delta$ provided

$$\beta \geq \sqrt{\frac{\log \delta^{-1}}{2s}} := H(s, \delta).$$

We are now ready to state the $\alpha$-UCB algorithm, where $\alpha > 1$ is the exploration parameter. At each time $t$, we pick

$$I_t \in \arg\max_{i=1,2} \left\{ \hat{\mu}_{i,T_i(t-1)} + \alpha H(T_i(t-1), 1/t) \right\}.$$

The following theorem shows that UCB achieves a pseudo-regret of the order of $O(\log n)$. Define $\Delta_i = \mu^* - \mu_i$ and $\Delta_* = \Delta_1 \lor \Delta_2$.

**Theorem 3.78 (Pseudo-regret of UCB).** In the two-arm stochastic bandit problem where the rewards are in $[0, 1]$ with distinct means, $\alpha$-UCB with $\alpha > 1$ achieves

$$\bar{R}_n \leq \frac{2\alpha^2}{\Delta_*} \log n + \Delta_* C_\alpha,$$

for some constant $C_\alpha \in (0, +\infty)$ depending only on $\alpha$.

This bound should not come entirely as a surprise. Indeed a simple, alternative approach to UCB is to (1) first pull each arm $m_n = o(n)$ times and then (2) use the arm with largest estimated mean for the remainder. Assuming there is a known lower bound on $\Delta_*$, then Hoeffding’s inequality (Theorem 2.40) guarantees that $m_n$ can be chosen of the order of $\frac{1}{\Delta_*^2} \log n$ to identify the largest mean with probability $1 - 1/n$. Because the rewards are bounded by 1, accounting for the contribution of the first phase and the probability of failure in the second phase, one gets a pseudo-regret of the order of $\Delta^* \frac{1}{\Delta_*^2} \log n + \Delta_* \approx \frac{1}{\Delta_*} \log n$. The UCB strategy, on the other hand, elegantly adapts to the gap $\Delta_*$ and the horizon $n$.

We break down the proof into a sequence of lemmas. We first rewrite the
pseudo-regret as

\[
\overline{R}_n = n\mu^* - \mathbb{E} \left[ \sum_{t=1}^n \mu_{i_t} \right] \\
= \mathbb{E} \left[ \sum_{t=1}^n (\mu^* - \mu_{i_t}) \right] \\
= \mathbb{E} \left[ \sum_{t=1}^n \sum_{i=1,2} 1\{I_t = i\} \Delta_i \right] \\
= \sum_{i=1,2} \Delta_i \mathbb{E}[T_i(n)].
\]  

(3.18)

Hence the problem boils down to bounding the expected number of times, \( \mathbb{E}[T_i(n)] \), that arm \( i \) is pulled. We will use the following sufficient condition. Let \( i^* \) be the optimal arm, that is, the one that achieves \( \mu^* \). Intuitively, if arm \( i \neq i^* \) is pulled, it is because our upper estimate of the optimal mean happens to be low, or our lower estimate of the mean of \( i \) happens to be high, or there is too much uncertainty in our estimate of \( \mu_i \).

**Lemma 3.79.** Under the \( \alpha \)-UCB strategy, if arm \( i \neq i^* \) is pulled at time \( t \) then at least one of the following events hold:

\[
\mathcal{E}_{t,1} = \{ \hat{\mu}_{i^*,T_i(t-1)} + \alpha \mathcal{H}(T_i(t-1), 1/t) \leq \mu^* \},
\]

(3.19)

\[
\mathcal{E}_{t,2} = \{ \hat{\mu}_{i,T_i(t-1)} - \alpha \mathcal{H}(T_i(t-1), 1/t) > \mu_i \},
\]

(3.20)

\[
\mathcal{E}_{t,3} = \{ \alpha \mathcal{H}(T_i(t-1), 1/t) > \frac{\Delta_i}{2} \}.
\]

(3.21)

**Proof.** We argue by contradiction. Assume all the conditions above are false. Then

\[
\hat{\mu}_{i^*,T_i(t-1)} + \alpha \mathcal{H}(T_i(t-1), 1/t) > \mu^*
\]

\[
= \mu_i + \Delta_i \\
\geq \hat{\mu}_{i,T_i(t-1)} + \alpha \mathcal{H}(T_i(t-1), 1/t).
\]

That implies that arm \( i \) would not be chosen. \( \blacksquare \)

Using the condition in Lemma 3.79, we get the following bound on \( \mathbb{E}[T_i(n)] \). Let

\[
u_n = \frac{2\alpha^2 \log n}{\Delta_i^2}.
\]
Lemma 3.80. Under the $\alpha$-UCB strategy, for $i \neq i^*$,

$$\mathbb{E}[T_i(n)] \leq u_n + \sum_{t=1}^{n} \mathbb{P}[\mathcal{E}_{t,1}] + \sum_{t=1}^{n} \mathbb{P}[\mathcal{E}_{t,2}].$$

Proof. For $i \neq i^*$,

$$\mathbb{E}[T_i(n)] = \mathbb{E} \left[ \sum_{t=1}^{n} 1_{\{I_t=i\}} \right] = \mathbb{E} \left[ \sum_{t=1}^{n} \left( 1_{\{I_t=i\} \cap \mathcal{E}_{t,1}} + 1_{\{I_t=i\} \cap \mathcal{E}_{t,2}} + 1_{\{I_t=i\} \cap \mathcal{E}_{t,3}} \right) \right],$$

where we used Lemma 3.79. The condition in $\mathcal{E}_{t,3}$ can be written equivalently as

$$\alpha \sqrt{\frac{\log t}{2T_i(t-1)}} > \frac{\Delta_i^2}{2} \iff T_i(t-1) < \frac{2\alpha^2 \log t}{\Delta_i^2}.$$

In particular, for all $t \leq n$, the event $\mathcal{E}_{t,3}$ implies that $T_i(t-1) < u_n$. As a result, since $T_i(t) = T_i(t-1) + 1$ when $I_t = i$, the event $\{I_t = i\} \cap \mathcal{E}_{t,3}$ can occur at most $u_n$ times and

$$\mathbb{E}[T_i(n)] \leq u_n + \mathbb{E} \left[ \sum_{t=1}^{n} \left( 1_{\{I_t=i\} \cap \mathcal{E}_{t,1}} + 1_{\{I_t=i\} \cap \mathcal{E}_{t,2}} \right) \right]$$

$$\leq u_n + \sum_{t=1}^{n} \mathbb{P}[\mathcal{E}_{t,1}] + \sum_{t=1}^{n} \mathbb{P}[\mathcal{E}_{t,2}],$$

which proves the claim.

It remains to bound $\mathbb{P}[\mathcal{E}_{t,1}]$ and $\mathbb{P}[\mathcal{E}_{t,2}]$. The random variable $T_i(t-1)$ depends in a potentially complex way on the past rewards $Z_s$, $s \leq t-1$. So in order to apply a concentration inequality to $\hat{\mu}_{i,T_i(t-1)}$, we use a rather blunt approach: we bound the worst deviation over all possible values in the support of $T_i(t-1)$. That is,

$$\mathbb{P}[\hat{\mu}_{i,T_i(t-1)} + \alpha H(T_i(t-1), 1/t) > \mu_i]$$

$$\leq \mathbb{P} \left[ \bigcup_{s \leq t-1} \{\hat{\mu}_{i,s} + \alpha H(s, 1/t) > \mu_i\} \right].$$

(3.22)
The maximal Azuma-Hoeffding inequality (Theorem 3.52) allows us to bound such a probability without a union bound. We re-write

\[
\mathbb{P}\left[ \bigcup_{s \leq t-1} \left\{ \hat{\mu}_{i,s} + \alpha H(s, 1/t) > \mu_i \right\} \right]
\]

\[
= \mathbb{P}\left[ \sup_{s \leq t-1} (\hat{\mu}_{i,s} - \mu_i) > \alpha H(s, 1/t) \right]
\]

\[
= \mathbb{P}\left[ \sup_{s \leq t-1} \sum_{r=1}^{s} (X_{i,r} - \mu_i) > \alpha s H(s, 1/t) \right]
\]

\[
= \mathbb{P}\left[ \sup_{s \leq t-1} \sum_{r=1}^{s} (X_{i,r} - \mu_i) > \alpha \sqrt{\frac{s \log t}{2}} \right]. \quad (3.23)
\]

Observe that left hand side of the inequality on the last line is the supremum of a martingale with increments bounded by 1. But the right hand side depends on \( s \). We could make the latter \( \sqrt{\frac{\log t}{2}} \) and apply the maximal Azuma-Hoeffding inequality. However a better bound can be derived by using what is known as the slicing method (or peeling method).

The slicing method is useful when bounding a weighted supremum. Here, define \( M_s = \sum_{r=1}^{s} (X_{i,r} - \mu_i) \) and \( \omega(s) = \sqrt{s} \). Our goal is to control probabilities of the form

\[
\mathbb{P}\left[ \sup_{s \leq t-1} \frac{M_s}{\omega(s)} \geq \beta \right].
\]

The idea is to divide the supremum into slices \( \gamma^{k-1} \leq s < \gamma^k, \ k \geq 1 \), where the constant \( \gamma > 1 \) will be optimized below. That is, fixing \( K_t = \left\lfloor \frac{\log t}{\log \gamma} \right\rfloor \) (which roughly solves \( \gamma^{K_t} = t \)),

\[
\mathbb{P}\left[ \sup_{1 \leq s < t} \frac{M_s}{\omega(s)} \geq \beta \right] \leq \sum_{k=1}^{K_t} \mathbb{P}\left[ \sup_{\gamma^{k-1} \leq s < \gamma^k} \frac{M_s}{\omega(s)} \geq \beta \right].
\]

Because \( \omega(s) \) is increasing, on each slice we can bound

\[
\mathbb{P}\left[ \sup_{\gamma^{k-1} \leq s < \gamma^k} \frac{M_s}{\omega(s)} \geq \beta \right] \leq \mathbb{P}\left[ \sup_{\gamma^{k-1} \leq s < \gamma^k} \frac{M_s}{\omega(\gamma^{k-1})} \geq \beta \right]
\]

\[
= \mathbb{P}\left[ \sup_{\gamma^{k-1} \leq s < \gamma^k} M_s \geq \beta \omega(\gamma^{k-1}) \right].
\]
Now we use the maximal Azuma-Hoeffding inequality (Theorem 3.52) to obtain
\[
P \left[ \sup_{s^{k-1} \leq s < \gamma^k} M_s \geq \beta \omega(\gamma^{k-1}) \right] \leq \exp \left( -\frac{2(\beta \omega(\gamma^{k-1}))^2}{\gamma^k} \right)
= \exp \left( -\frac{2 \beta^2}{\gamma} \right).
\]
Plugging this back above we get
\[
P \left[ \sup_{1 \leq s < t} \frac{M_s}{\omega(s)} \geq \beta \right] \leq \frac{\log t}{\log \gamma} \exp \left( -\frac{2 \beta^2}{\gamma} \right),
\text{(3.24)}
\]
Combining (3.22), (3.23), and (3.24) with \( \beta = \alpha \sqrt{\frac{\log t}{2}} \), we obtain
\[
P[\hat{\mu}_i, T_i(t-1) + \alpha H(T_i(t-1), 1/t) > \mu_i] \leq \frac{\log t}{\log \gamma} \exp \left( -\frac{\alpha^2 \log t}{\gamma} \right)
= \frac{\log t}{\log \gamma} \frac{1}{t^{\alpha^2/\gamma}}.
\]
Lemma 3.81. For any \( \gamma > 1 \), it holds that
\[
P[\mathcal{E}_{t,1}] \leq \frac{1}{\log \gamma} t^{-\alpha^2/\gamma} \log t,
\]
and similarly for \( P[\mathcal{E}_{t,2}] \).

We are ready to prove the main result.

Proof of Theorem 3.78. By (3.18) and Lemmas 3.79, 3.80 and 3.81, we have
\[
\overline{R}_n = \sum_{i=1}^{\Delta} \Delta_i E[T_i(n)] \leq \Delta \left( u_n + 2 \sum_{i=1}^{n} \frac{1}{\log \gamma} t^{-\alpha^2/\gamma} \log t \right).
\]
Recalling that \( \alpha > 1 \), we can choose \( \gamma > 1 \) such that \( \alpha^2/\gamma > 1 \). In that case, the series on the right hand side is summable and there is \( C_\alpha \in (0, +\infty) \) such that
\[
\overline{R}_n \leq \Delta \left( u_n + C_\alpha \right).
\]
That proves the claim.

Remark 3.82. A slightly better—and provably optimal—multiplicative constant in the pseudo-regret bound has been obtained by [GC11] using a variant of UCB called KL-UCB. The matching lower bound is due to [LR85]. See also [BCB12, Sections 2.3-2.4]. Further improvements can be obtained by using Bernstein’s rather than Hoeffding’s inequality [AMS09].
3.3 Electrical networks

In this section we develop a classical link between random walks and electrical networks. The electrical interpretation is merely a useful physical analogy. The mathematical substance of the connection starts with the following well-known observation.

Let \( (X_t) \) be a Markov chain with transition matrix \( P \) on a finite or countable state space \( V \). For two disjoint subsets \( A, Z \) of \( V \), the probability of hitting \( A \) before \( Z \)

\[
h(x) = \mathbb{P}_x[\tau_A < \tau_Z],
\]

seen as a function of the starting point \( x \in V \), is harmonic on \( W := (A \cup Z)^c \) in the sense that

\[
h(x) = \sum_y P(x, y) h(y), \quad \forall x \in W,
\]

where \( h \equiv 1 \) (respectively \( \equiv 0 \)) on \( A \) (respectively \( Z \)). Indeed by the Markov property, after one step of the chain, for \( x \in W \)

\[
\mathbb{P}_x[\tau_A < \tau_Z] = \sum_{y \notin A \cup Z} P(x, y) \mathbb{P}_y[\tau_A < \tau_Z]
+ \sum_{y \in A} P(x, y) \cdot 1 + \sum_{y \in Z} P(x, y) \cdot 0
= \sum_y P(x, y) \mathbb{P}_y[\tau_A < \tau_Z].
\]

Quantities such as (3.25) arise naturally, for instance in the study of recurrence, and the connection to potential theory, the study of harmonic functions, proves fruitful in that context—and beyond—as we outline in this section.

