Chapter 3

Martingales and potentials

In this chapter we turn to *martingales*, which play a central role in probability theory. We illustrate their use in a number of applications to the analysis of discrete stochastic processes. After some background on stopping times and a brief review of basic martingale properties and results in Section 3.1, we develop two major directions. In Section 3.2, we show how martingales can be used to derive a substantial generalization of our previous concentration inequalities—from the *sums* of independent random variables we focused on in Chapter 2 to *nonlinear functions* with Lipschitz properties. In particular, we give several applications of the method of bounded differences to random graphs. We also discuss bandit problems in machine learning. In the second thread in Section 3.3, we give an introduction to *potential theory* and *electrical network theory* for Markov chains. This toolkit in particular provides bounds on hitting times for random walks on networks, with important implications in the study of recurrence among other applications. We also introduce Wilson's remarkable method for generating uniform spanning trees.

3.1 Background

We begin with a quick review of stopping times and martingales. Along the way, we prove a few useful results. In particular, we derive some bounds on hitting times and cover times of Markov chains.

Throughout, $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{Z}_+}, \mathbb{P})$ is a filtered space. See Appendix B for a formal definition. Recall that, intuitively, the σ-algebra $\mathcal{F}_t$ in the filtration $(\mathcal{F}_t)_t$ represents "'the information known at time t." All time indices are discrete (in $\mathbb{Z}_+$

Version: December 20, 2023
Modern Discrete Probability: An Essential Toolkit

unless stated otherwise). We will also use the notation $\overline{\mathbb{Z}}_+ := \{0, 1, \ldots, +\infty\}$ to allow time $+\infty$.

3.1.1 Stopping times

Definitions Roughly speaking, a stopping time is a random time whose value is determined by a rule not depending on the future. Formally:

Definition 3.1.1 (Stopping time)**.** *A random variable $\tau : \Omega \to \overline{\mathbb{Z}}_+$ is called a* stopping time *if*

stopping time

$$\{\tau \leq t\} \in \mathcal{F}_t, \ \forall t \in \overline{\mathbb{Z}}_+,$$

or, equivalently,

$$\{\tau = t\} \in \mathcal{F}_t, \ \forall t \in \overline{\mathbb{Z}}_+.$$

To see the equivalence above, note that $\{\tau = t\} = \{\tau \leq t\} \setminus \{\tau \leq t-1\}$, and $\{\tau \leq t\} = \cup_{i \leq t}\{\tau = i\}$.

Example 3.1.2 (Hitting time)**.** Let $(A_t)_{t \in \mathbb{Z}_+}$, with values in $(E, \mathcal{E})$, be adapted and let $B \in \mathcal{E}$. 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. ◀

hitting time

Let τ be a stopping time. Denote by $\mathcal{F}_\tau$ the set of all events F such that, $\forall t \in \overline{\mathbb{Z}}_+$, $F \cap \{\tau = t\} \in \mathcal{F}_t$. Intuitively, the σ-algebra $\mathcal{F}_\tau$ captures the information up to time τ. The following lemmas help clarify the definition of $\mathcal{F}_\tau$.

Lemma 3.1.3. *$\mathcal{F}_\tau = \mathcal{F}_s$ if $\tau := s$, $\mathcal{F}_\tau = \mathcal{F}_\infty = \sigma(\cup_t \mathcal{F}_t)$ if $\tau := +\infty$ and $\mathcal{F}_\tau \subseteq \mathcal{F}_\infty$ for any stopping time τ.*

Proof. In the first case, note that $F \cap \{\tau = t\}$ is empty if $t \neq s$ and is F if $t = s$. So if $F \in \mathcal{F}_\tau$ then $F = F \cap \{\tau = s\} \in \mathcal{F}_s$ by definition of $\mathcal{F}_\tau$, and if $F \in \mathcal{F}_s$ then $F = F \cap \{\tau = t\} \in \mathcal{F}_t$ for all t by definition of τ. So we have proved both inclusions. This works also for $t = +\infty$. For the third claim note that, for any $F \in \mathcal{F}_\tau$,

$$F = \cup_{t \in \overline{\mathbb{Z}}_+} F \cap \{\tau = t\} \in \mathcal{F}_\infty,$$

again by definition of $\mathcal{F}_\tau$. ■

Lemma 3.1.4. *If (X_t) is adapted and τ is a stopping time then $X_\tau \in \mathcal{F}_\tau$ (where we assume that $X_\infty \in \mathcal{F}_\infty$, e.g., by setting $X_\infty := \liminf X_n$).*

Proof. For $B \in \mathcal{E}$,

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

by definition of τ. That shows X_τ is measurable with respect to $\mathcal{F}_\tau$ as claimed. ■

Lemma 3.1.5. *If σ, τ are stopping times then $\mathcal{F}_{\sigma\wedge\tau} \subseteq \mathcal{F}_\tau$.*

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

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

Indeed, the expression in parentheses is in $\mathcal{F}_s \subseteq \mathcal{F}_t$ by definition of $\mathcal{F}_{\sigma\wedge\tau}$ and $\{\tau = t\} \in \mathcal{F}_t$. ■

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

Definition 3.1.6 (First visit and return)**.** *The* first visit time *and* first return time *to $x \in V$ are* *first return*

$$\tau_x := \inf\{t \ge 0 \,:\, X_t = x\} \quad \textit{and} \quad \tau_x^+ := \inf\{t \ge 1 \,:\, X_t = x\}.$$

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

Definition 3.1.7 (Cover time)**.** *Assume V is finite. The* cover time *of (X_t) is the first time that all states have been visited, that is,* *cover time*

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

Strong Markov property Let (X_t) be a Markov chain with transition matrix P and initial distribution μ. Let $\mathcal{F}_t = \sigma(X_0, \dots, X_t)$. Recall that the Markov property (Theorem 1.1.18) says that, given the present, the future is independent of the past. The Markov property naturally extends to stopping times. Let τ be a stopping time with $\mathbb{P}[\tau < +\infty] > 0$. In its simplest form we have:

$$\mathbb{P}[X_{\tau+1} = y \,|\, \mathcal{F}_\tau] = \mathbb{P}_{X_\tau}[X_{\tau+1} = y] = P(X_\tau, y).$$

In words, the chain "starts fresh" at a stopping time with the state at that time as starting point. More generally:

Theorem 3.1.8 (Strong Markov property)**.** *Let $f_t : V^\infty \to \mathbb{R}$ be a sequence of measurable functions, uniformly bounded in t and let $F_t(x) := \mathbb{E}_x[f_t((X_s)_{s\ge0})]$. On $\{\tau < +\infty\}$,*

$$\mathbb{E}[f_\tau((X_{\tau+t})_{t\ge0}) \,|\, \mathcal{F}_\tau] = F_\tau(X_\tau).$$

Throughout, when we say that two random variables Y, Z are equal on an event B, we mean formally that $Y\mathbf{1}_B = Z\mathbf{1}_B$ almost surely.

Proof of Theorem 3.1.8. We use that

$$\mathbb{E}[f_\tau((X_{\tau+t})_{t\geq 0}) \,|\, \mathcal{F}_\tau]\mathbf{1}_{\tau<+\infty} = \mathbb{E}[f_\tau((X_{\tau+t})_{t\geq 0})\mathbf{1}_{\tau<+\infty} \,|\, \mathcal{F}_\tau].$$

Let $A \in \mathcal{F}_\tau$. Summing over the possible values of τ, using the tower property (Lemma B.6.16) and then the Markov property

$$\begin{aligned}
&\mathbb{E}[f_\tau((X_{\tau+t})_{t\geq 0})\mathbf{1}_{\tau<+\infty}; A] \\
&= \mathbb{E}[f_\tau((X_{\tau+t})_{t\geq 0}); A \cap \{\tau < +\infty\}] \\
&= \sum_{s\geq 0} \mathbb{E}[f_s((X_{s+t})_{t\geq 0}); A \cap \{\tau = s\}] \\
&= \sum_{s\geq 0} \mathbb{E}[\mathbb{E}[f_s((X_{s+t})_{t\geq 0}); A \cap \{\tau = s\} \,|\, \mathcal{F}_s]] \\
&= \sum_{s\geq 0} \mathbb{E}[\mathbf{1}_{A\cap\{\tau=s\}}\mathbb{E}[f_s((X_{s+t})_{t\geq 0}) \,|\, \mathcal{F}_s]] \\
&= \sum_{s\geq 0} \mathbb{E}[\mathbf{1}_{A\cap\{\tau=s\}}F_s(X_s)] \\
&= \sum_{s\geq 0} \mathbb{E}[F_s(X_s); A \cap \{\tau = s\}] \\
&= \mathbb{E}[F_\tau(X_\tau); A \cap \{\tau < +\infty\}] \\
&= \mathbb{E}[F_\tau(X_\tau)\mathbf{1}_{\tau<+\infty}; A],
\end{aligned}$$

where, on the fifth line, we used that $A \cap \{\tau = s\} \in \mathcal{F}_s$ by definition of $\mathcal{F}_\tau$ and taking out what is known (Lemma B.6.13). The definition of the conditional expectation (Theorem B.6.1) concludes the proof. ∎

The following typical application of the strong Markov property (Theorem 3.1.8) is useful.

Theorem 3.1.9 (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 of $\mathcal{F}_\tau$ and is symmetric about 0. In particular, it has

probability at least $1/2$ of being greater or equal to 0 by the first moment principle (Theorem 2.2.1), an event which implies that S_t is greater than or equal to b. Hence

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

(Exercise 3.1 asks for a more formal proof.) ■

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

Theorem 3.1.10 (Reflection principle: simple random walk)**.** *Let (S_t) be simple random walk on $\mathbb{Z}$ started at 0. Then, $\forall a, b, t > 0$,*

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

and

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

Proof. For the first claim, reflect the sub-path after the first visit to b across the line $y = b$. Summing over $a > 0$ and rearranging gives the second claim. ■

We record another related result that will be useful later.

Theorem 3.1.11 (Ballot theorem)**.** *In an election with n voters, candidate A gets α 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 $\tau_y^0 := 0$)

k-th return

$$\tau_y^k := \inf\{t > \tau_y^{k-1} \,:\, X_t = y\}.$$

In particular, $\tau_y^1 = \tau_y^+$. Define $\rho_{xy} := \mathbb{P}_x[\tau_y^+ < +\infty]$. Then by the strong Markov property (and induction)

$$\mathbb{P}_x[\tau_y^k < +\infty] = \rho_{xy}\rho_{yy}^{k-1}. \tag{3.1.1}$$

(Exercise 3.2 asks for a more formal proof.) Letting

$$N_y := \sum_{t>0} \mathbf{1}_{\{X_t = y\}} = \sum_{k \geq 1} \mathbf{1}_{\{\tau_y^k < +\infty\}},$$

be the number of visits to y after time 0, by linearity

$$\mathbb{E}_x[N_y] = \frac{\rho_{xy}}{1-\rho_{yy}}. \tag{3.1.2}$$

When $\rho_{yy} < 1$, we have $\mathbb{E}_y[N_y] < +\infty$ by (3.1.2), and in particular $\tau_y^k = +\infty$ for some k. Or $\rho_{yy} = 1$ and, starting at $x = y$, we have $\tau_y^k < +\infty$ almost surely for all k by (3.1.1). That leads us to the following dichotomy.

Definition 3.1.12 (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*. Let x be recurrent. If in addition $\mathbb{E}_x[\tau_x^+] < +\infty$, we say that x is* positive recurrent*; otherwise we say that it is* null recurrent*. A chain is* recurrent *(or* transient*, or* positive recurrent*, or* null recurrent*) if all its states are.*

recurrent

Recurrence is "contagious" in the following sense.

Lemma 3.1.13. *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 $\rho_{xy} > 0$. This definition is consistent with (and generalizes to sets) the one we gave in Section 1.1.2. Recall that we have the following decomposition theorem.

Theorem 3.1.14 (Decomposition theorem)**.** *Let $R := \{x : \rho_{xx} = 1\}$ be the recurrent states of the chain. Then R can be written as a disjoint union $\cup_j R_j$ where each R_j is closed and irreducible.*

Example 3.1.15 (Simple random walk on $\mathbb{Z}$)**.** Consider simple random walk (S_t) on $\mathbb{Z}$ started at 0. The chain is clearly irreducible so it suffices to check the type of state 0 by Lemma 3.1.13. First note the periodicity of this chain. So we look at S_{2t}. Then by Stirling's formula (see Appendix A)

$$\mathbb{P}[S_{2t} = 0] \quad = \quad \binom{2t}{t} 2^{-2t} \sim 2^{-2t} \frac{(2t)^{2t}}{(t^t)^2} \frac{\sqrt{2t}}{\sqrt{2\pi t}} \sim \frac{1}{\sqrt{\pi t}}.$$

Thus

$$\mathbb{E}[N_0] = \sum_{t>0} \mathbb{P}[S_t = 0] = +\infty,$$

and the chain is recurrent. ◀

Return times are closely related to stationary measures. We recall the following standard results without proof. We gave an alternative proof of the existence of a unique stationary distribution in the finite, irreducible case in Theorem 1.1.24.

Theorem 3.1.16. *Let x be a recurrent state. Then the following defines a stationary measure*

$$\mu_x(y) := \mathbb{E}_x \left[\sum_{0 \le t < \tau_x^+} \mathbf{1}_{\{X_t = y\}} \right].$$

Theorem 3.1.17. *If (X_t) is irreducible and recurrent, then the stationary measure is unique up to a constant multiple.*

Theorem 3.1.18. *If there is a stationary distribution π then all states y that have $\pi(y) > 0$ are recurrent.*

Theorem 3.1.19. *If (X_t) is irreducible and has a stationary distribution π, then*

$$\pi(x) = \frac{1}{\mathbb{E}_x \tau_x^+}.$$

Theorem 3.1.20. *If (X_t) is irreducible, then the following are equivalent.*

(i) There is a stationary distribution.