First we re-write (3.26) to reveal the electrical interpretation. For this we switch to reversible chains. Recall that a reversible Markov chain is equivalent to a random walk on a network \( \mathcal{N} = (G, c) \) where the edges of \( G \) correspond to transitions of positive probability. If the chain is reversible with respect to a stationary measure \( \pi \), then the edge weights are \( c(x, y) = \pi(x) P(x, y) \). In this notation (3.26) becomes

\[
h(x) = \frac{1}{c(x)} \sum_{y \sim x} c(x, y) h(y), \quad \forall x \in (A \cup Z)^c,
\]

where \( c(x) := \sum_{y \sim x} c(x, y) = \pi(x) \). In words, \( h(x) \) is the weighted average of its neighboring values. Now comes the electrical analogy: if one interprets \( c(x, y) \) as a conductance, a function satisfying (3.28) is known as a voltage or potential function. The voltages at \( A \) and \( Z \) are 1 and 0 respectively. We show in the next subsection by a martingale argument that, under appropriate conditions, such a voltage
exists and is unique. To see why martingales come in, let $\mathcal{F}_t = \sigma(X_0, \ldots, X_t)$ and $\tau^* := \tau_{A \cup Z}$. Notice that, by a one-step calculation again, (3.26) implies that
\[
h(X_{t \wedge \tau^*}) = \mathbb{E} \left[ h(X_{(t+1) \wedge \tau^*}) \mid \mathcal{F}_t \right], \quad \forall t \geq 0,
\]
i.e., $(h(X_{t \wedge \tau^*}))_t$ is a martingale with respect to $(\mathcal{F}_t)$.

### 3.3.1 Martingales and the Dirichlet problem

Although the rest of Section 3.3 is concerned with reversible Markov chains, the current subsection applies to the non-reversible case as well. The following definition will be useful below. Let $\mathcal{Z}$ be a stopping time for a Markov chain $(X_t)$. The Green function of the chain stopped at $\mathcal{Z}$ is given by
\[
G_{\mathcal{Z}}(x, y) = \mathbb{E}_x \left[ \sum_{0 \leq t < \mathcal{Z}} 1 \{ X_t = y \} \right], \quad x, y \in V
\]
i.e., it is the expected number of visits to $y$ before $\mathcal{Z}$ when started at $x$. We will use the notation $h|_{\mathcal{Z}}$ for the function $h$ restricted to the subset $\mathcal{Z}$.

### Existence and uniqueness of a harmonic extension

We begin with a general problem.

**Theorem 3.83** (Existence and uniqueness). Let $P$ be an irreducible transition matrix on a finite or countable state space $V$. Let $W$ be a finite, proper subset of $V$ and let $h : W^c \to \mathbb{R}$ be a bounded function on $W^c = V \setminus W$. Then there exists a unique extension of $h$ to $W$ that is harmonic on $W$, i.e., it satisfies
\[
h(x) = \sum_y P(x, y) h(y), \quad \forall x \in W. \tag{3.29}
\]
The solution is given by $h(x) = \mathbb{E}_x[h(X_{\tau_{W^c}})]$.

**Proof.** We first argue about uniqueness. Suppose $h$ is defined over all of $V$ and satisfies (3.29). Let $\tau^* := \tau_{W^c}$. Then the process $(h(X_{t \wedge \tau^*}))_t$ is a martingale: on $\{\tau^* \leq t\}$,
\[
\mathbb{E}[h(X_{(t+1) \wedge \tau^*}) \mid \mathcal{F}_t] = h(X_{\tau^*}) = h(X_{t \wedge \tau^*}),
\]
and on $\{\tau^* > t\}$
\[
\mathbb{E}[h(X_{(t+1) \wedge \tau^*}) \mid \mathcal{F}_t] = \sum_y P(X_t, y) h(y) = h(X_t) = h(X_{t \wedge \tau^*}).
\]
Because \( W \) is finite and the chain is irreducible, we have \( \tau^* < +\infty \) a.s. See Lemma 3.25. Moreover the process is bounded because \( h \) is bounded on \( W^c \) and \( W \) is finite. Hence by the bounded convergence theorem (or the optional stopping theorem)

\[
h(x) = \mathbb{E}_x[h(X_0)] = \mathbb{E}_x[h(X_{t \wedge \tau^*})] \to \mathbb{E}_x[h(X_{\tau^*})], \quad \forall x \in W,
\]

which implies that \( h \) is unique.

For the existence, simply define

\[
h(x) = \mathbb{E}_x[h(X_{\tau^*})], \quad \forall x \in W,
\]

and use the Markov property as in (3.27).

For some insights on what happens when the assumptions of Theorem 3.83 are not satisfied, see Exercise 3.3. For an alternative proof of uniqueness based on the maximum principle, see Exercise 3.4.

The previous result is related to the classical Dirichlet problem in partial differential equations. To see the connection, note first that the proof above still works if one only specifies \( h \) on the outer boundary of \( W \)

\[
\partial_N W = \{z \in V \setminus W : \exists y \in W, P(y, z) > 0\}.
\]

Introduce the \textit{Laplacian} operator on \( \mathcal{N} \)

\[
\Delta_{\mathcal{N}} f(x) = \left[ \sum_y P(x, y)f(y) \right] - f(x) = \sum_y P(x, y)[f(y) - f(x)].
\]

We have proved that, under the assumptions of Theorem 3.83, there exists a unique solution to

\[
\begin{aligned}
\Delta_{\mathcal{N}} f(x) &= 0, \quad \forall x \in W, \\
f(x) &= h(x), \quad \forall x \in \partial_N W,
\end{aligned}
\]

and that solution is given by \( f(x) = \mathbb{E}_x[h(X_{\tau^*_W})] \), for \( x \in W \cup \partial_N W \). The system (3.30) is called a \textit{Dirichlet problem}. The Laplacian above can be interpreted as a discretized version of the standard Laplacian. For instance, for simple random walk on \( \mathbb{Z} \) (with \( \pi \equiv 1 \)), \( \Delta_{\mathcal{N}} f(x) = \frac{1}{2} \{[f(x + 1) - f(x)] - [f(x) - f(x - 1)]\} \)
which is a discretized second derivative.
Applications  Before developing the electrical network theory, we point out that Theorem 3.83 has many more applications. One of its consequences is that harmonic functions on finite, connected networks are constant.

**Corollary 3.84.** Let $P$ be an irreducible transition matrix on a finite state space $V$. If $h$ is harmonic on all of $V$, then it is constant.

**Proof.** Fix the value of $h$ at an arbitrary vertex $z$ and set $W = V \setminus \{z\}$.

Applying Theorem 3.83, for all $x \in W$, $h(x) = \mathbb{E}_x[h(X_{\tau_W})] = h(z)$.

As an example of application of this corollary, we prove the following surprising result: in a finite, irreducible Markov chain, the expected time to hit a target chosen at random according to the stationary distribution does not depend on the starting point.

**Theorem 3.85 (Random target lemma).** Let $(X_t)$ be an irreducible Markov chain on a finite state space $V$ with transition matrix $P$ and stationary distribution $\pi$. Then

$$h(x) := \sum_{y \in V} \pi(y) \mathbb{E}_x[\tau_y]$$

does not in fact depend on $x$.

**Proof.** By assumption, $\mathbb{E}_x[\tau_y] < +\infty$ for all $x, y$. By Corollary 3.84, it suffices to show that $h(x) := \sum_{y} \pi(y) \mathbb{E}_x[\tau_y]$ is harmonic on all of $V$. As before, it is natural to expand $\mathbb{E}_x[\tau_y]$ according to the first step of the chain,

$$\mathbb{E}_x[\tau_y] = \mathbb{1}_{\{x \neq y\}} \left( 1 + \sum_z P(x, z) \mathbb{E}_z[\tau_y] \right).$$

Substituting into $h(x)$ gives

$$h(x) = (1 - \pi(x)) + \sum_{z \neq x} \sum_y \pi(y) P(x, z) \mathbb{E}_z[\tau_y]$$

$$= (1 - \pi(x)) + \sum_z P(x, z) \left( h(z) - \pi(x) \mathbb{E}_x[\tau_x] \right).$$

Rearranging, we get

$$\Delta_N h(x) = \pi(x) \left( 1 + \sum_z P(x, z) \mathbb{E}_z[\tau_x] \right) - 1 = 0,$$

where we used $1/\pi(x) = \mathbb{E}_x[\tau_x^+] = 1 + \sum_z P(x, z) \mathbb{E}_z[\tau_x]$.  

150
3.3.2 Basic electrical network theory

We now develop the basic theory of electrical networks and their connections to random walks. We begin with a few definitions.

Definitions Let \( \mathcal{N} = (G, c) \) be a finite or countable network. Throughout this section we assume that \( \mathcal{N} \) is connected and locally finite. In the context of electrical networks, edge weights are called conductances. The reciprocal of the conductances are called resistances and are denoted by \( \rho(x, y) = \frac{1}{c(x, y)} \), for all \( x \sim y \). Both \( c \) and \( \rho \) are symmetric. For an edge \( e = \{x, y\} \) we also write \( c(e) := c(x, y) \) and \( \rho(e) := \rho(x, y) \). Recall that the transition matrix of the random walk on \( \mathcal{N} \) satisfies

\[
P(x, y) = \frac{c(x, y)}{\sum_{y \sim x} c(x, y)}.
\]

Let \( A, Z \) be disjoint, non-empty subsets of \( V \) such that \( W := (A \cup Z)^c \) is finite. For our purposes it will suffice to take \( A \) to be a singleton, i.e. \( A = \{a\} \) for some \( a \). Then \( a \) is called the source and \( Z \) is called the sink-set, or sink for short.

As an immediate corollary of Theorem 3.83, we obtain the existence and uniqueness of a voltage function, defined formally in the next corollary. It will be useful to consider voltages taking an arbitrary value at \( a \), but we always set the voltage on \( Z \) to 0. Note in the definition below that if \( v \) is a voltage with value \( v_0 \) at \( a \), then \( \tilde{v}(x) = v(x)/v_0 \) is a voltage with value 1 at \( a \).

**Corollary 3.86 (Voltage).** Fix \( v_0 > 0 \). Let \( \mathcal{N} = (G, c) \) be a finite or countable, connected network with \( G = (V, E) \). Let \( A := \{a\}, Z \) be disjoint non-empty subsets of \( V \) such that \( W = (A \cup Z)^c \) is non-empty and finite. Then there exists a unique voltage, i.e., a function \( v \) on \( V \) such that \( v \) is harmonic on \( W \)

\[
v(x) = \frac{1}{c(x)} \sum_{y \sim x} c(x, y)v(y), \quad \forall x \in W,
\]

where \( c(x) = \sum_{y \sim x} c(x, y) \), and

\[
v(a) = v_0 \quad \text{and} \quad v|_Z \equiv 0.
\]

Moreover

\[
\frac{v(x)}{v_0} = \mathbb{P}_x[\tau_a < \tau_Z],
\]

for the corresponding random walk on \( \mathcal{N} \).

**Proof.** Set \( h(x) = v(x) \) on \( A \cup Z \). Theorem 3.83 gives the result. \( \blacksquare \)
Let $v$ be a voltage function on $\mathcal{N}$ with source $a$ and sink $Z$. The Laplacian-based formulation of harmonicity, (3.30), can be interpreted in flow terms as follows. We define the current function $i(x, y) := c(x, y)[v(x) - v(y)]$ or, equivalently, $v(x) - v(y) = r(x, y)i(x, y)$. The latter definition is usually referred to as Ohm’s “law.” Notice that the current function is defined on ordered pairs of vertices and is anti-symmetric, i.e., $i(x, y) = -i(y, x)$. In terms of the current function, the harmonicity of $v$ is then expressed as

$$\sum_{y \sim x} i(x, y) = 0, \quad \forall x \in W,$$

i.e., $i$ is a flow on $W$. This set of equations is known as Kirchhoff’s node law. We also refer to these constraints as flow-conservation constraints. To be clear, the current function is not just any flow. It is a flow that can be written as a potential difference according to Ohm’s law. Such a current also satisfies Kirchhoff’s cycle law: if $x_1 \sim x_2 \sim \cdots \sim x_k \sim x_{k+1} = x_1$ is a cycle, then

$$\sum_{j=1}^{k} i(x_j, x_{j+1}) r(x_j, x_{j+1}) = 0,$$

as can be seen by substituting Ohm’s law. The strength of the current is defined as

$$\|i\| = \sum_{y \sim a} i(a, y).$$

The definition of $i(x, y)$ ensures that the flow out of the source is nonnegative as $P_y[\tau_a < \tau_Z] \leq 1 = P_a[\tau_a < \tau_Z]$ for all $y \sim a$. Note that by multiplying the voltage by a constant we obtain a current which is similarly scaled. Up to that scaling, the current function is unique from the uniqueness of the voltage. We will often consider the unit current where we scale $v$ and $i$ so as to enforce that $\|i\| = 1$.

**Remark 3.87.** Note that the definition of the current depends crucially on the reversibility of the chain, i.e., on the fact that $c(x, y) = c(y, x)$. For non-reversible chains, it is not clear how to interpret the system (3.30) as flow conservation as it involves only the outgoing transitions (which in general are not related to the incoming transitions).

Summing up the previous paragraph, to determine the voltage it suffices to find functions $v$ and $i$ that simultaneously satisfy Ohm’s law and Kirchhoff’s node law. Here is an example.

**Example 3.88** (Network reduction: birth-and-death chain). Let $\mathcal{N}$ be the line on \{0, 1, \ldots, $n$\} with $j \sim k \iff |j - k| = 1$ and arbitrary (positive) conductances
on the edges. Let \((X_t)\) be the corresponding walk. We use the principle above to compute \(\mathbb{P}_x[\tau_0 < \tau_n]\) for \(1 \leq x \leq n - 1\). Consider the voltage function \(v\) when \(v(0) = 1\) and \(v(n) = 0\) with current \(i\). The desired quantity is \(v(x)\) by Corollary 3.86. Note that because \(i\) is a flow on \(N\), the flow into every vertex equals the flow out of that vertex, and we must have \(i(y, y + 1) = i(0, 1) = ||i||\) for all \(y\). To compute \(v(x)\), we note that it remains the same if we replace the path \(0 \sim 1 \sim \cdots \sim x\) with a single edge of resistance \(R_{0,x} = r(0, 1) + \cdots + r(x - 1, x)\).

Indeed leave the voltage unchanged on the remaining nodes and define the current on the new edge as \(||i||\). Kirchhoff’s node law is automatically satisfied by the argument above. To check Ohm’s law on the new “super-edge,” note that on the original network \(N\)

\[
v(0) - v(x) = (v(0) - v(1)) + \cdots + (v(x - 1) - v(x))
= r(x - 1, x) i(x - 1, x) + \cdots + r(0, 1) i(0, 1)
= [r(0, 1) + \cdots + r(x - 1, x)] ||i||
= R_{0,x} ||i||.
\]

Ohm’s law is also satisfied on every other edge because nothing has changed there. That proves the claim. We do the same reduction on the other side of \(x\) by replacing \(x \sim x + 1 \sim \cdots \sim n\) with a single edge of resistance \(R_{x,n} = r(x, x + 1) + \cdots + r(n - 1, n)\). See Figure 3.88. Because the voltage at \(x\) has not changed, we can compute \(v(x) = \mathbb{P}_x[\tau_0 < \tau_n]\) directly on the reduced network, where it is now a straightforward computation. Indeed, starting at \(x\), the reduced walk jumps to 0 with probability proportional to the conductance on the new super-edge \(0 \sim x\) (or the reciprocal of the resistance), i.e.,

\[
\mathbb{P}_x[\tau_0 < \tau_n] = \frac{R_{0,x}^{-1}}{R_{0,x}^{-1} + R_{x,n}^{-1}}
= \frac{R_{x,n}}{R_{x,n} + R_{0,x}}
= \frac{r(x, x + 1) + \cdots + r(n - 1, n)}{r(0, 1) + \cdots + r(n - 1, n)}.
\]

Some special cases:

- **Simple random walk.** In the case of simple random walk, all resistances are equal and we get

\[
\mathbb{P}_x[\tau_0 < \tau_n] = \frac{n - x}{n}.
\]
• gambler’s ruin. the gambler’s ruin example corresponds to taking \( c(j, j + 1) = (p/q)^j \) or \( r(j, j + 1) = (q/p)^j \), for some \( 0 < p < 1 \). in this case we obtain

\[
\mathbb{P}_x[\tau_0 < \tau_n] = \frac{\sum_{j=0}^{n-1}(q/p)^j}{\sum_{j=0}^{n-1}(q/p)^j} = \frac{(q/p)^x(1 - (q/p)^{n-x})}{1 - (q/p)^n} = \frac{(p/q)^{n-x} - 1}{(p/q)^n - 1},
\]

when \( p \neq q \) (otherwise we get back the simple random walk case).

the above example illustrates the series law: resistances in series add up. there is a similar parallel law: conductances in parallel add up. to formalize these laws, one needs to introduce multigraphs. this is straightforward, but to avoid complicating the notation further we will not do this here. (see the “bibliographic remarks” for more.) another useful network reduction technique is shorting, in which we identify, or glue together, vertices with the same voltage while keeping existing edges. here is an example.

**example 3.89** (network reduction: binary tree). let \( \mathcal{N} \) be the rooted binary tree with \( n \) levels \( \mathbb{T}_2^n \) and equal conductances on all edges. let \( 0 \) be the root. pick an arbitrary leaf and denote it by \( n \). the remaining vertices on the path between 0
and \( n \), which we refer to as the main path, will be denoted by \( 1, \ldots, n-1 \) moving away from the root. We claim that, for all \( 0 < x < n \), it holds that

\[
P_x[\tau_0 < \tau_n] = (n - x)/n.
\]

Indeed let \( v \) be the voltage with values 1 and 0 at \( a = 0 \) and \( Z = \{ z \} \) with \( z = n \) respectively. Let \( i \) be the corresponding current. Notice that, for each \( 0 \leq y < n \), the current—as a flow—has nowhere to go on the subtree \( T_y \) hanging from \( y \) away from the main path. The leaves of the subtree are dead ends. Hence the current must be 0 on \( T_y \) and by Ohm’s law the voltage must be constant on it, i.e., every vertex in \( T_y \) has voltage \( v(y) \). Imagine collapsing all vertices in \( T_y \), including \( y \), into a single vertex (and removing the self-loops so created). Doing this for every vertex on the main path results in a new reduced network which is formed of a single path as in Example 3.88. Note that the voltage and the current can be taken to be the same as they were previously on the main path. Indeed, with this choice, Ohm’s law is automatically satisfied. Moreover, because there is no current on the hanging subtrees in the original network, Kirchhoff’s node law is also satisfied on the reduced network, as no current is lost. Hence the answer can be obtained from Example 3.88. That proves the claim. (You should convince yourself that this result is obvious from a probabilistic point of view.)

We gave a probabilistic interpretation of the voltage. What about the current? The following result says that, roughly speaking, \( i(x, y) \) is the net traffic on the edge \( \{ x, y \} \) from \( x \) to \( y \). We start with an important formula for the voltage at \( a \). For the walk started at \( a \), we use the shorthand

\[
P[a \to Z] := P_a[\tau_Z < \tau_a],
\]

for the escape probability.

**Lemma 3.90.** Let \( v \) be a voltage on \( \mathcal{N} \) with source \( a \) and sink \( Z \). Let \( i \) be the associated current. Then

\[
\frac{v(a)}{\|i\|} = \frac{1}{c(a) P[a \to Z]}.
\]

(3.34)
Proof. Using the usual one-step trick,

\[
P[a \to Z] = \sum_{x} P(a, x) P_x[\tau_Z < \tau_a] = \sum_{x} c(a, x) \left( 1 - \frac{v(x)}{v(a)} \right) = \frac{1}{c(a)v(a)} \sum_{x} c(a, x)[v(a) - v(x)] = \frac{1}{c(a)v(a)} \sum_{x} i(a, x),
\]

where we used Corollary 3.86 on the second line and Ohm’s law on the last line. Rearranging gives the result.

\[\blacksquare\]

Theorem 3.91 (Probabilistic interpretation of the current). For \(x \sim y\), let \(N^Z_{x \to y}\) be the number of transitions from \(x\) to \(y\) up to the time of the first visit to the sink \(Z\) for the random walk on \(N\) started at \(a\). Let \(v\) be the voltage corresponding to the unit current \(i\). Then the following formulas hold:

\[v(x) = \frac{\mathcal{G}_{\tau_Z}(a, x)}{c(x)}, \quad \forall x, \quad (3.35)\]

and

\[i(x, y) = \mathbb{E}_a[N^Z_{x \to y} - N^Z_{y \to x}], \quad \forall x \sim y.\]

Proof. We prove the formula for the voltage by showing that \(v(x)\) as defined above is harmonic on \(W = V \setminus \{a\} \cup Z\). Note first that \(\mathcal{G}_{\tau_Z}(a, z) = 0\) for all \(z \in Z\) by definition, or \(0 = v(z) = \frac{\mathcal{G}_{\tau_Z}(a, z)}{c(z)}\). Moreover, to compute \(\mathcal{G}_{\tau_Z}(a, a)\), note that the number of visits to \(a\) before the first visit to \(Z\) is geometric with success probability \(\mathbb{P}[a \to Z]\) by the strong Markov property and hence

\[\mathcal{G}_{\tau_Z}(a, a) = \frac{1}{\mathbb{P}[a \to Z]},\]

and by the previous lemma \(v(a) = \frac{\mathcal{G}_{\tau_Z}(a, a)}{c(a)}\), as required. To establish the formula for \(x \in W\), we compute the quantity \(\frac{1}{c(x)} \sum_{y \sim x} \mathbb{E}_a[N^Z_{y \to x}]\) in two ways. First, because each visit to \(x \in W\) must enter through one of \(x\)’s neighbors (including itself in the presence of a self-loop), we get

\[\frac{1}{c(x)} \sum_{y \sim x} \mathbb{E}_a[N^Z_{y \to x}] = \frac{\mathcal{G}_{\tau_Z}(a, x)}{c(x)}. \quad (3.36)\]
On the other hand,

\[
E_a[N_{y \to x}^Z] = E_a \left[ \sum_{0 \leq t < \tau_Z} 1_{\{X_t = y, X_{t+1} = x\}} \right]
\]

\[
= \sum_{t \geq 0} P_a [X_t = y, X_{t+1} = x, \tau_Z > t]
\]

\[
= \sum_{t \geq 0} P_a [\tau_Z > t] P_a [X_t = y | \tau_Z > t] P(y, x)
\]

\[
= P(y, x) E_a \left[ \sum_{0 \leq t < \tau_Z} 1_{\{X_t = y\}} \right]
\]

\[
= P(y, x) \mathcal{G}_{\tau_Z}(a, y),
\] (3.37)

so that, summing over \(y\), we obtain this time

\[
\frac{1}{c(x)} \sum_{y \sim x} E_a[N_{y \to x}^Z] = \frac{1}{c(x)} \sum_{y \sim x} P(y, x) \mathcal{G}_{\tau_Z}(a, y) = \sum_{y \sim x} P(x, y) \frac{\mathcal{G}_{\tau_Z}(a, y)}{c(y)},
\] (3.38)

where we used reversibility. Equating (3.36) and (3.38) shows that \(\mathcal{G}_{\tau_Z}(a, x)\) is harmonic on \(W\) and hence must be equal to the voltage function by Corollary 3.86.

Finally, by (3.37),

\[
E_a[N_{x \to y}^Z - N_{y \to x}^Z] = P(x, y) \mathcal{G}_{\tau_Z}(a, x) - P(y, x) \mathcal{G}_{\tau_Z}(a, y)
\]

\[
= P(x, y) v(x) c(x) - P(y, x) v(y) c(y)
\]

\[
= c(x, y) [v(x) - v(y)]
\]

\[
= i(x, y).
\]

That concludes the proof.

Remark 3.92. Formula (3.35) relies crucially on reversibility. Indeed assume the chain has stationary distribution \(\pi\). Then, in probabilistic terms, (3.35) reads

\[
\pi(x) P_x [\tau_a < \tau_Z] = \frac{\mathcal{G}_{\tau_Z}(a, x)}{\pi(a) P[a \rightarrow Z]},
\]

where we used (3.33) and (3.34), and the fact that the current has unit strength. This is not true in general for non-reversible chains. Take for instance a deterministic walk on a directed cycle of size \(n\) with \(x\) on the directed path from \(a\) to \(Z = \{z\}\). In that case the l.h.s. is 0 but the r.h.s. is \(n\).
Example 3.93 (Network reduction: binary tree (continued)). Recall the setting of Example 3.89. We argued that the current on side edges, i.e., edges of subtrees hanging from the main path, is 0. This is clear from the probabilistic interpretation of the current: in a walk from \( a \) to \( z \), any traversal of a side edge must be undone at a later time.

The network reduction techniques illustrated above are useful. But the power of the electrical network perspective is more apparent in what comes next: the definition of the effective resistance and, especially, its variational characterization.

Effective resistance \hspace{1cm} Before proceeding further, let us recall our original motivation. Let \( \mathcal{N} = (G, c) \) be a countable, locally finite, connected network and let \( (X_t) \) be the corresponding walk. Recall that a vertex \( a \) in \( G \) is transient if \( \Pr[a \xrightarrow{\tau_a^+} +\infty] < 1 \).

To relate this to our setting, consider an exhaustive sequence of induced subgraphs \( G_n \) of \( G \) which for our purposes is defined as: \( G_0 \) contains only \( a \), \( G_n \subseteq G_{n+1} \), \( G = \bigcup_n G_n \), and every \( G_n \) is finite and connected. Such a sequence always exists by iteratively adding the neighbors of the previous vertices and using that \( G \) is locally finite and connected. Let \( Z_n \) be the set of vertices of \( G \) not in \( G_n \). Then, by Lemma 3.25, \( \Pr[a \xrightarrow{\tau_a^+} +\infty] = 0 \) for all \( n \) by our assumptions on \( (G_n) \).

Hence, the remaining possibilities are

\[
1 = \Pr[a \xrightarrow{\exists n, \tau_a^+ < \tau_{Z_n}}] + \Pr[a \xrightarrow{\forall n, \tau_{Z_n} < \tau_a^+}]
\]

\[
= \Pr[a \xrightarrow{\tau_a^+ < +\infty}] + \lim_n \Pr[a \rightarrow Z_n].
\]

Therefore \( a \) is transient if and only if \( \lim_n \Pr[a \rightarrow Z_n] > 0 \). Note that the limit exists because the sequence of events \( \{\tau_{Z_n} < \tau_a^+\} \) is decreasing by construction. By a sandwiching argument the limit also does not depend on the exhaustive sequence. (Exercise.) Hence we define

\[
\Pr[a \rightarrow \infty] := \lim_n \Pr[a \rightarrow Z_n] > 0.
\]

We use Lemma 3.90 to characterize this limit using electrical network notions.