(ii) All states are positive recurrent.

(iii) There is a positive recurrent state.

We have seen previously that, in the irreducible, positive recurrent, aperiodic case, there is convergence to stationarity (see Theorem 1.1.33). In the transient and null recurrent cases, there is no stationary distribution to converge to by Theorem 3.1.20. Instead, we have the following.

Theorem 3.1.21 (Convergence of P^t: transient and null recurrent cases)**.** *If P is an irreducible chain which is either transient or null recurrent, we have for all x, y that*

$$\lim_t P^t(x, y) = 0.$$

Proof. We only prove the transient case. In that case, we showed in (3.1.2) that

$$\sum_t P^t(x, y) = \mathbb{E}_x \left[\sum_t \mathbf{1}_{\{X_t = y\}} \right] = \mathbb{E}_x[N_y] < +\infty.$$

Hence $P^t(x, y) \to 0$. ∎

A useful identity A slight generalization of the "cycle trick" used in the proof of Theorem 3.1.16 gives a useful identity.

Definition 3.1.22 (Green function)**.** *Let σ be a stopping time for a Markov chain (X_t). The* Green function *of the chain stopped at σ is given by*

Green function

$$\mathscr{G}_\sigma(x,y) = \mathbb{E}_x\left[\sum_{0\le t<\sigma} \mathbf{1}_{\{X_t=y\}}\right], \qquad x,y\in V, \tag{3.1.3}$$

that is, it is the expected number of visits to y before *σ when started at x.*

Lemma 3.1.23 (Occupation measure identity)**.** *Consider an irreducible, positive recurrent Markov chain $(X_t)_{t\ge 0}$ with transition matrix P and stationary distribution π. Let x be a state and σ be a stopping time such that $\mathbb{E}_x[\sigma] < +\infty$ and $\mathbb{P}_x[X_\sigma = x] = 1$. For any y,*

$$\mathscr{G}_\sigma(x,y) = \pi_y\, \mathbb{E}_x[\sigma].$$

Proof. By the uniqueness of the stationary measure up to constant multiple (Theorem 3.1.17), it suffices to show that $\mathscr{G}_\sigma(x,y)$ satisfies the system for a stationary measure as a function of y

$$\sum_y \mathscr{G}_\sigma(x,y)P(y,z) = \mathscr{G}_\sigma(x,z), \quad \forall z, \tag{3.1.4}$$

and use the fact that

$$\sum_y \mathscr{G}_\sigma(x,y) = \sum_y \mathbb{E}_x\left[\sum_{0\le t<\sigma} \mathbf{1}_{\{X_t=y\}}\right] = \mathbb{E}_x[\sigma].$$

To check (3.1.4), because $X_\sigma = X_0$ almost surely, observe that

$$\begin{aligned}
\mathscr{G}_\sigma(x,z) &= \mathbb{E}_x\left[\sum_{0\le t<\sigma} \mathbf{1}_{X_t=z}\right] \\
&= \mathbb{E}_x\left[\sum_{0\le t<\sigma} \mathbf{1}_{X_{t+1}=z}\right] \\
&= \sum_{t\ge 0} \mathbb{P}_x[X_{t+1}=z, \sigma > t].
\end{aligned}$$

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

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

which establishes (3.1.4) and proves the claim. ∎

Here is a typical application of this lemma.

Corollary 3.1.24. *In the setting of Lemma 3.1.23, for all $x \neq y$,*

$$\mathbb{P}_x[\tau_y < \tau_x^+] = \frac{1}{\pi_x(\mathbb{E}_x[\tau_y] + \mathbb{E}_y[\tau_x])}.$$

Proof. Let σ be the time of the first visit to x after the first visit to y. Then $\mathbb{E}_x[\sigma] = \mathbb{E}_x[\tau_y] + \mathbb{E}_y[\tau_x] < +\infty$, where we used that the chain is irreducible and positive recurrent. By the strong Markov property, the number of visits to x before the first visit to y is geometric with success probability $\mathbb{P}_x[\tau_y < \tau_x^+]$ (where, here, a visit to x is a "failed trial"). Moreover the number of visits to x after the first visit to y but before σ is 0 by definition. Hence $\mathscr{G}_\sigma(x,x)$ is the mean of the geometric distribution, namely $1/\mathbb{P}_x[\tau_y < \tau_x^+]$. Applying the occupation measure identity gives the result. ∎

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 any hitting time converges to 0 exponentially fast.

Lemma 3.1.25. *Let (X_t) be a finite, irreducible Markov chain with state space V. For any subset of states $A \subseteq V$ and initial distribution μ:*

(i) It holds that $\mathbb{E}_\mu[\tau_A] < +\infty$ (and, in particular, $\tau_A < +\infty$ a.s.).

(ii) Letting $\bar{\mathrm{t}}_A := \max_x \mathbb{E}_x[\tau_A]$, we have the tail bound

$$\mathbb{P}_\mu[\tau_A > t] \leq \exp\left(-\left\lfloor \frac{t}{\lceil e\,\bar{\mathrm{t}}_A \rceil} \right\rfloor\right).$$

Proof. For any positive integer m, for some distribution θ over the state space V, by the strong Markov property (Theorem 3.1.8)

$$\mathbb{P}_\mu[\tau_A > ms \,|\, \tau_A > (m-1)s] = \mathbb{P}_\theta[\tau_A > s] \leq \max_x \mathbb{P}_x[\tau_A > s] =: \alpha_s.$$

Choose a positive integer 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 $\alpha_s < 1$.

By the multiplication rule and the monotonicity of the events $\{\tau_A > rs\}$ over r, we have

$$\mathbb{P}_\mu[\tau_A > ms] = \mathbb{P}_\mu[\tau_A > s] \prod_{r=2}^{m} \mathbb{P}_\mu[\tau_A > rs \,|\, \tau_A > (r-1)s].$$

Therefore, $\mathbb{P}_\mu[\tau_A > ms] \leq \alpha_s^m$, which in turn implies

$$\mathbb{P}_\mu[\tau_A > t] \leq \alpha_s^{\lfloor \frac{t}{s} \rfloor}. \tag{3.1.5}$$

The result for the expectation follows from

$$\mathbb{E}_\mu[\tau_A] = \sum_{t \geq 0} \mathbb{P}_\mu[\tau_A > t] \leq \sum_t \alpha_s^{\lfloor \frac{t}{s} \rfloor} < +\infty,$$

since $\alpha_s < 1$.