But, first, here comes the key definition. In Lemma 3.90, \( v(a) \) can be thought of as the potential difference between the source and the sink, and \( ||i|| \) can be thought of as the total current flowing through the network from the source to the sink. Hence, viewing the network as a single \( \text{“super-edge,”} \) Equation (3.34) is the analogue of Ohm’s law if we interpret \( c(a) \Pr[a \rightarrow Z] \) as a \( \text{“conductance.”} \)

Definition 3.94 (Effective resistance and conductance). Let \( \mathcal{N} = (G, c) \) be a finite or countable, locally finite, connected network. Let \( A = \{a\} \) and \( Z \) be disjoint
non-empty subsets of the vertex set $V$ such that $W := V \setminus (A \cup Z)$ is finite. Let $v$ be a voltage from source $a$ to sink $Z$ and let $i$ be the corresponding current. The effective resistance between $a$ and $Z$ is defined as

$$R(a \leftrightarrow Z) := \frac{1}{c(a) \mathbb{P}[a \rightarrow Z]} = \frac{v(a)}{\|i\|},$$

where the rightmost equality holds by Lemma 3.90. The reciprocal is called the effective conductance and denoted by $\mathcal{C}(a \leftrightarrow Z) := 1/R(a \leftrightarrow Z)$.

Going back to recurrence, for an exhaustive sequence $(G_n)$ with $(Z_n)$ as above, it is natural to define

$$R(a \leftrightarrow \infty) := \lim_n R(a \leftrightarrow Z_n),$$

where, once again, the limit does not depend on the choice of exhaustive sequence.

**Theorem 3.95 (Recurrence and resistance).** Let $\mathcal{N} = (G, c)$ be a countable, locally finite, connected network. Vertex $a$ (and hence all vertices) in $\mathcal{N}$ is transient if and only if $R(a \leftrightarrow \infty) < +\infty$.

**Proof.** This follows immediately from the definition of the effective resistance. Recall that, on a connected network, all states have the same type (recurrent or transient).

Note that the network reduction techniques we discussed previously leave both the voltage and the current strength unchanged on the reduced network. Hence they also leave the effective resistance unchanged.

**Example 3.96 (Gambler’s ruin chain revisited).** Extend the gambler’s ruin chain of Example 3.88 to all of $\mathbb{Z}_+$. We determine when this chain is transient. Because it is irreducible, all states have the same type and it suffices to look at 0. Consider the exhaustive sequence obtained by letting $G_n$ be the graph restricted to $[0, n-1]$ and letting $Z_n = [n, +\infty)$. To compute the effective resistance $R(0 \leftrightarrow Z_n)$, we use the same reduction as in Example 3.88, except that this time we reduce the network all the way to a single edge. That edge has resistance

$$R(0 \leftrightarrow Z_n) = \sum_{j=0}^{n-1} r(j, j + 1) = \sum_{j=0}^{n-1} (q/p)^j = \frac{(q/p)^n - 1}{(q/p) - 1},$$

when $p \neq q$, and similarly it has value $n$ in the $p = q$ case. Hence

$$R(0 \leftrightarrow \infty) = \begin{cases} +\infty, & p \leq 1/2, \\ \frac{p}{2p-1}, & p > 1/2. \end{cases}$$

So 0 is transient if and only if $p > 1/2$. 

159
Example 3.97 (Biased walk on the $b$-ary tree). Fix $\lambda \in (0, +\infty)$. Consider the rooted, infinite $b$-ary tree with conductance $\lambda^j$ on all edges between level $j - 1$ and $j$, $j \geq 1$. We determine when this chain is transient. Because it is irreducible, all states have the same type and it suffices to look at the root. Denote the root by 0. For an exhaustive sequence, let $G_n$ be the root together with the first $n - 1$ levels. Let $Z_n$ be as before. To compute $R(0 \leftrightarrow Z_n)$: 1) glue together all vertices of $Z_n$; 2) glue together all vertices on the same level of $G_n$; 3) replace parallel edges with a single edge whose conductance is the sum of the conductances; 4) let the current on this edge be the sum of the currents; and 5) leave the voltages unchanged. Note that Ohm’s law and Kirchhoff’s node law are still satisfied. Hence we have not changed the effective resistance. (This is an application of the parallel law.) The reduced network is now a line. Denote the new vertices $0, 1, \ldots, n$. The conductance on the edge between $j$ and $j + 1$ is $b^{j+1}\lambda^j = b(b\lambda)^j$. So this is the chain from the previous example with $(p/q) = b\lambda$ where all conductances are scaled by a factor of $b$. Hence

$$R(0 \leftrightarrow \infty) = \begin{cases} +\infty, & b\lambda \leq 1, \\ \frac{1}{b(1-(b\lambda)^{-1})}, & b\lambda > 1. \end{cases}$$

So the root is transient if and only if $b\lambda > 1$. \hfill $\blacktriangle$

3.3.3 Bounding the effective resistance

The examples we analyzed so far were atypical in that it was possible to reduce the network down to a single edge using simple rules and read off the effective resistance. In general, we need more robust techniques to bound the effective resistance. The following two variational principles provide a powerful approach for this purpose.

Variational principles Recall that flow $\theta$ from source $a$ to sink $Z$ on a countable, locally finite, connected network $\mathcal{N} = (G, c)$ is a function on pairs of adjacent vertices such that: $\theta$ is anti-symmetric, i.e., $\theta(x, y) = -\theta(y, x)$ for all $x \sim y$; and it satisfies the flow-conservation constraint $\sum_{y \sim x} \theta(x, y) = 0$ on all vertices $x$ except those in $\{a\} \cup Z$. The strength of the flow is $||\theta|| := \sum_{y \sim a} \theta(a, y)$. The current is a special flow—one that can be written as a potential difference according to Ohm’s law. As we show next, it can also be characterized as a flow minimizing a certain energy. Specifically, the energy of a flow $\theta$ is defined as

$$E(\theta) = \frac{1}{2} \sum_{x,y} r(x, y) [\theta(x, y)]^2.$$
The proof of the variational principle we present here employs a neat trick, convex duality. In particular, it reveals that the voltage and current are dual in the sense of convex analysis.

**Theorem 3.98 (Thomson’s principle).** Let \( N = (G, c) \) be a finite, connected network. The effective resistance between source \( a \) and sink \( Z \) is characterized by

\[
\mathcal{R}(a \leftrightarrow Z) = \inf \{ \mathcal{E}(\theta) : \theta \text{ is a unit flow between } a \text{ and } Z \}. \tag{3.39}
\]

The unique minimizer is the unit current.

**Proof.** It will be convenient to work in matrix form. Choose an arbitrary orientation of \( N \), i.e., replace each edge \( \{x, y\} \) with either \( h_{x,y} \) or \( h_{y,x} \). Let \( 
abla \) be the corresponding directed graph. Think of the flow \( \theta \) as a vector with one component for each oriented edge. Then the flow constraint can be written as a linear system

\[
A \theta = b.\]

Here the matrix \( A \) has a column for each edge and a row for each vertex except those in \( Z \). The entries of \( A \) are of the form \( A_{x,(x,y)} = 1 \), \( A_{y,(x,y)} = -1 \), and 0 otherwise. The vector \( b \) has 0s everywhere except for \( b_a = 1 \). Let \( r \) be the vector of resistances and let \( R \) be the diagonal matrix with diagonal \( r \). In matrix form the optimization problem (3.39) reads

\[
\mathcal{E}^* = \inf \{ \theta' R \theta : A \theta = b \},
\]

where \( \theta \) denotes the transpose.

Introduce the Lagrangian

\[
\mathcal{L}(\theta; h) := \theta' R \theta - 2h' (A \theta - b),
\]

where \( h \) has an entry for all vertices except those in \( Z \). (The reason for the factor of 2 will be clear below.) For all \( h \),

\[
\mathcal{E}^* \geq \inf_{\theta} \mathcal{L}(\theta; h),
\]

because those \( \theta \)s with \( A \theta = b \) make the second term vanish in \( \mathcal{L}(\theta; h) \). Since \( \mathcal{L}(\theta; h) \) is strictly convex as a function of \( \theta \), the solution is characterized by the usual optimality condition which in this case reads \( 2R \theta - 2A' h = 0 \), or

\[
\theta = R^{-1} A' h. \tag{3.40}
\]

Substituting into the Lagrangian and simplifying, we have proved that

\[
\mathcal{E}(\theta) \geq \mathcal{E}^* \geq -h' AR^{-1} A' h + 2h' b =: \mathcal{L}^*(h), \quad \forall h \text{ and flow } \theta. \tag{3.41}
\]
This inequality is a statement of weak duality.

To show that a flow $\theta$ is optimal it suffices to find $h$ such that the l.h.s. in (3.41) equals $\mathcal{E}(\theta) = \theta^R \theta$. Not surprisingly, when $\theta$ is the unit current, the suitable $h$ turns out to be the corresponding voltage. To see this, observe that $A' h$ is the vector of neighboring node differences

$$A' h = (h(x) - h(y))_{(x,y) \in G}.$$  \hspace{1cm} (3.42)

Hence the optimality condition (3.40) is nothing but Ohm’s law in matrix form. Therefore, if $i$ is the unit current and $v$ is the associated voltage in vector form, it holds that

$$\mathcal{L}^* (v) = \mathcal{L}(i; v) = \mathcal{E}(i),$$

where the first equality follows from the optimality of $v$ and the second equality follows from the fact that $Ai = b$. So we must have $\mathcal{E}(i) = \mathcal{E}^*$. As for uniqueness, note that two minimizers $\theta, \theta'$ satisfy

$$\mathcal{E}^* = \frac{\mathcal{E}(\theta) + \mathcal{E}(\theta')}{2} = \mathcal{E} \left( \frac{\theta + \theta'}{2} \right) + \mathcal{E} \left( \frac{\theta - \theta'}{2} \right).$$

The first term on the r.h.s. is greater or equal than $\mathcal{E}^*$ because the average of two unit flows is still a unit flow. The second term is nonnegative by definition. Hence the latter must be zero and $\theta = \theta'$.

To conclude the proof, it remains to compute the optimal value. The matrix

$$\Delta_N := AR^{-1} A',$$

can be interpreted as the Laplacian operator of Section 3.3.1 in matrix form, i.e., for each row $x$ it takes a conductance-weighted average of the neighboring values and subtracts the value at $x$

$$(AR^{-1} A' v)_x = \sum_{y: (x,y) \in G} \left[ c(x,y) (v(x) - v(y)) \right]$$

$$- \sum_{y: (y,x) \in G} \left[ c(y,x) (v(y) - v(x)) \right]$$

$$= \sum_{y \sim x} \left[ c(x,y) (v(x) - v(y)) \right],$$

where we used (3.42) and the fact that $r(x,y)^{-1} = c(x,y)$ and $c(x,y) = c(y,x)$. So $\Delta_N v$ is zero everywhere except for the row corresponding to $a$ where it is

$$\sum_{y \sim a} c(a,y)(v(a) - v(y)) = \sum_{y \sim a} i(a,y) = 1,$$

162
where we used Ohm’s law and the fact that the current has unit strength. We have finally
\[
E^\ast = \mathcal{L}^\ast(v) = -v'AR^{-1}A'v + 2v'b = -v(a) + 2v(a) = v(a) = \mathcal{R}(a \leftrightarrow Z),
\]
by (3.90), where used again that \( \|i\| = 1 \).

Observe that the convex combination \( \alpha \) minimizing the sum of squares \( \sum_j \alpha_j^2 \) is uniform. In a similar manner, Thomson’s principle stipulates roughly speaking that the more the flow can be spread out over the network, the lower is the effective resistance (penalizing flow on edges with higher resistance). Pólya’s theorem below provides a vivid illustration. Here is a simple example suggesting that, in a sense, the current is indeed a well distributed flow.

**Example 3.99** (Random walk on the complete graph). Let \( \mathcal{N} \) be the complete graph on \( \{1, \ldots, n\} \) with unit resistances, and let \( a = 1 \) and \( Z = \{n\} \). Assume \( n > 2 \). The effective resistance is straightforward to compute in this case. Indeed, the escape probability (with a slight abuse of notation) is
\[
\mathbb{P}[1 \to n] = \frac{1}{n-1} + \frac{1}{2} \left( 1 - \frac{1}{n-1} \right) = \frac{n}{2(n-1)},
\]
as we either jump to \( n \) immediately or jump to one of the remaining nodes, in which case we reach \( n \) first with probability \( 1/2 \) by symmetry. Hence, since \( c(1) = n - 1 \), we get
\[
\mathcal{R}(1 \leftrightarrow n) = \frac{2}{n},
\]
from the definition of the effective resistance. We now look for the optimal flow. Putting a flow of \( 1 \) on the edge \((1, n)\) gives an upper bound of \( 1 \), which is far from the optimal \( \frac{2}{n} \). Spreading the flow a bit more by pushing \( 1/2 \) through the edge \((1, n)\) and \( 1/2 \) through the path \( 1 \sim 2 \sim n \) gives the slightly better bound \( 1/4 + 2(1/4) = 3/4 \). Taking this further, putting a flow of \( \frac{1}{n-1} \) on \((1, n)\) as well as on each two-edge path to \( n \) through the remaining neighbors of \( 1 \) gives the yet improved bound
\[
\frac{1}{(n-1)^2} + 2(n-2)\frac{1}{(n-1)^2} = \frac{2n - 3}{(n-1)^2} = \frac{2}{n} \cdot \frac{2n^2 - 3n}{2n^2 - 4n + 2} > \frac{2}{n},
\]
when \( n > 2 \). Because the direct path from \( 1 \) to \( n \) has a somewhat lower resistance, the optimal flow is obtained by increasing the flow on that edge slightly. Namely, for a flow \( \alpha \) on \((1, n)\) we get an energy of \( \alpha^2 + 2(n-2)\left(\frac{1}{n-2} \right)^2 \) which is minimized at \( \alpha = \frac{2}{n} \) where it is indeed
\[
\left( \frac{2}{n} \right)^2 + \frac{2}{n-2} \left( \frac{n-2}{n} \right)^2 = \frac{2}{n} \left( \frac{2}{n} + \frac{n-2}{n} \right) = \frac{2}{n}.
\]
The matrix \( \Delta_N = AR^{-1}A' \) in the proof of Thomson’s principle is the Laplacian matrix. As we noted above, because \( A'h \) is the vector of neighboring node differences, we have

\[
h' \Delta_N h = \frac{1}{2} \sum_{x,y} c(x, y)[h(y) - h(x)]^2,
\]

where we implicitly fix \( h|_Z \equiv 0 \), which is called the Dirichlet energy. Thinking of \( \nabla_N := A' \) as a discrete gradient operator, the Dirichlet energy can be interpreted as the weighted norm of the gradient of \( h \). The following dual to Thomson’s principle is essentially a reformulation of the Dirichlet problem. Exercise 3.6 asks for a proof.