Now that we have established that $\bar{\mathrm{t}}_A < +\infty$, by Markov's inequality (Theorem 2.1.1),

$$\alpha_s = \max_x \mathbb{P}_x[\tau_A > s] \leq \frac{\bar{\mathrm{t}}_A}{s}.$$

for all non-negative integers s. Plugging back into (3.1.5) gives $\mathbb{P}_\mu[\tau_A > t] \leq \left(\frac{\bar{\mathrm{t}}_A}{s}\right)^{\lfloor \frac{t}{s} \rfloor}$. By differentiating with respect to s, it can be checked that a good choice for s is $\lceil e\, \bar{\mathrm{t}}_A \rceil$. Simplifying gives the second claim. ∎

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 $\tau_{\mathrm{cov}} := \max_y \tau_y$. We bound the mean cover time in terms of

$$\bar{\mathrm{t}}_{\mathrm{hit}} := \max_{x \neq y} \mathbb{E}_x \tau_y.$$

Claim 3.1.26.

$$\max_x \mathbb{E}_x[\tau_{\mathrm{cov}}] \leq (3 + \log n) \lceil e\, \bar{\mathrm{t}}_{\mathrm{hit}} \rceil.$$

Proof. By a union bound over all states to be visited and Lemma 3.1.25,

$$\max_x \mathbb{P}_x[\tau_{\mathrm{cov}} > t] \le \min\left\{1, n \exp\left(-\left\lfloor \frac{t}{\lceil e\,\bar{\mathrm{t}}_{\mathrm{hit}}\rceil}\right\rfloor\right)\right\}.$$

Summing over $t \in \mathbb{Z}_+$ and appealing to the sum of a geometric series,

$$\max_x \mathbb{E}_x[\tau_{\mathrm{cov}}] \le (\log n + 1)\lceil e\,\bar{\mathrm{t}}_{\mathrm{hit}}\rceil + \frac{1}{1-e^{-1}}\lceil e\,\bar{\mathrm{t}}_{\mathrm{hit}}\rceil,$$

where the first term on the right-hand side comes from the fact that until $t \ge (\log n + 1)\lceil e\,\bar{\mathrm{t}}_{\mathrm{hit}}\rceil$ the upper bound above is 1. The factor $\lceil e\,\bar{\mathrm{t}}_{\mathrm{hit}}\rceil$ in the second term on the right-hand side comes from the fact that we must break up the series into blocks of size $\lceil e\,\bar{\mathrm{t}}_{\mathrm{hit}}\rceil$. Simplifying gives the claim. ∎

The previous proof should be reminiscent of that of Theorem 2.4.21.

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

Theorem 3.1.27 (Matthews' cover time bounds)**.** *Let*

$$\underline{\mathrm{t}}^{A}_{\mathrm{hit}} := \min_{x,y\in A,\ x\neq y} \mathbb{E}_x\tau_y,$$

and $h_n := \sum_{m=1}^{n} \frac{1}{m}$. *Then*

$$\max_x \mathbb{E}_x[\tau_{\mathrm{cov}}] \le h_n\,\bar{\mathrm{t}}_{\mathrm{hit}}, \tag{3.1.6}$$

and

$$\min_x \mathbb{E}_x[\tau_{\mathrm{cov}}] \ge \max_{A\subseteq V} h_{|A|-1}\,\underline{\mathrm{t}}^{A}_{\mathrm{hit}}. \tag{3.1.7}$$

Clearly, $\max_{x\neq y} \underline{\mathrm{t}}^{\{x,y\}}_{\mathrm{hit}}$ is a lower bound on the worst expected cover time. Lower bound (3.1.7) says that a tighter bound is obtained by finding a larger subset of states A that are "far away" from each other.

We sketch the proof of the lower bound for $A = V$, which we assume is $[n]$ without loss of generality. The other cases are similar. Let $(J_1, \dots, J_n)$ be a uniform random ordering of V, let $C_m := \max_{i\le m} \tau_{J_i}$, and let L_m be the last state visited among $J_1, \dots, J_m$. Then for $m \ge 2$

$$\mathbb{E}_x[C_m - C_{m-1} \mid J_1, \dots, J_m, \{X_t, t \le C_{m-1}\}] \ge \underline{\mathrm{t}}^{V}_{\mathrm{hit}}\,\mathbf{1}_{\{L_m = J_m\}}.$$

By symmetry, $\mathbb{P}[L_m = J_m] = \frac{1}{m}$. To see this, first pick the set of vertices corresponding to $\{J_1, \dots, J_m\}$, wait for all of those vertices to be visited, then pick the ordering. Moreover observe that $\mathbb{E}_x C_1 \ge (1 - \frac{1}{n})\underline{\mathrm{t}}^{V}_{\mathrm{hit}}$ where the factor of $(1 - \frac{1}{n})$ accounts for the probability that $J_1 \neq x$. Taking expectations above and summing over m gives the result.

Exercise 3.3 asks for a proof that the bounds above cannot in general be improved up to smaller order terms.

3.1.3 Martingales

Definition Martingales are an important class of stochastic processes that correspond intuitively to the "probabilistic version of a monotone sequence." They hide behind many processes and have properties that make them powerful tools in the analysis of processes where they have been identified. Formally:

Definition 3.1.28 (Martingale)**.** *An adapted process* $(M_t)_{t\geq 0}$ *with* $\mathbb{E}|M_t| < +\infty$ *for all* t *is a* martingale *if*

martingale

$$\mathbb{E}[M_{t+1} \,|\, \mathcal{F}_t] = M_t, \qquad \forall t \geq 0.$$

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

Recall that adapted (Definition B.7.5) simply means that $M_t \in \mathcal{F}_t$, that is, roughly speaking M_t is "known at time t." Note that for a martingale, by the tower property (Lemma B.6.16), we have $\mathbb{E}[M_t \,|\, \mathcal{F}_s] = M_s$ for all $t > s$, and similarly (with inequalities) for supermartingales and submartingales.

We start with a straightforward example.

Example 3.1.29 (Sums of i.i.d. random variables with mean 0)**.** Let $X_0, X_1, \ldots$ be i.i.d. integrable, centered random variables, $\mathcal{F}_t = \sigma(X_0, \ldots, X_t)$, $S_0 = 0$, and $S_t = \sum_{i=1}^t X_i$. Note that $\mathbb{E}|S_t| < \infty$ by the triangle inequality. By taking out what is known and the role of independence lemma (Lemma B.6.14) we obtain

$$\mathbb{E}[S_t \,|\, \mathcal{F}_{t-1}] = \mathbb{E}[S_{t-1} + X_t \,|\, \mathcal{F}_{t-1}] = S_{t-1} + \mathbb{E}[X_t] = S_{t-1},$$

which proves that (S_t) is a martingale. ◀

Martingales however are richer than random walks with centered steps. For instance mixtures of such random walks are also martingales.