**Theorem 3.100** (Dirichlet’s principle). Let \( \mathcal{N} = (G, c) \) be a finite, connected network. The effective conductance between source \( a \) and sink \( Z \) is characterized by

\[
\mathcal{C}(a \leftrightarrow Z) = \inf \left\{ \frac{1}{2} \sum_{x,y} c(x, y)[h(y) - h(x)]^2 : h(a) = 1, h|_Z \equiv 0 \right\}.
\]

The unique minimizer is the voltage \( v \) with \( v(a) = 1 \).

The following lower bound is a typical application of Thomson’s principle. See Pólya’s theorem below for an example of its use.

**Definition 3.101** (Cutset). On a finite graph, a cutset separating \( a \) from \( Z \) is a set of edges \( \Pi \) such that any path between \( a \) and \( Z \) must include at least one edge in \( \Pi \). Similarly, on a countable network, a cutset separating \( a \) from \( \infty \) is a set of edges that must be crossed by any infinite self-avoiding path from \( a \).

**Corollary 3.102** (Nash-Williams inequality). Let \( \mathcal{N} \) be a finite, connected network and let \( \{\Pi_j\}_{j=1}^n \) be a collection of disjoint cutsets separating source \( a \) from sink \( Z \). Then

\[
\mathcal{R}(a \leftrightarrow Z) \geq \sum_{j=1}^n \left( \sum_{e \in \Pi_j} c(e) \right)^{-1}.
\]

Similarly, if \( \mathcal{N} \) is a countable, locally finite, connected network, then for any collection \( \{\Pi_j\}_j \) of finite, disjoint cutsets separating \( a \) from \( \infty \),

\[
\mathcal{R}(a \leftrightarrow \infty) \geq \sum_j \left( \sum_{e \in \Pi_j} c(e) \right)^{-1}.
\]
Proof. We will need the following lemma.

**Lemma 3.103.** Let $\mathcal{N}$ be finite. For any flow $\theta$ between $a$ and $Z$ and any cutset $\Pi$ separating $a$ from $Z$, it holds that

$$\sum_{e \in \Pi} |\theta(e)| \geq \|\theta\|.$$  

Proof. Intuitively, the flow out of $a$ must cross $\Pi$ to reach $Z$. Formally, let $W_{\Pi}$ be the set of vertices reachable from $a$ without crossing $\Pi$, let $Z_{\Pi}$ be the set of vertices not in $W_{\Pi}$ that are incident with an edge in $\Pi$ and let $V_{\Pi} = W_{\Pi} \cup Z_{\Pi}$. For $x \in W_{\Pi}\{a\}$, note that by definition of a cutset $x \notin Z$. Moreover, all neighbors of $x$ in $V$ are in fact in $V_{\Pi}$: if $y \sim x$ is not in $Z_{\Pi}$ then it is reachable from $a$ through $x$ without crossing $\Pi$ and therefore it is in $W_{\Pi}$. Hence,

$$\sum_{y \in V_{\Pi} : y \sim x} \theta(x, y) = \sum_{y \in V : y \sim x} \theta(x, y) = 0,$$  \hspace{1cm} (3.43)

or in other words $\theta$ is a flow from $a$ to $Z_{\Pi}$ on the graph $G_{\Pi}$ induced by $V_{\Pi}$. By the same argument, this flow has strength

$$\sum_{y \in V_{\Pi} : y \sim a} \theta(a, y) = \sum_{y \in V : y \sim a} \theta(a, y) = \|\theta\|.$$  \hspace{1cm} (3.44)

By (3.43) and (3.44) and the anti-symmetry of $\theta$,

$$0 = \sum_{x \in V_{\Pi}} \sum_{y \sim x} \theta(x, y)$$
$$= \|\theta\| + \sum_{x \in Z_{\Pi}} \sum_{y \sim x} \theta(x, y)$$
$$= \|\theta\| + \sum_{x \in Z_{\Pi}} \sum_{y \in V_{\Pi}} \theta(x, y) + \sum_{x \in Z_{\Pi}} \sum_{y \in W_{\Pi}} \theta(x, y)$$
$$= \|\theta\| + \sum_{x \in Z_{\Pi}} \sum_{y \in W_{\Pi}} \theta(x, y)$$
$$\geq \|\theta\| - \sum_{e \in \Pi} |\theta(e)|,$$

as $x \in Z_{\Pi}, y \in W_{\Pi}, y \sim x$ implies $\{x, y\} \in \Pi$. That concludes the proof. \hfill \Box

Returning to the proof of the claim, consider the case where $\mathcal{N}$ is finite. For any
unit flow from $a$ to $Z$, by Cauchy-Schwarz and the lemma above

$$\sum_{e \in \Pi_j} c(e) \sum_{e \in \Pi_j} r(e) |\theta(e)|^2 \geq \left( \sum_{e \in \Pi_j} \sqrt{c(e)r(e)} |\theta(e)| \right)^2$$

$$= \left( \sum_{e \in \Pi_j} |\theta(e)| \right)^2$$

$$\geq 1.$$ 

Summing over $j$, using the disjointness of the cutsets, and rearranging gives the result in the finite case.

The infinite case follows from a similar argument. Note that, after removing a finite cutset $\Pi$ separating $a$ from $\infty$, the connected component containing $a$ must be finite by definition of $\Pi$.

Another typical application of Thomson’s principle is the following monotonicity property (which is not obvious from a probabilistic point of view).

**Corollary 3.104.** Adding an edge to a finite, connected network cannot increase the effective resistance between a source $a$ and a sink $Z$. In particular, if the added edge is not incident to $a$, then $P[a \to Z]$ cannot decrease.

**Proof.** The additional edge enlarges the space of possible flows, so by Thomson’s principle it can only lower the resistance or leave it as is. The second statement follows from the definition of the effective resistance.

More generally:

**Corollary 3.105 (Rayleigh’s principle).** Let $N$ and $N'$ be two networks on the same finite, connected graph $G$ such that, for each edge in $G$, the resistance in $N'$ is greater than it is in $N$. Then, for any source $a$ and sink $Z$,

$$R_N(a \leftrightarrow Z) \leq R_{N'}(a \leftrightarrow Z).$$

**Proof.** Compare the energies of an arbitrary flow on $N$ and $N'$, and apply Thomson’s principle.

**Application to recurrence** Combining Theorem 3.95 and Thomson’s principle, we derive a flow-based criterion for recurrence. To state the result, it is convenient to introduce the notion of a unit flow $\theta$ from source $a$ to $\infty$ on a countable, locally finite network: $\theta$ is anti-symmetric, it satisfies the flow-conservation constraint on all vertices but $a$, and $|\theta| := \sum_{y \sim a} \theta(a,y) = 1$. Note that the energy $E(\theta)$ of such a flow is well defined in $[0, \infty]$.
Theorem 3.106 (Recurrence and finite-energy flows). Let \( \mathcal{N} = (G, c) \) be a countable, locally finite, connected network. Vertex \( a \) (and hence all vertices) in \( \mathcal{N} \) is transient if and only if there is a unit flow from \( a \) to \( \infty \) of finite energy.

Proof. One direction is immediate. Suppose such a flow exists and has energy bounded by \( B < +\infty \). Let \( (G_n) \) be an exhaustive sequence with associated sinks \( (Z_n) \). A unit flow from \( a \) to \( Z_n \) yields, by projection, a unit flow from \( a \) to \( Z_n \). This projected flow also has energy bounded by \( B \). Hence Thomson’s principle implies \( \mathcal{R}(a \leftrightarrow Z_n) \leq B \) for all \( n \) and transience follows from Theorem 3.95.

Proving the other direction involves producing a flow to \( 1 \). Suppose \( a \) is transient and let \( (G_n) \) be an exhaustive sequence as above. Then Theorem 3.95 implies that \( \mathcal{R}(a \leftrightarrow Z_n) \leq \mathcal{R}(a \leftrightarrow 1) < B \) for some \( B < +\infty \) and Thomson’s principle guarantees in turn the existence of a flow \( \theta_n \) from \( a \) to \( Z_n \) with energy bounded by \( B \). In particular there is a unit current \( i_n \), and associated voltage \( v_n \), of energy bounded by \( B \). So it remains to use the sequence of current flows \( (i_n) \) to construct a flow to \( \infty \) on the infinite network. The technical point is to show that the limit of \( (i_n) \) exists and is indeed a flow. For this, consider the random walk on \( \mathcal{N} \) started at \( a \). Let \( Y_n(x) \) be the number of visits to \( x \) before hitting \( Z_n \) the first time. By the monotone convergence theorem, \( E_a Y_n(x) \rightarrow E_a Y_\infty(x) \) where \( Y_\infty(x) \) is the total number of visits to \( x \). By (3.35), \( E_a Y_n(x) = c(x)v_n(x) \). So we can now define

\[
v_\infty(x) := \lim_n v_n(x),
\]

and then

\[
i_\infty(x, y) := c(x, y)[v_\infty(y) - v_\infty(x)] = \lim_n c(x, y)[v_n(y) - v_n(x)] = \lim_n i_n(x, y),
\]

by Ohm’s law (when \( n \) is large enough that both \( x \) and \( y \) are in \( G_n \)). Because \( i_n \) is a flow for all \( n \), by taking limits in the flow-conservation constraints we see that so is \( i_\infty \). Note that the partial sums

\[
\sum_{x, y \in G_n} c(x, y)[i_\infty(x, y)]^2 = \lim_{\ell \geq n} \sum_{x, y \in G_n} c(x, y)[i_\ell(x, y)]^2 \leq \limsup_{\ell \geq n} \mathcal{E}(i_\ell) < B,
\]

uniformly in \( n \). Because the l.h.s. converges to the energy of \( i_\infty \) by the monotone convergence theorem, we are done.

Example 3.107 (Random walk on trees: recurrence). To be written. See [Per99, Theorem 13.1].

\[\text{\footnote{Requires: Section 2.3.3.}}\]

We can now prove the following classical result.
Theorem 3.108 (Pólya’s theorem). Random walk on $\mathbb{L}^d$ is recurrent for $d \leq 2$ and transient for $d \geq 3$.

We prove the theorem for $d = 2, 3$ using the tools developed in this section. The other cases follow by Rayleigh’s principle. Of course, there are elementary proofs of this result. But we will show below that the electrical network approach has the advantage of being robust to the details of the lattice. For a different argument, see Exercise 2.7.

Proof of Theorem 3.108. The case $d = 2$ follows from the Nash-Williams inequality by letting $\Pi_j$ be the set of edges connecting vertices of $\ell^\infty$ norm $j$ and $j + 1$. (Recall that the $\ell^\infty$ norm is defined as $\|x\|_\infty = \max_j |x_j|$.) Using the fact that all conductances are 1, that $|\Pi_j| = O(j)$, and that $\sum_j j^{-1}$ diverges, recurrence is established.

Now consider the case $d = 3$ and let $a = 0$. We construct a finite-energy flow to $\infty$ using the method of random paths. Note that a simple way to produce a unit flow to $\infty$ is to push a flow of 1 through an infinite self-avoiding path. Taking this a step further, let $\mu$ be a probability measure on infinite self-avoiding paths and define the anti-symmetric function

$$\theta(x, y) := \mathbb{E}[1_{(x, y) \in \Gamma} - 1_{(y, x) \in \Gamma}] = \mathbb{P}(x, y) \in \Gamma) - \mathbb{P}(y, x) \in \Gamma),$$

where $\Gamma$ is a random path distributed according to $\mu$, oriented away from 0. Observe that $\sum_{y \sim x} [1_{(x, y) \in \Gamma} - 1_{(y, x) \in \Gamma}] = 0$ for any $x \neq 0$ because vertices visited by $\Gamma$ are entered and exited exactly once. That same sum is 1 at $x = 0$. Hence $\theta$ is a unit flow to $\infty$. Finally, for edge $e = \{x, y\}$, let

$$\mu(e) := \mathbb{P}(x, y) \in \Gamma \text{ or } (y, x) \in \Gamma) = \mathbb{P}(x, y) \in \Gamma) + \mathbb{P}(y, x) \in \Gamma) \geq \theta(x, y),$$

where we used that a self-avoiding path $\Gamma$ cannot visit both $(x, y)$ and $(y, x)$. Thomson’s principle gives the following bound.

Claim 3.109 (Method of random paths).

$$\mathcal{R}(0 \leftrightarrow \infty) \leq \sum_e [\mu(e)]^2. \quad (3.45)$$

For a measure $\mu$ concentrated on a single path, the sum above is infinite. To obtain a useful bound, what we need is a large collection of spread out paths. On the lattice $\mathbb{L}^3$, we construct $\mu$ as follows. Let $U$ be a uniformly random point on the unit sphere in $\mathbb{R}^3$ and let $\gamma$ be the ray from 0 to $\infty$ going through $U$. Imagine centering a unit cube around each point in $\mathbb{Z}^3$ whose edges are aligned with the...
axes. Then $\gamma$ traverses an infinite number of such cubes. Let $\Gamma$ be the corresponding self-avoiding path in the lattice $\mathbb{L}^3$. To see that this procedure indeed produces a path observe that $\gamma$, upon exiting a cube around a point $z \in \mathbb{Z}^3$, enters the cube of a neighboring point $z' \in \mathbb{Z}^3$ through a face corresponding to the edge between $z$ and $z'$ on the lattice $\mathbb{L}^3$ (unless it goes through a corner of the cube, but this has probability 0). To argue that $\mu$ distributes its mass among sufficiently spread out paths, we bound the probability that a vertex is visited by $\gamma$. Let $z$ be an arbitrary vertex in $\mathbb{Z}^3$. Because the sphere of radius $\|z\|_2$ around the origin in $\mathbb{R}^3$ has area $O(\|z\|_2^2)$ and its intersection with the unit cube centered around $z$ has area $O(1)$, it follows that

$$P[z \in \Gamma] = O\left(\frac{1}{\|z\|_2^2}\right).$$

That immediately implies a similar bound on the probability that an edge is visited by $\Gamma$. Moreover:

**Lemma 3.110.** There are $O(j^2)$ edges with an endpoint at $\ell^2$ distance within $[j, j + 1]$ from the origin.

**Proof.** Consider a ball of $\ell^2$ radius $1/2$ centered around each vertex of $\ell^2$ norm within $[j, j + 1]$. These balls are non-intersecting and have total volume $\Omega(N_j)$ where $N_j$ is the number of such vertices. On the other hand, the volume of the shell of $\ell^2$ inner and outer radii $j - 1/2$ and $j + 3/2$ centered around the origin is

$$\frac{4}{3} \pi (j + 3/2)^3 - \frac{4}{3} \pi (j - 1/2)^3 = O(j^2),$$

hence $N_j = O(j^2)$. Finally note that each vertex has 6 incident edges.