Example 3.1.30 (Mixtures of random walks)**.** Consider again the setting of Example 3.1.29. This time assume that X_0 is uniformly distributed in $\{1, 2\}$ and define

$$R_t = X_0 S_t, \quad t \geq 0.$$

Then, because (S_t) is a martingale,

$$\mathbb{E}[R_t \,|\, \mathcal{F}_{t-1}] = X_0 \mathbb{E}[S_t \,|\, \mathcal{F}_{t-1}] = X_0 S_{t-1} = R_{t-1},$$

so (R_t) is also a martingale.

Further examples Martingales can also be a little more hidden. Here are two examples.

Example 3.1.31 (Variance of a sum of i.i.d. random variables)**.** Consider again the setting of Example 3.1.29 with $\sigma^2 := \mathrm{Var}[X_1] < \infty$. Define

$$M_t = S_t^2 - t\sigma^2.$$

Note that by the triangle inequality and the fact that S_t has mean zero and is a sum of independent random variables

$$\mathbb{E}|M_t| \leq \sum_{i=1}^{t} \mathrm{Var}[X_i] + t\sigma^2 \leq 2t\sigma^2 < +\infty.$$

Moreover, arguing similarly to the previous example, and using the fact that both X_t and S_{t-1} are square integrable

$$\begin{aligned} \mathbb{E}[M_t \,|\, \mathcal{F}_{t-1}] &= \mathbb{E}[(X_t + S_{t-1})^2 - t\sigma^2 \,|\, \mathcal{F}_{t-1}] \\ &= \mathbb{E}[X_t^2 + 2X_t S_{t-1} + S_{t-1}^2 - t\sigma^2 \,|\, \mathcal{F}_{t-1}] \\ &= \sigma^2 + 0 + S_{t-1}^2 - t\sigma^2 \\ &= M_{t-1}, \end{aligned}$$

which proves that (M_t) is a martingale. ◀

Example 3.1.32 (Eigenvectors of a transition matrix)**.** Let $(X_t)_{t\geq 0}$ be a finite Markov chain with state space V and transition matrix P, and let $(\mathcal{F}_t)_{t\geq 0}$ be the corresponding filtration. Suppose $f : V \to \mathbb{R}$ is such that

$$\sum_j P(i,j) f(j) = \lambda f(i), \ \forall i \in S.$$

In other words, f is a (right) eigenvector of P with eigenvalue λ. Define

$$M_t = \lambda^{-t} f(X_t).$$

Note that by the finiteness of the state space

$$\mathbb{E}|M_t| < +\infty,$$

and that further by the Markov property

$$\begin{aligned} \mathbb{E}[M_t \,|\, \mathcal{F}_{t-1}] &= \lambda^{-t}\mathbb{E}[f(X_t) \,|\, \mathcal{F}_{t-1}] \\ &= \lambda^{-t} \sum_j P(X_{t-1}, j) f(j) \\ &= \lambda^{-t} \cdot \lambda f(X_{t-1}) \\ &= M_{t-1}. \end{aligned}$$

That is, (M_t) is a martingale. ◀

Or we can create martingales out of thin air. We give two important examples that will appear later.

Example 3.1.33 (Doob martingale: accumulating data)**.** Let X with $\mathbb{E}|X| < +\infty$. Define $M_t = \mathbb{E}[X \,|\, \mathcal{F}_t]$. Note that $\mathbb{E}|M_t| \leq \mathbb{E}|X| < +\infty$ by Jensens' inequality, and

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

by the tower property. This is known as a *Doob martingale*. Intuitively this process tracks our expectation of the unobserved X as "more information becomes available." See the co-called "exposure martingales" in Section 3.2.3 for a concrete illustration of this idea. ◀

Doob martingale

Example 3.1.34 (Martingale transform)**.** Let $(X_t)_{t\geq 1}$ be an integrable, adapted process and let $(C_t)_{t\geq 1}$ be a bounded, predictable process. Recall that predictable (Definition B.7.6) means $C_t \in \mathcal{F}_{t-1}$ for all t, that is, roughly speaking C_t is "known at time $t-1$." Define

$$N_t = \sum_{i\leq t} (X_i - \mathbb{E}[X_i \,|\, \mathcal{F}_{i-1}])C_i.$$

Then

$$\mathbb{E}|N_t| \leq \sum_{i\leq t} 2\mathbb{E}|X_t|K < +\infty,$$

where we used that $|C_t| < K$ for all $t \geq 1$, and

$$\begin{aligned} \mathbb{E}[N_t - N_{t-1} \,|\, \mathcal{F}_{t-1}] &= \mathbb{E}[(X_t - \mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}])C_t \,|\, \mathcal{F}_{t-1}] \\ &= C_t(\mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}] - \mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}]) \\ &= 0, \end{aligned}$$

by taking out what is known. So (N_t) is a martingale.

When (X_t) is itself a martingale (in which case $\mathbb{E}[X_i \,|\, \mathcal{F}_{i-1}] = X_{i-1}$ in the definition of N_t), this is a sort of "stochastic (Stieltjes) integral." When, instead, (X_t) is a supermartingale (respectively submartingale) and (C_t) is nonnegative and bounded, then the same computation shows that

$$N_t = \sum_{i\leq t} (X_i - X_{i-1})C_i,$$

defines a supermartingale (respectively submartingale). ◀

As implied by the next lemma, an immediate consequence of Jensen's inequality (in its conditional version of Lemma B.6.12), submartingales naturally arise as convex functions of martingales.

Lemma 3.1.35. *If $(M_t)_{t\geq 0}$ is a martingale and ϕ is a convex function such that $\mathbb{E}|\phi(M_t)| < +\infty$ for all t, then $(\phi(M_t))_{t\geq 0}$ is a submartingale. Moreover, if $(M_t)_{t\geq 0}$ is a submartingale and ϕ 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.*

Martingales and stopping times A fundamental reason explaining the utility of martingales in analyzing a variety of stochastic processes is that they play nicely with stopping times, in particular, through what is known as the *optional stopping theorem* (in its various forms). We will encounter many applications of this important result. First a definition:

Definition 3.1.36. *Let (M_t) be an adapted process and σ be a stopping time. Then*

$$M_t^\sigma(\omega) := M_{\sigma(\omega)\wedge t}(\omega),$$

is M_t stopped at σ.

stopped process

Lemma 3.1.37. *Let (M_t) be a supermartingale and σ be a stopping time. Then the stopped process (M_t^σ) is a supermartingale and in particular*

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

The same result holds with equalities if (M_t) is a martingale, and with inequalities in the opposite direction if (M_t) is a submartingale.

Proof. Note that

$$M_t^\sigma - M_0 = \sum_{i\leq t} C_i(X_i - X_{i-1}),$$

with $C_i = \mathbf{1}\{i \leq \sigma\} \in \mathcal{F}_{i-1}$ (which is nonnegative and bounded) and $X_i = M_i$ for all i, and use Example 3.1.34 to conclude that $\mathbb{E}[M_{\sigma\wedge t}] \leq \mathbb{E}[M_0]$.