Plugging these bounds into (3.45), we get

$$\mathcal{R}(0 \leftrightarrow \infty) \leq \sum_j O(j^2) \cdot [O(1/j^2)]^2 = O\left(\sum_j j^{-2}\right) < +\infty.$$  

Transience follows from Theorem 3.106. (This argument clearly does not work on $\mathbb{L}$ where there are only two rays. You should convince yourself that it does not work on $\mathbb{L}^2$ either. But see Exercise 3.7.)

Finally we derive a useful general result illustrating the robustness reaped from Thomson’s principle. At a high level, a rough embedding from $\mathcal{N}$ to $\mathcal{N}'$ is a mapping of the edges of $\mathcal{N}$ to paths of $\mathcal{N}'$ of comparable overall resistance that do not overlap much. The formal definition follows. As we will see, the purpose of a rough embedding is to allow a flow on $\mathcal{N}$ to be morphed into a flow on $\mathcal{N}'$ of comparable energy.
Definition 3.111 (Rough embedding). Let \( \mathcal{N} = (G, c) \) and \( \mathcal{N}' = (G', c') \) be networks with resistances \( r \) and \( r' \) respectively. We say that a map \( \phi \) from the vertices of \( G \) to the vertices of \( G' \) is a rough embedding if there are constants \( \alpha, \beta < +\infty \) and a map \( \Phi \) defined on the edges of \( G \) such that:

1. for every edge \( e = \{x, y\} \) in \( G \), \( \Phi(e) \) is a non-empty, self-avoiding path of edges of \( G' \) between \( \phi(x) \) and \( \phi(y) \) such that

\[
\sum_{e' \in \Phi(e)} r'(e') \leq \alpha r(e),
\]

2. for every edge \( e' \) in \( G' \), there are no more than \( \beta \) edges in \( G \) whose image under \( \Phi \) contains \( e' \).

(The map \( \phi \) need not be a bijection.)

We say that two networks are roughly equivalent if there exist rough embeddings between them, one in each direction.

Example 3.112 (Independent coordinates walk). Let \( \mathcal{N} = \mathbb{L}^d \) with unit resistances and let \( \mathcal{N}' \) be the network on the subset of \( \mathbb{Z}^d \) corresponding to \( (Y_{t}^{(1)}, \ldots, Y_{t}^{(d)}) \), where the \( (Y_{t}^{(i)}) \)s are independent simple random walks on \( \mathbb{Z} \) started at 0. Note that \( \mathcal{N}' \) contains only those points of \( \mathbb{Z}^d \) with coordinates of identical parities. We claim that the networks \( \mathcal{N} \) and \( \mathcal{N}' \) are roughly equivalent.

- \( \mathcal{N} \) to \( \mathcal{N}' \): Consider the map \( \phi \) which associates to each \( x \in \mathcal{N} \) a closest point in \( \mathcal{N}' \) chosen in some arbitrary manner. For \( \Phi \), associate to each edge \( e = \{x, y\} \in \mathcal{N} \) a shortest path in \( \mathcal{N}' \) between \( \phi(x) \) and \( \phi(y) \), again chosen arbitrarily. If \( \phi(x) = \phi(y) \), choose an arbitrary, non-empty, shortest cycle through \( \phi(x) \).

- \( \mathcal{N}' \) to \( \mathcal{N} \): Consider the map \( \phi \) which associates to each \( x \in \mathcal{N}' \) the corresponding point \( x \) in \( \mathcal{N} \). Construct \( \Phi \) similarly to the previous case.

See Exercise 3.9 for an important generalization of the previous example. Our main result about roughly equivalent networks is that they have the same type.

Theorem 3.113 (Recurrence and rough equivalence). Let \( \mathcal{N} \) and \( \mathcal{N}' \) be roughly equivalent, locally finite, connected networks. Then \( \mathcal{N} \) is transient if and only if \( \mathcal{N}' \) is transient.
Proof. Assume $\mathcal{N}$ is transient and let $\theta$ be a unit flow from some $a$ to $\infty$ of finite energy. The existence of this flow is guaranteed by Theorem 3.106. Let $\phi, \Phi$ be a rough embedding with parameters $\alpha$ and $\beta$.

The basic idea of the proof is to map the flow $\theta$ onto $\mathcal{N}'$ using $\Phi$. Because flows are directional, it will be convenient to think of edges as being directed. Recall that $\langle x, y \rangle$ denotes the directed edge from $x$ to $y$. For $e = \{x, y\}$ in $\mathcal{N}$, we write $\langle x', y' \rangle \in \overrightarrow{\Phi}(x, y)$ to mean that $\{x', y'\} \in \Phi(e)$ and that $x'$ is visited before $y'$ in the path $\Phi(e)$ from $\phi(x)$ to $\phi(y)$. (If $\phi(x) = \phi(y)$, choose an arbitrary orientation of the cycle $\Phi(e)$ for $\overrightarrow{\Phi}(x, y)$ and the reversed orientation for $\overleftarrow{\Phi}(y, x)$.) Then define, for $x', y'$ with $\{x', y'\}$ in $\mathcal{N}'$,

$$\theta'(x', y') := \sum_{\langle x, y \rangle : \langle x', y' \rangle \in \overrightarrow{\Phi}(x, y)} \theta(x, y). \quad (3.46)$$

See Figure 3.5.

We claim that $\theta'$ is a flow to $\infty$ of finite energy on $\mathcal{N}'$. We first check that $\theta'$ is a flow.

1. (Anti-symmetry) By construction, $\theta'(y', x') = -\theta'(x', y')$, i.e., $\theta'$ is antisymmetric, because $\theta$ itself is anti-symmetric.
2. (Flow conservation) Next we check the flow-conservation constraints. Fix $z'$ in $N'$. By Condition 2 in Definition 3.111, there are finitely many edges $e$ in $N$ such $\Phi(e)$ visits $z'$. Let $e = \{x, y\}$ be such an edge. There are two cases:

- Assume first that $\phi(x), \phi(y) \neq z'$ and let $\langle u', z' \rangle, \langle z', w' \rangle$ be the directed edges incident with $z'$ on the path $\Phi(e)$ oriented from $\phi(x)$ to $\phi(y)$. Observe that, in the definition of $\theta'$, $\langle y, x \rangle$ contributes $\theta(y, x) = -\theta(x, y)$ to $\theta'(z', u')$ and $\langle x, y \rangle$ contributes $\theta(x, y)$ to $\theta'(z', w')$. So these contributions cancel out in the flow-conservation constraint for $z'$, i.e., in the sum $\sum_{v' \sim z'} \theta'(z', v')$.

- If instead $e = \{x, y\}$ is such that $\phi(x) = z'$, let $\langle z', w' \rangle$ be the first edge on the path $\Phi(e)$ from $\phi(x)$ to $\phi(y)$. Edge $\langle x, y \rangle$ contributes $\theta(x, y)$ to $\theta'(z', w')$. (A similar statement applies to $\phi(y) = z'$.)

From the two cases above, summing over all paths visiting $z'$ gives

$$\sum_{v' \sim z'} \theta'(z', v') = \sum_{z: \phi(z) = z'} \left( \sum_{v' \sim z} \theta(z, v) \right).$$

Because $\theta$ is a flow, the sum on the r.h.s. is 0 unless $a \in \phi^{-1}(\{z'\})$ in which case it is 1. We have shown that $\theta'$ is a unit flow from $\phi(a)$ to $\infty$.

It remains to bound the energy of $\theta'$. By (3.46), Cauchy-Schwarz, and Condition 2 in Definition 3.111,

$$\theta'(x', y')^2 = \left[ \sum_{\langle x, y \rangle: \langle x', y' \rangle \in \Phi(x, y)} \theta(x, y) \right]^2 \leq \left[ \sum_{\langle x, y \rangle: \langle x', y' \rangle \in \Phi(x, y)} 1 \right] \left[ \sum_{\langle x, y \rangle: \langle x', y' \rangle \in \Phi(x, y)} \theta(x, y)^2 \right] \leq \beta \sum_{\langle x, y \rangle: \langle x', y' \rangle \in \Phi(x, y)} \theta(x, y)^2.$$
Summing over all pairs and using Condition 1 in Definition 3.111 gives
\[ \frac{1}{2} \sum_{x',y'} r'(x',y') \theta'(x',y')^2 \leq \beta \frac{1}{2} \sum_{x',y'} r'(x',y') \sum_{(x,y):(x',y') \in \Phi(x,y)} \theta(x,y)^2 \]
\[ = \beta \frac{1}{2} \sum_{x,y} \theta(x,y)^2 \sum_{(x',y') \in \Phi(x,y)} r'(x',y') \]
\[ \leq \alpha \beta \frac{1}{2} \sum_{x,y} r(x,y) \theta(x,y)^2, \]
which is finite by assumption. That concludes the proof.

**Example 3.114** (Independent coordinates walk (continued)). Consider again the networks \( \mathcal{N} \) and \( \mathcal{N}' \) in Example 3.112. Because they are roughly equivalent, they have the same type. This leads to yet another proof of Pólya’s theorem. Recall that, because the number of returns to 0 is geometric with success probability equal to the escape probability, random walk on \( \mathcal{N}' \) is recurrent if and only if the expected number of visits to 0 is finite. By independence of the coordinates, this expectation can be written as
\[ \sum_{t \geq 0} \left( \mathbb{P}_0 \left[ Y_{2t}^{(1)} = 0 \right] \right)^d = \sum_{t \geq 0} \left( \frac{2t}{t} \right)^d = \sum_{t \geq 0} \Theta(t^{-d/2}), \]
where we used Stirling’s formula. The r.h.s. is finite if and only if \( d \geq 3 \). That implies random walk on \( \mathcal{N}' \) is transient under the same condition. By rough equivalence, the same is true of \( \mathcal{N} \).

**Other applications** So far we have emphasized applications to recurrence. Here we show that electrical network theory can also be used to bound certain hitting times. In Sections 3.3.5 and 3.3.6, we give further applications beyond random walks on graphs.

An application of Lemma 3.24 gives another probabilistic interpretation of the effective resistance—and a useful formula.

**Theorem 3.115** (Commute time identity). Let \( \mathcal{N} = (G,c) \) be a finite, connected network with vertex set \( V \). For \( x \neq y \), let the commute time \( \tau_{x,y} \) be the time of the first return to \( x \) after the first visit to \( y \). Then
\[ \mathbb{E}_x[\tau_{x,y}] = \mathbb{E}_x[\tau_y] + \mathbb{E}_y[\tau_x] = c_{\mathcal{N}} R(x \leftrightarrow y), \]
where \( c_{\mathcal{N}} = 2 \sum_{e=(x,y) \in \mathcal{N}} c(e) \).
Proof. This follows immediately from Lemma 3.24 and the definition of the effective resistance. Specifically,

\[
\mathbb{E}_x[\tau_y] + \mathbb{E}_y[\tau_x] = \frac{1}{\pi_x(\mathbb{P}_x[\tau_y < \tau_x^+])} = \frac{1}{\{e(x)/(2\sum_{e=(x,y)\in\mathcal{N}} c(e))\} \mathbb{P}_x[\tau_y < \tau_x^+]} = c_N \mathcal{R}(x \leftrightarrow y).
\]

Example 3.116 (Random walk on the torus). Consider random walk on the \(d\)-dimensional torus \(L_d^n\) with unit resistances. We use the commute time identity to lower bound the mean hitting time \(\mathbb{E}_x[\tau_y]\) for arbitrary vertices \(x \neq y\) at graph distance \(k\) on \(L_d^n\). To use Theorem 3.115, note that by symmetry \(\mathbb{E}_x[\tau_y] = \mathbb{E}_y[\tau_x]\) so that

\[
\mathbb{E}_x[\tau_y] = \frac{1}{2} c_N \mathcal{R}(x \leftrightarrow y) = n^d \mathcal{R}(x \leftrightarrow y). \tag{3.47}
\]

To simplify, assume \(n\) is odd and identify the vertices of \(L_d^n\) with the box

\[
B := \{-(n-1)/2, \ldots, (n-1)/2\}^d,
\]

in \(\mathbb{L}^d\) centered at \(x = 0\). The rest of the argument is essentially identical to the first half of the proof of Theorem 3.108. Let \(\partial B_j^\infty = \{z \in \mathbb{L}^d : ||z||_\infty = j\}\) and let \(\Pi_j\) be the set of edges between \(\partial B_j^\infty\) and \(\partial B_{j+1}^\infty\). Note that on \(B\) the \(\ell_1\) norm of \(y\) is at most \(k\). Since the \(\ell_\infty\) norm is at least \(1/d\) times the \(\ell_1\) norm on \(\mathbb{L}^d\), there exists \(J = O(k)\) such that all \(\Pi_j\)s, \(j \leq J\), are cutsets separating \(x\) from \(y\). By the Nash-Williams inequality

\[
\mathcal{R}(x \leftrightarrow y) \geq \sum_{0 \leq j \leq J} |\Pi_j|^{-1} = \sum_{0 \leq j \leq J} \Omega\left(j^{-d(1-\delta)}\right) = \begin{cases} \Omega(\log k), & d = 2 \\ \Omega(1), & d \geq 3. \end{cases}
\]

From (3.47), we get:

Claim 3.117.

\[
\mathbb{E}_x[\tau_y] = \begin{cases} \Omega(n^d \log k), & d = 2 \\ \Omega(n^d), & d \geq 3. \end{cases}
\]

Remark 3.118. The bounds in the previous example are tight up to constants. See [LPW06, Proposition 10.13].