On the other hand,

$$M_t - M_t^\sigma = \sum_{i\leq t}(1 - C_i)(X_i - X_{i-1}).$$

So the other inequality follows from the same argument. ∎

Theorem 3.1.38 (Doob's optional stopping theorem)**.** *Let (M_t) be a supermartingale and σ be a stopping time. Then M_σ is integrable and*

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

if any of the following conditions hold:

(i) *σ is bounded;*

(ii) *(M_t) is uniformly bounded and σ is almost surely finite;*

(iii) *$\mathbb{E}[\sigma] < +\infty$ and (M_t) has bounded increments (i.e., there is $c > 0$ such that $|M_t - M_{t-1}| \leq c$ a.s. for all t);*

(iv) *(M_t) is nonnegative and σ is almost surely finite.*

The first three imply equality above if (M_t) is a martingale.

Proof. Case (iv) is Fatou's lemma (Proposition B.4.14). We prove (iii). We leave the proof of the other claims as an exercise (see Exercise 3.5).

From Lemma 3.1.37, we have

$$\mathbb{E}[M_{\sigma\wedge t} - M_0] \leq 0. \tag{3.1.8}$$

Furthermore the assumption that $\mathbb{E}[\sigma] < +\infty$ implies that $\sigma < +\infty$ almost surely. Hence we seek to take a limit as $t \to +\infty$ *inside the expectation*. To justify swapping limit and expectation, note that by a telescoping sum

$$\begin{aligned} |M_{\sigma\wedge t} - M_0| &\leq \left| \sum_{s \leq \sigma\wedge t} (M_s - M_{s-1}) \right| \\ &\leq \sum_{s \leq \sigma} |M_s - M_{s-1}| \\ &\leq c\sigma. \end{aligned}$$

The claim now follows from dominated convergence (Proposition B.4.14). Equality holds if (M_t) is a martingale. ∎

Although the optional stopping theorem (Theorem 3.1.38) is useful, one often works directly with Lemma 3.1.37 and applies suitable limit theorems (see Proposition B.4.14). The following martingale-based proof of Wald's first identity provides an illustration.

Theorem 3.1.39 (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*

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

Proof. We first prove the result for nonnegative X_is. By Example 3.1.29, $S_t - t\mu$ is a martingale and Lemma 3.1.37 implies that $\mathbb{E}[S_{\tau\wedge t} - \mu(\tau \wedge t)] = 0$, or

$$\mathbb{E}[S_{\tau\wedge t}] = \mu\,\mathbb{E}[\tau \wedge t].$$

Note that, in the nonnegative case, we have $S_{\tau\wedge t} \uparrow S_\tau$ and $\tau \wedge t \uparrow \tau$. Thus, by monotone convergence (Proposition B.4.14), the claim $\mathbb{E}[S_\tau] = \mu\,\mathbb{E}[\tau]$ follows in that case.

Consider now the general case. Again, $\mathbb{E}[S_{\tau\wedge t}] = \mu\,\mathbb{E}[\tau\wedge t]$ and $\mathbb{E}[\tau\wedge t] \uparrow \mathbb{E}[\tau]$. Applying the previous argument to the sum of nonnegative random variables $R_t = \sum_{s=1}^t |X_s|$ shows that $\mathbb{E}[R_\tau] = \mathbb{E}|X_1|\,\mathbb{E}[\tau] < +\infty$ by assumption. Since $|S_{\tau\wedge t}| \le R_\tau$ for all t by the triangle inequality, dominated convergence (Proposition B.4.14) implies $\mathbb{E}[S_{\tau\wedge t}] \to \mathbb{E}[S_\tau]$ and we are done. ∎

We also recall Wald's second identity. The proof, which we omit, uses the martingale in Example 3.1.31.

Theorem 3.1.40 (Wald's second identity)**.** *Let $X_1, X_2, \ldots \in L^2$ be i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathrm{Var}[X_1] = \sigma^2$ and let $\tau \in L^1$ be a stopping time. Let $S_t = \sum_{s=1}^t X_s$. Then*

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

We illustrate Wald's identities on the *gambler's ruin* problem that is characteristic of applications of stopping times in Markov chains. We consider the "unbiased" and "biased" cases separately.

gambler's ruin

Example 3.1.41 (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 $-\infty < a < 0 < b < +\infty$, where the first visit time τ_x was defined in Definition 3.1.6.

Claim 3.1.42. *We have:*

(i) $\tau < +\infty$ *almost surely;*

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

(iii) $\mathbb{E}[\tau] = -ab$;

(iv) $\tau_a < +\infty$ *almost surely but* $\mathbb{E}[\tau_a] = +\infty$.

Proof. We prove the claims in order.

(i) We argue that in fact $\mathbb{E}[\tau] < \infty$. That follows immediately from the exponential tail of hitting times in Lemma 3.1.25 for the chain $(S_{\tau\wedge t})$ whose (effective) state space, $\{a, a+1, \ldots, b\}$, is finite.

(ii) By Wald's first identity (Theorem 3.1.39) and (i), we have $\mathbb{E}[S_\tau] = 0$ or

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

that is, using $\mathbb{P}[S_\tau = a] = 1 - \mathbb{P}[S_\tau = b] = \mathbb{P}[\tau_a < \tau_b]$,

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

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

(iii) Because $\sigma^2 = 1$, Wald's second identity (Theorem 3.1.40) says that $\mathbb{E}[S_\tau^2] = \mathbb{E}[\tau]$. Furthermore, we have by (ii)

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

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

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

That concludes the proof. ■

Note that (iv) above shows that the L^1 condition on the stopping time in Wald's second identity (Theorem 3.1.40) is necessary. Indeed we have shown $a^2 = \mathbb{E}[S_{\tau_a}^2] \neq \sigma^2\mathbb{E}[\tau_a] = +\infty$. ◀

Example 3.1.43 (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 $S_t = \sum_{i=1}^t X_i$ where the X_is are i.i.d. in $\{-1, +1\}$ with $\mathbb{P}[X_1 = 1] = p$. Let again $\tau := \tau_a \wedge \tau_b$ where $a < 0 < b$. Define $q := 1 - p$, $\delta := p - q > 0$, and $\phi(x) := (q/p)^x$.

Claim 3.1.44. *We have:*

(i) $\tau < +\infty$ *almost surely;*

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

(iii) $\mathbb{E}[\tau_b] = \frac{b}{2p-1}$*;*

(iv) $\tau_a = +\infty$ *with positive probability.*

Proof. Let $\psi_t(x) := x - \delta t$. We use *two* martingales: $(\phi(S_t))$ and $(\psi_t(S_t))$. Observe that indeed both processes are clearly integrable and

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

and

$$\mathbb{E}[\psi_t(S_t) \,|\, \mathcal{F}_{t-1}] = p[S_{t-1} + 1 - \delta t] + q[S_{t-1} - 1 - \delta t] = \psi_{t-1}(S_{t-1}).$$

(i) This claim follows by the same argument as in the unbiased case.

(ii) Note that $(\phi(S_t))$ is a nonnegative, bounded martingale since $q < p$ by assumption. By Lemma 3.1.37 and dominated convergence (Proposition B.4.14),

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

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

$$\mathbb{P}[\tau_a < +\infty] = \frac{1}{\phi(a)} < 1, \tag{3.1.9}$$

so that $\tau_a = +\infty$ with positive probability. On the other hand, $\mathbb{P}[\tau_b < \tau_a] = 1 - \mathbb{P}[\tau_a < \tau_b] = \frac{\phi(0)-\phi(a)}{\phi(b)-\phi(a)}$, and taking $a \to -\infty$

$$P[\tau_b < +\infty] = 1.$$

(iii) By Lemma 3.1.37 applied to $(\psi_t(S_t))$,

$$0 = \mathbb{E}[S_{\tau_b \wedge t} - \delta(\tau_b \wedge t)]. \tag{3.1.10}$$

By monotone convergence (Proposition B.4.14), $\mathbb{E}[\tau_b \wedge t] \uparrow \mathbb{E}[\tau_b]$. Furthermore, observe that $-\inf_t S_t \geq 0$ almost surely since $S_0 = 0$. Moreover, for $x \geq 0$, by (3.1.9)

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

so that $\mathbb{E}[-\inf_t S_t] = \sum_{x \geq 1} \mathbb{P}[-\inf_t S_t \geq x] < +\infty$. Hence, in (3.1.10), we can use dominated convergence (Proposition B.4.14) with

$$|S_{\tau_b \wedge t}| \leq \max\{b, -\inf_t S_t\},$$

and the fact that $\tau_b < +\infty$ almost surely from (ii) to deduce that $\mathbb{E}[\tau_b] = \frac{\mathbb{E}[S_{\tau_b}]}{p-q} = \frac{b}{2p-1}$.

(iv) That claim was proved in (ii).

That concludes the proof. ■

Note that, in (iii) above, in order to apply Wald's first identity directly we would have had to prove that $\tau_b \in L^1$ first. ◀

We also obtain the following maximal version of Markov's inequality (Theorem 2.1.1).

Theorem 3.1.45 (Doob's submartingale inequality)**.** *Let (M_t) be a nonnegative submartingale. Then, for $b > 0$,*

$$\mathbb{P}\left[\sup_{0\le s\le t} M_s \ge b\right] \le \frac{\mathbb{E}[M_t]}{b}.$$

Observe that a naive application of Markov's inequality implies only that

$$\sup_{0\le s\le t} \mathbb{P}[M_s \ge b] \le \frac{\mathbb{E}[M_t]}{b},$$

where we used that $\mathbb{E}[M_s] \le \mathbb{E}[M_t]$ for all $0 \le s \le t$ for a submartingale. Introducing an appropriate stopping time immediately gives something stronger. (Exercise 3.6 asks for the supermartingale version of this.)

Proof. Let σ be the first time that $M_t \ge b$. Then the event of interest can be characterized as

$$\left\{\sup_{0\le s\le t} M_s \ge b\right\} = \{M_{\sigma\wedge t} \ge b\}.$$

By Markov's inequality,

$$\mathbb{P}[M_{\sigma\wedge t} \ge b] \le \frac{\mathbb{E}[M_{\sigma\wedge t}]}{b}.$$

Lemma 3.1.37 implies that $\mathbb{E}[M_{\sigma\wedge t}] \le \mathbb{E}[M_t]$, which concludes the proof. ■

One consequence of the previous bound is a strengthening of Chebyshev's inequality (Theorem 2.1.2) for sums of independent random variables.

Corollary 3.1.46 (Kolmogorov's maximal inequality)**.** *Let $X_1, X_2, \ldots$ be independent random variables with $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}[X_i] < +\infty$. Define $S_t = \sum_{i\le t} X_i$. Then, for $\beta > 0$,*

$$\mathbb{P}\left[\max_{i\le t} |S_i| \ge \beta\right] \le \frac{\mathrm{Var}[S_t]}{\beta^2}.$$

Proof. By Example 3.1.29, (S_t) is a martingale. By Lemma 3.1.35, (S_t^2) is hence a (nonnegative) submartingale. The result follows from Doob's submartingale inequality (Theorem 3.1.45). ■

Convergence Finally another fundamental result about martingales is the following convergence theorem, which we state without proof. We give a quick application below.

Theorem 3.1.47 (Convergence theorem)**.** *Let (M_t) be a supermartingale bounded in L^1. Then (M_t) converges almost surely to a finite limit M_∞. Moreover, letting $M_\infty := \limsup_t M_t$, then $M_\infty \in \mathcal{F}_\infty$ and $\mathbb{E}|M_\infty| < +\infty$.*

Corollary 3.1.48 (Convergence of non-negative supermartingales)**.** *If (M_t) is a non-negative supermartingale then M_t converges almost surely to a finite limit M_∞ with $\mathbb{E}[M_\infty] \leq \mathbb{E}[M_0]$.*

Proof. By the supermartingale property, (M_t) is bounded in L^1 since

$$\mathbb{E}|M_t| = \mathbb{E}[M_t] \leq \mathbb{E}[M_0], \ \forall t.$$

Then we use the martingale convergence theorem (Theorem 3.1.47) and Fatou's lemma (Proposition B.4.14). ■

Example 3.1.49 (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 tth draw. Let

Pólya's urn

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

Define M_t to be the fraction of green balls after the tth draw. Then

$$\begin{aligned}
\mathbb{E}[M_t \,|\, \mathcal{F}_{t-1}] &= \frac{R_{t-1}}{G_{t-1}+R_{t-1}} \frac{G_{t-1}}{G_{t-1}+R_{t-1}+1} \\
&\qquad + \frac{G_{t-1}}{G_{t-1}+R_{t-1}} \frac{G_{t-1}+1}{G_{t-1}+R_{t-1}+1} \\
&= \frac{G_{t-1}}{G_{t-1}+R_{t-1}} \\
&= M_{t-1}.
\end{aligned}$$

Since $M_t \geq 0$ and is a martingale, we have $M_t \to M_\infty$ almost surely. In fact, Exercise 3.4 asks for a proof that

$$\mathbb{P}[G_t = m+1] = \binom{t}{m} \frac{m!(t-m)!}{(t+1)!} = \frac{1}{t+1}.$$

So taking a limit as $t \to +\infty$

$$\mathbb{P}[M_t \leq x] = \frac{\lfloor x(t+2) - 1 \rfloor}{t+1} \to x.$$

That is, (M_t) converges in distribution to a uniform random variable on $[0,1]$. ◀

Convergence of the expectation in general requires stronger conditions. A simple case is boundedness in L^2. Before stating the result, we derive a key property of martingales in L^2 which will be useful later.