174
3.3.4 Random walk on supercritical percolation clusters

In this section, we apply the random paths approach to random walk on percolation clusters.§

To be written. See [LP, Section 5.5].

3.3.5 Uniform spanning trees: Wilson’s method

In this section, we describe an application of electrical network theory to uniform spanning trees.

Uniform spanning trees Let $G = (V, E)$ be a finite connected graph. Recall that a spanning tree is a subtree of $G$ containing all its vertices. A uniform spanning tree is a spanning tree $T$ chosen uniformly at random among all spanning trees of $G$. (The reader interested only in Wilson’s method for generating uniform spanning trees may jump ahead to the second half of this section.)

A fundamental property of uniform spanning trees is the following negative correlation between edges.

Claim 3.119.

$$
\mathbb{P}[e \in T | e' \in T] \leq \mathbb{P}[e \in T], \quad \forall e \neq e' \in G.
$$

This property is perhaps not surprising. For one, the number of edges in a spanning tree is fixed, so the inclusion of $e'$ makes it seemingly less likely for other edges to be present. Yet proving Claim 3.119 is non-trivial. The only known proof relies on the electrical network perspective developed in this section. The key to the proof is a remarkable formula for the inclusion of an edge in a uniform spanning tree.

Theorem 3.120 (Kirchhoff’s resistance formula). Let $G = (V, E)$ be a finite, connected graph and let $\mathcal{N}$ be the network on $G$ with unit resistances. If $T$ is a uniform spanning tree on $G$, then for all $e = \{x, y\}$

$$
\mathbb{P}[e \in T] = R(x \leftrightarrow y).
$$

Before explaining how this formula arises, we show that it implies Claim 3.119.

Proof of Claim 3.119. By Bayes’ rule and a short calculation, we can instead prove

$$
\mathbb{P}[e \in T | e' \notin T] \geq \mathbb{P}[e \in T], \quad (3.48)
$$

\footnote{Requires: Section 2.3.}

175
unless \( \mathbb{P}[e' \in T] \in \{0, 1\} \) or \( \mathbb{P}[e \in T] \in \{0, 1\} \) in which case the claim is vacuous. (In fact these probabilities cannot be 0. Why? Can they be equal to 1?) Picking a uniform spanning tree on \( \mathcal{N} \) conditioned on \( \{e' \notin T\} \) is the same as picking a uniform spanning tree on the modified network \( \mathcal{N}' \) where \( e' \) is removed. By Rayleigh’s principle,

\[
\mathcal{R}_{N'}(x \leftrightarrow y) \geq \mathcal{R}_N(x \leftrightarrow y),
\]

and Kirchhoff’s resistance formula gives (3.48).

**Remark 3.121.** More generally, thinking of a uniform spanning tree \( T \) as a random subset of edges, the law of \( T \) has the property of negative associations, defined as follows. An event \( A \subseteq 2^E \) is said to be increasing if \( \omega \cup \{e\} \in A \) whenever \( \omega \in A \). The event \( A \) is said to depend only on \( F \subseteq E \) if for all \( \omega_1, \omega_2 \in 2^E \) that agree on \( F \), either both are in \( A \) or neither is. The law, \( \mathbb{P}_T \), of \( T \) has negative associations in the sense that for any two increasing events \( A \) and \( B \) that depend only on disjoint sets of edges, we have \( \mathbb{P}_T[A \cap B] \leq \mathbb{P}_T[A] \mathbb{P}_T[B] \), i.e., \( A \) and \( B \) are negatively correlated. See [LP, Exercise 4.6]. (To see why the events considered depend on disjoint edges, see for instance what happens when \( A \subseteq B \).)

Let \( e = \{x, y\} \). To get some insight into Kirchhoff’s resistance formula, we first note that, if \( i \) is the unit current from \( x \) to \( y \) and \( v \) is the associated voltage, by definition of the effective resistance

\[
\mathcal{R}(x \leftrightarrow y) = \frac{v(x)}{\|i\|} = c(e)(v(x) - v(y)) = i(x, y),
\]

(3.49)

where we used Ohm’s law as well as the fact that \( c(e) = 1 \), \( v(y) = 0 \), and \( \|i\| = 1 \). Note the difference between \( \|i\| \) and \( i(x, y) \). Although \( \|i\| = 1 \), \( i(x, y) \) is only the current along the edge to \( y \). Furthermore by the probabilistic interpretation of the current, with \( Z = \{y\} \),

\[
i(x, y) = \mathbb{E}_x[N_{x \rightarrow y}^Z - N_{y \leftarrow x}^Z] = \mathbb{P}_x[(x, y) \text{ is traversed before } \tau_y].
\]

(3.50)

Indeed, started at \( x \), \( N_{y \rightarrow x}^Z = 0 \) and \( N_{x \rightarrow y}^Z \in \{0, 1\} \). Kirchhoff’s resistance formula is then established by relating the random walk on \( \mathcal{N} \) to the probability that \( e \) is present in a uniform spanning tree \( T \). To do this we introduce a random-walk-based algorithm for generating uniform spanning trees. This rather miraculous procedure, known as **Wilson’s method**, is of independent interest. For a classical connection between random walks and spanning trees, see Exercise 3.11.

**Wilson’s method** It will be somewhat more transparent to work in a more general context. Let \( \mathcal{N} = (G, c) \) be a finite, connected network on \( G \) with arbitrary
conductances and define the **weight** of a spanning tree $T$ on $\mathcal{N}$ as

$$W(T) = \prod_{e \in T} c(e).$$

With a slight abuse, we continue to call a tree $T$ picked at random among all spanning trees of $G$ with probability proportional to $W(T)$ a “uniform” spanning tree on $\mathcal{N}$.

To state Wilson’s method, we need the notion of **loop erasure**. Let $P = x_0 \sim \cdots \sim x_k$ be a path in $\mathcal{N}$. The loop erasure of $P$ is obtained by removing cycles in the order they appear. That is, let $j^*$ be the smallest $j$ such that $x_j = x_{\ell}$ for some $\ell < j$. Remove the subpath $x_{\ell+1} \sim \cdots \sim x_j$ from $P$, and repeat. The resulting self-avoiding path is denoted by $\text{LE}(P)$.

Let $\rho$ be an arbitrary vertex of $G$, which refer to as the root, and let $T_0$ be the subtree made up of $\rho$ alone. Order arbitrarily the vertices $v_0, \ldots, v_{n-1}$ of $G$, starting with the root $v_0 := \rho$. Wilson’s method constructs an increasing sequence of subtrees as follows. See Figure 3.6. Let $T := T_0$.

1. Let $v$ be the vertex of $G$ not in $T$ with lowest index. Perform random walk on $\mathcal{N}$ started at $v$ until the first visit to a vertex of $T$. Let $P$ be the resulting path.
2. Add the loop erasure \( \text{LE}(P) \) to \( T \).

3. Repeat until all vertices of \( G \) are in \( T \).

Let \( T_0, \ldots, T_m \) be the sequence of subtrees produced by Wilson’s method.

**Claim 3.122.** Forgetting the root, \( T_m \) is a uniform spanning tree on \( \mathcal{N} \).

This claim is far from obvious. Before proving it, we finish the proof of Kirchhoff’s resistance formula.

**Proof of Theorem 3.120.** From (3.49) and (3.50), it suffices to prove that, for \( e = \{x, y\} \),

\[
\mathbb{P}_x \left[ \langle x, y \rangle \text{ is traversed before } \tau_y \right] = \mathbb{P}[e \in T],
\]

where the probability on the l.h.s. refers to random walk on \( \mathcal{N} \) started at \( x \) and the probability on the r.h.s. refers to a uniform spanning tree \( T \) on \( \mathcal{N} \). Generate \( T \) using Wilson’s method started at root \( \rho = y \) with the choice \( v_1 = x \). If the sample path from \( x \) to \( y \) during the first iteration of Wilson’s method includes \( \langle x, y \rangle \), then the loop erasure is simply \( x \sim y \) and \( e \) is in \( T \). On the other hand, if the sample path from \( x \) to \( y \) does not include \( \langle x, y \rangle \), then \( e \) cannot be used at a later stage because it would create a cycle. That immediately proves the claim.

It remains to prove the claim.

**Proof of Claim 3.122.** The idea of the proof is to cast Wilson’s method in the more general framework of cycle popping algorithms. We begin by explaining how such algorithms work.

Let \( P \) be the transition matrix corresponding to random walk on \( \mathcal{N} = (G, c) \) with \( G = (V, E) \). To each vertex \( x \neq \rho \) in \( V \), we assign an independent stack of directed edges

\[
S^x_0 := (\langle x, Y^x_1 \rangle, \langle x, Y^x_2 \rangle, \ldots)
\]

where each \( Y^x_j \) is chosen independently at random from the distribution \( P(x, \cdot) \). In particular all \( Y^x_j \)'s are neighbors of \( x \) in \( \mathcal{N} \). The index \( j \) in \( \langle x, Y^x_j \rangle \) is usually referred to as the color of the edge. It keeps track of the position of the edge in the original stack. (Picture \( S^x \) as a spring-loaded plate dispenser located on vertex \( x \).)

We consider a process which involves popping edges off the stacks. We use the notation \( S^x \) to denote the current stack at \( x \). The initial assignment of the stack is \( S^x := S^x_0 \) as above. Given the current stacks \( (S^x) \), we call visible graph the directed graph over \( V \) with edges \( \text{Top}(S^x) \) for all \( x \neq \rho \), where \( \text{Top}(S^x) \) is the first edge in the current stack \( S^x \). The latter are referred to as visible edges. We denote the current visible graph by \( \overrightarrow{G} \). Note that \( \overrightarrow{G} \) has out-degree 1 for
all $x \neq \rho$ and the root has out-degree 0. In particular all (undirected) cycles in $\overrightarrow{G}_0$ are in fact directed cycles. Indeed, a set of edges forming a cycle that is not directed must have a vertex of out-degree 2. Recall the following characterization (see Lemma A.8): a cycle-free undirected graph with $n$ vertices and $n - 1$ edges is a spanning tree. Hence, if there is no cycle in $\overrightarrow{G}_0$ then it must be a spanning tree where, furthermore, all edges point towards the root. Such a tree is also known as a spanning arborescence.

As the name suggests, a cycle popping algorithm proceeds by popping cycles in $\overrightarrow{G}_0$ off the tops of the stacks until a spanning arborescence is produced. That is, at every iteration, if $\overrightarrow{G}_0$ contains at least one cycle, then a cycle $\overrightarrow{C}$ is picked according to some rule, the top of each stack in $\overrightarrow{C}$ is popped, and a new visible graph $\overrightarrow{G}_0$ is revealed. See Figure 3.7 for an illustration.

With these definitions in place, the proof of the claim involves the following steps.

1. **Wilson’s method is a cycle popping algorithm.** We can think of the initial stacks $(S_0^x)$ as corresponding to picking—ahead of time—all potential transitions in the random walks used by Wilson’s method. With this representation, Wilson’s method boils down to a recipe for choosing which cycle to pop next. Indeed, at each iteration, we start from a vertex $v$ not in the current tree $T$. Following the visible edges from $v$ traces a path whose distribution is that of random walk on $\mathcal{N}$. Loop erasure then corresponds to popping cycles. We pop only those visible edges on the removed cycles as they originate from vertices that will be visited again by the algorithm and for which a new transition will then be needed. Those visible edges in the remaining loop-erased path are not popped—they are part of the final arborescence.

2. **The popping order does not matter.** We just argued that Wilson’s method is a cycle popping algorithm. In fact we claim that any cycle popping algorithm, i.e., no matter what popping choices are made along the way, produces the same final arborescence. To make this precise, we identify the popped cycles uniquely. This is where the colors come in. A colored cycle is a directed cycle over $\overrightarrow{V}$ made of colored edges from the stacks (not necessarily of the same color and not necessarily in the current visible graph). We say that a colored cycle $\overrightarrow{C}$ is poppable for a visible graph $\overrightarrow{G}_0$ if there exists a sequence of colored cycles $\overrightarrow{C}_1, \ldots, \overrightarrow{C}_r = \overrightarrow{C}$ that can be popped in that order starting from $\overrightarrow{G}_0$. Note that, by this definition, $\overrightarrow{C}_1$ is a directed cycle in $\overrightarrow{G}_0$. Now we claim that if $\overrightarrow{C}_1'$ were popped first instead of $\overrightarrow{C}_1$, producing the new visible graph $\overrightarrow{G}_0'$, then $\overrightarrow{C}$ would still be poppable for $\overrightarrow{G}_0'$. This claim implies
Figure 3.7: A realization of a cycle popping algorithm (from top to bottom). In all three figures, the underlying graph is $G$ while the arrows depict the visible edges.
that, in any cycle popping algorithm, either an infinite number of cycles are popped or eventually all poppable cycles are popped—indeed of the order—producing the same outcome. To prove the claim, note first that if \( \overrightarrow{C}_1' = \overrightarrow{C}_1 \) or if \( \overrightarrow{C}_1' \) does not share a vertex with any of \( \overrightarrow{C}_1, \ldots, \overrightarrow{C}_r \), there is nothing to prove. So let \( \overrightarrow{C}_j \) be the first cycle in the sequence sharing a vertex with \( \overrightarrow{C}_1 \), say \( x \). Let \( \langle x, y \rangle_c \) and \( \langle x, y' \rangle_{c'} \) be the colored edges emanating from \( x \) in \( \overrightarrow{C}_j \) and \( \overrightarrow{C}_1 \) respectively. By definition, \( x \) is not on any of \( \overrightarrow{C}_1, \ldots, \overrightarrow{C}_{j-1} \) so the edge originating from \( x \) is not popped by that sequence and we must have \( \langle x, y \rangle_c = \langle x, y' \rangle_{c'} \) as colored edges. In particular, the vertex \( y \) is also a shared vertex of \( \overrightarrow{C}_j \) and \( \overrightarrow{C}_1 \), and the same argument applies to it. Proceeding by induction leads to the conclusion that \( \overrightarrow{C}_1' = \overrightarrow{C}_j \) as colored cycles. But then \( \overrightarrow{C} \) is clearly poppable for the visible graph resulting from popping \( \overrightarrow{C}_1' \) first, because it can be popped with the rearranged sequence \( \overrightarrow{C}_1' = \overrightarrow{C}_j, \overrightarrow{C}_1, \ldots, \overrightarrow{C}_{j-1}, \overrightarrow{C}_{j+1}, \ldots, \overrightarrow{C}_r = \overrightarrow{C} \), where we used the fact that \( \overrightarrow{C}_1' \) does not share a vertex with \( \overrightarrow{C}_1, \ldots, \overrightarrow{C}_{j-1} \).

3. **Termination occurs in finite time almost surely.** We have shown so far that, in any cycle popping algorithm, either an infinite number of cycles are popped or eventually all poppable cycles are popped. But Wilson’s method—a cycle popping algorithm as we have shown—stops after a finite amount of time with probability 1. Indeed, because the network is finite and connected, the random walk started at each iteration hits the current \( T \) in finite time almost surely (by Lemma 3.25). To sum up, all cycle popping algorithms terminate and produce the same spanning arborescence. It remains to compute the distribution of the outcome.

4. **The arborescence has the desired distribution.** Let \( \mathcal{A} \) be the spanning arborescence produced by any cycle popping algorithm on the stacks \( (S_0^x) \). To compute the distribution of \( \mathcal{A} \), we first compute the distribution of a particular cycle popping realization leading to \( \mathcal{A} \). Because the popping order does not matter, by “realization” we mean a collection \( \mathcal{C} \) of colored cycles together with a final spanning arborescence \( \mathcal{A} \). Notice that what lies in the stacks under \( \mathcal{A} \) is not relevant to the realization, i.e., the same outcome is produced no matter what is under \( \mathcal{A} \). So, from the distribution of the stacks, the probability of observing \( (\mathcal{C}, \mathcal{A}) \) is simply the product of the transitions corresponding to the directed edges in \( \mathcal{C} \) and \( \mathcal{A} \), i.e.,

\[
\prod_{\overrightarrow{e} \in \mathcal{C} \cup \mathcal{A}} P(\overrightarrow{e}) = \Psi(\mathcal{A}) \prod_{\overrightarrow{c} \in \mathcal{C}} \Psi(\overrightarrow{c}),
\]

181
where the function $\Psi$ returns the product of the transition probabilities of a set of directed edges. Thanks to the product form on the r.h.s., summing over all possible $C$s gives that the probability of producing $A$ is proportional to $\Psi(A)$. For this argument to work though, there are two small details to take care of. First, note that we want the probability of the uncolored arborescence. But observe that, in fact, there is no need to keep track of the colors on the edges of $A$ because these are determined by $C$. Secondly, we need for the collection of possible $C$s not to vary with $A$. But it is clear that any arborescence could lie under any $C$.

To see that we are done, let $T$ be the undirected spanning tree corresponding to the outcome, $A$, of Wilson’s method. Then, because $P(x, y) = \frac{e^{c(x,y)}}{e^{c(x)}}$, we get

$$\Psi(A) = \frac{W(T)}{\prod_{x \neq \rho} c(x)},$$

where note that the denominator does not depend on $T$. So if we forget the orientation of $A$, which is determined by the root, we get a spanning tree whose distribution is proportional to $W(T)$, as required.

### 3.3.6 Ising model on trees: the reconstruction problem

To be written. See [Per99, Section 16].

### Exercises

**Exercise 3.1** (Azuma-Hoeffding: a second proof). This exercise leads the reader through an alternative proof of the Azuma-Hoeffding inequality.

- a) Show that for all $x \in [-1, 1]$ and $a > 0$
  $$e^{ax} \leq \cosh a + x \sinh a.$$

- b) Use a Taylor expansion to show that for all $x$
  $$\cosh x \leq e^{x^2/2}.$$

- c) Let $X_1, \ldots, X_n$ be (not necessarily independent) random variables such that, for all $i$, $|X_i| \leq c_i$ for some constant $c_i < +\infty$ and
  $$\mathbb{E}[X_{i_1} \cdots X_{i_k}] = 0, \quad \forall 1 \leq k \leq n, \forall 1 \leq i_1 < \cdots < i_k \leq n. \quad (3.51)$$
Show, using a) and b), that for all $b > 0$

$$\Pr \left[ \sum_{i=1}^{n} X_i \geq b \right] \leq \exp \left( -\frac{b^2}{2 \sum_{i=1}^{n} c_i^2} \right).$$

d) Prove that c) implies the Azuma-Hoeffding inequality as stated in Theorem 3.52.

e) Show that the random variables in Exercise 2.5 do not satisfy (3.51) (without using the claim in part b) of that exercise).

**Exercise 3.2** (Kirchhoff’s laws). Consider a finite, connected network with a source and a sink. Show that an anti-symmetric function on the edges satisfying Kirchhoff’s two laws is a current function (i.e., it corresponds to a voltage function through Ohm’s law).

**Exercise 3.3** (Dirichlet problem: non-uniqueness). Let $(X_t)$ be the birth-and-death chain on $\mathbb{Z}_+$ with $P(x, x + 1) = p$ and $P(x, x - 1) = 1 - p$ for all $x \geq 1$, and $P(0, 1) = 1$, for some $0 < p < 1$. Fix $h(0) = 1$.

a) When $p > 1/2$, show that there is more than one bounded extension of $h$ to $\mathbb{Z}_+\setminus\{0\}$ that is harmonic on $\mathbb{Z}_+\setminus\{0\}$. [Hint: Consider $P_x[\tau_0 = +\infty].$]

b) When $p \leq 1/2$, show that there exists a unique bounded extension of $h$ to $\mathbb{Z}_+\setminus\{0\}$ that is harmonic on $\mathbb{Z}_+\setminus\{0\}$.

**Exercise 3.4** (Maximum principle). Let $\mathcal{N} = (G, c)$ be a finite or countable, connected network with $G = (V, E)$. Let $W$ be a finite, connected, proper subset of $V$.

a) Let $h : V \to \mathbb{R}$ be a function on $V$. Prove the maximum principle: if $h$ is harmonic on $W$, i.e., it satisfies

$$h(x) = \frac{1}{c(x)} \sum_{y \sim x} c(x, y) h(y), \quad \forall x \in W,$$

and if $h$ achieves its supremum on $W$, then $h$ is constant on $W \cup \partial_V W$, where

$$\partial_V W = \{ z \in V \setminus W : \exists y \in W, y \sim z \}.$$

b) Let $h : W^c \to \mathbb{R}$ be a bounded function on $W^c := V \setminus W$. Let $h_1$ and $h_2$ be extensions of $h$ to $W$ that are harmonic on $W$. Use part a) to prove that $h_1 \equiv h_2$. 

183
Exercise 3.5 (Effective resistance: metric). Show that effective resistances between pairs of vertices form a metric.

Exercise 3.6 (Dirichlet principle: proof). Prove Theorem 3.100.

Exercise 3.7 (Random walk on $\mathbb{L}^2$: effective resistance). Consider random walk on $\mathbb{L}^2$, which we showed is recurrent. Let $(G_n)$ be the exhaustive sequence corresponding to vertices at distance at most $n$ from the origin and let $Z_n$ be the corresponding sink-set. Show that $\mathcal{R}(0 \leftrightarrow Z_n) = \Theta(\log n)$. [Hint: Use the Nash-Williams inequality and the method of random paths.]

Exercise 3.8 (Random walk on regular graphs: effective resistance). Let $G$ be a $d$-regular graph with $n$ vertices and $d > n/2$. Let $\mathcal{N}$ be the network $(G, c)$ with unit conductances. Let $a$ and $z$ be arbitrary distinct vertices.

a) Show that there are at least $2d - n$ vertices $x \neq a, z$ such that $a \sim x \sim z$ is a path.

b) Prove that

$$\mathcal{R}(a \leftrightarrow z) \leq \frac{2rn}{2r - n}.$$ 

Exercise 3.9 (Rough isometries). Graphs $G = (V, E)$ and $G' = (V', E')$ are roughly isometric (or quasi-isometric) if there is a map $\phi : V \to V'$ and constants $0 < \alpha, \beta < +\infty$ such that for all $x, y \in V$

$$\alpha^{-1}d(x, y) - \beta \leq d'(\phi(x), \phi(y)) \leq \alpha d(x, y) + \beta,$$

where $d$ and $d'$ are the graph distances on $G$ and $G'$ respectively, and furthermore all vertices in $G'$ are within distance $\beta$ of the image of $V$. Let $\mathcal{N} = (G, c)$ and $\mathcal{N}' = (G', c')$ be countable, connected networks with uniformly bounded conductances, resistances and degrees. Prove that if $G$ and $G'$ are roughly isometric then $\mathcal{N}$ and $\mathcal{N}'$ are roughly equivalent. [Hint: Start by proving that being roughly isometric is an equivalence relation.]

Exercise 3.10 (Random walk on the cycle: hitting time). Use the commute time identity to compute $E_x[\tau_y]$ in Example 3.116 in the case $d = 1$. Give a second proof using a direct martingale argument.

Exercise 3.11 (Markov chain tree theorem). Let $P$ be the transition matrix of a finite, irreducible Markov chain with stationary distribution $\pi$. Let $G$ be the directed graph corresponding to the positive transitions of $P$. For an arborescence $A$ of $G$, define its weight as

$$\Psi(A) = \prod_{\bar{e} \in A} P(\bar{e}).$$
Consider the following process on spanning arborescences over $G$. Let $\rho$ be the root of the current spanning arborescence $\mathcal{A}$. Pick an outgoing edge $\vec{e} = \langle \rho, x \rangle$ of $\rho$ according to $P(\rho, \cdot)$. Add $\vec{e}$ to $\mathcal{A}$. This creates a cycle. Remove the edge of this cycle originating from $x$, producing a new arborescence $\mathcal{A}'$ with root $x$. Repeat the process.

a) Show that this chain is irreducible.

b) Show that $\Psi$ is a stationary measure for this chain.

c) Prove the Markov chain tree theorem: The stationary distribution $\pi$ of $P$ is proportional to

$$\pi_x = \sum_{\mathcal{A} : \text{root}(\mathcal{A}) = x} \Psi(\mathcal{A}).$$

Bibliographic remarks

Section 3.1 The textbooks [Dur10] and [Wil91] contain excellent introductions to martingales. The upper bound in Theorem 3.28 was first proved by Matthews [Mat88]. Section 3.1.4 is based on [Per09, Sections 2 and 3].

Section 3.2 The Azuma-Hoeffding inequality is due to Hoeffding [Hoe63] and Azuma [Azu67]. The version of the inequality in Exercise 3.1 is from [Ste97]. The method of bounded differences has its origins in the works of Yurinskii [Yur76], Maurey [Mau79], Milman and Schechtman [MS86], Rhee and Talagrand [RT87], and Shamir and Spencer [SS87]. It was popularized by McDiarmid [McD89]. Example 3.63 is taken from [MU05, Section 12.5]. Claim 3.68 is due to Shamir and Spencer [SS87]. The 2-point concentration result alluded to in Section 3.2.3 is due to Alon and Krivelevich [AK97]. For the full story on the chromatic number of Erdős-Rényi graphs, see [JLR11, Chapter 7]. Claim 3.73 is due to Bollobás, Riordan, Spencer, and Tusnády [BRST01]. It confirmed simulations of Barabási and Albert [BA99]. The expectation was analyzed by Dorogovtsev, Mendes, and Samukhin [DMS00]. For much more on preferential attachment models see [Dur06] and [vdH14]. General references on the concentration of measure phenomenon and concentration inequalities are [Led01] and [BLM13]. See [BCB12] for an introduction to bandit problems. The slicing argument in Section 3.2.5 is based on [Bub10]. A more general discussion of the slicing method, whose best known application is the proof of the law of the iterated logarithm (e.g. [Wil91, Section 14.7]), can be found in [vH].
Section 3.3  The material in Sections 3.3.1-3.3.5 borrows heavily from [LPW06, Chapters 9, 10], [AF, Chapters 2, 3] and, especially, [LP, Sections 2.1-2.6, 4.1-4.2, 5.5]. The classical reference on potential theory and its probabilistic counterpart is [Doo01]. For the discrete case and the electrical network point of view, the book of Doyle and Snell is excellent [DS84]. In particular the series and parallel laws are defined and illustrated. See also [KSK76]. For an introduction to convex optimization and duality, see e.g. [BV04]. The Nash-Williams inequality is due to Nash-Williams [NW59]. The result in Example 3.107 is due to R. Lyons [Lyo90]. An elementary proof of Pólya’s theorem can be found in [Dur10, Section 4.2]. The flow we used in the proof of Pólya’s theorem is essentially due to T. Lyons [Lyo83]. Theorem 3.113 is due to Kanai [Kan86]. The commute time identity was proved by Chandra, Raghavan, Ruzzo, Smolensky and Tiwari [CRR+89]. Wilson’s method is due to Wilson [Wil96]. A related method for generating uniform spanning trees was introduced by Aldous [Ald90] and Broder [Bro89]. A connection between loop-erased random walks and uniform spanning trees has previously been established by Pemantle [Pem91] using the Aldous-Broder method. For more on negative correlation in uniform spanning trees, see e.g. [LP, Section 4.2]. For a proof of the matrix tree theorem using Wilson’s method, see [KRS]. For a discussion of the running time of Wilson’s method and other spanning tree generation approaches, see [Wil96].