Lemma 3.1.50 (Orthogonality of increments)**.** *Let (M_t) be a martingale with $M_t \in L^2$ for all t. Let $s \leq t \leq u \leq v$. Then,*

$$\langle M_t - M_s, M_v - M_u \rangle = 0,$$

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

Proof. Use $M_u = \mathbb{E}[M_v \mid \mathcal{F}_u]$ and $M_t - M_s \in \mathcal{F}_u$, and apply the L^2 characterization of the conditional expectation (Theorem B.6.2). ∎

In words, martingale increments over disjoint time intervals are uncorrelated (provided the second moment exists). Note that this is weaker than the independence of increments of random walks. (See Section 3.2.1 for more discussion on this.)

Theorem 3.1.51 (Convergence of martingales bounded in L^2)**.** *Let (M_t) be a martingale with $M_t \in L^2$ for all t. Then (M_t) is bounded in L^2 if and only if*

$$\sum_{k \geq 1} \mathbb{E}[(M_t - M_{t-1})^2] < +\infty.$$

When this is the case, M_t converges almost surely and in L^2 to a finite limit M_∞, and furthermore

$$\mathbb{E}[M_t] \to \mathbb{E}[M_\infty] < +\infty,$$

as $t \to +\infty$.

Proof. Writing M_t as a telescoping sum of increments, the orthogonality of increments (Lemma 3.1.50) implies

$$\begin{aligned} \mathbb{E}[M_t^2] &= \mathbb{E}\left[\left(M_0 + \sum_{s=1}^{t} (M_s - M_{s-1})\right)^2\right] \\ &= \mathbb{E}[M_0^2] + \sum_{s=1}^{t} \mathbb{E}[(M_s - M_{s-1})^2], \end{aligned}$$

proving the first claim.

By the monotonicity of norms (Lemma B.4.16), (M_t) being bounded in L^2 implies that (M_t) is bounded in L^1 which, in turn, implies that M_t converges almost

surely to a finite limit M_∞ with $\mathbb{E}|M_\infty| < +\infty$ by Theorem 3.1.47. Then using Fatou's lemma (Proposition B.4.14) in

$$\mathbb{E}[(M_{t+s} - M_t)^2] = \sum_{t+1 \leq i \leq t+s} \mathbb{E}[(M_i - M_{i-1})^2],$$

gives

$$\mathbb{E}[(M_\infty - M_t)^2] \leq \sum_{t+1 \leq i} \mathbb{E}[(M_i - M_{i-1})^2].$$

The right-hand side goes to 0 as $t \to +\infty$ since the full series is finite, which proves the second claim.

The last claim follows from Lemmas B.4.16 and B.4.17. ∎

3.1.4 ▷ *Percolation: critical regime on infinite d-regular tree*

Consider bond percolation (see Definition 1.2.1) on the infinite d-regular tree $\mathbb{T}_d$ rooted at a vertex 0. In Section 2.3.3, we showed that

$$p_c(\mathbb{T}_d) = \sup\{p \in [0,1] \,:\, \mathbb{P}_p[|\mathcal{C}_0| = +\infty] = 0\} = \frac{1}{d-1},$$

where recall that $\mathcal{C}_0$ is the open cluster of the root. Here we consider the critical case, that is, we set density $p = \frac{1}{d-1}$. (The same results apply to the infinite b-ary tree $\widehat{\mathbb{T}}_b$ with $d = b + 1$.) Assume $d \geq 3$ (since $d = 2$ is simply a path).

First:

Claim 3.1.52. *$|\mathcal{C}_0| < +\infty$ almost surely.*

Let $X_n := |\partial_n \cap \mathcal{C}_0|$, where ∂_n are the n-th level vertices. In Section 2.3.3, we proved the same claim in the subcritical case using the first moment method. It does not work here because

$$\mathbb{E}X_n = d(d-1)^{n-1}p^n = \frac{d}{d-1} \nrightarrow 0.$$

Instead we use a martingale argument which will be generalized when we discuss branching processes in Section 6.1.

Proof of Claim 3.1.52. Let $b := d - 1$ be the branching ratio. Because the root has a different number of children, 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

$$\mathbb{E}[Z_n \,|\, \mathcal{F}_{n-1}] = bpZ_{n-1} = Z_{n-1}.$$

So (Z_n) is a nonnegative, integer-valued martingale and it converges almost surely to a finite limit by Corollary 3.1.48. (In particular, $\mathbb{E}[Z_n] = 1$, which will be useful below.) But, clearly, for any integer $k > 0$ and $N \geq 0$

$$\mathbb{P}[Z_n = k,\ \forall n \geq N] = 0,$$

so it must be that the limit is 0 almost surely. In other words, Z_n is eventually 0 for all n large enough. This is true for every child of the root. Hence the open cluster of the root is finite almost surely. ■

On the other hand:

Claim 3.1.53.

$$\mathbb{E}|\mathcal{C}_0| = +\infty.$$

Proof. Consider the descendant subtree, T_1, of the first child of the root, which we denote by 1. Let $\widetilde{\mathcal{C}}_1$ be the open cluster of 1 in T_1. As we showed in the previous claim, the expected number of vertices on any level of T_1 is 1. So $\mathbb{E}|\widetilde{\mathcal{C}}_1| = +\infty$ by summing over the levels. ■

3.2 Concentration for martingales and applications

The Chernoff-Cramér method extends naturally to martingales. This observation leads to powerful new tail bounds 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: a function $f(X_1, \ldots, X_n)$ of many independent random variables that is not too sensitive to any of its coordinates tends to be close to its mean.

3.2.1 Azuma-Hoeffding inequality

The main result of this section is the following generalization of Hoeffding's inequality (Theorem 2.4.10).

Theorem 3.2.1 (Maximal Azuma-Hoeffding inequality)**.** *Let $(Z_t)_{t\in\mathbb{Z}_+}$ be a martingale with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}_+}$. Assume that there are predictable processes (A_t) and (B_t) (i.e., $A_t, B_t \in \mathcal{F}_{t-1}$) and constants $0 < c_t < +\infty$ such that: for all $t \geq 1$, almost surely,*

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

Then, for all $\beta > 0$,

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

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

Proof of Theorem 3.2.1. As in the Chernoff-Cramér method, we start by applying Markov's inequality (Theorem 2.1.1). Here we use the maximal version for submartingales, Doob's submartingale inequality (Theorem 3.1.45). First notice that e^{sx} is increasing and convex for $s > 0$, so that by Lemma 3.1.35 the process $(e^{s(Z_t - Z_0)})_t$ is a submartingale. Hence, for $s > 0$, by Theorem 3.1.45

$$\begin{aligned}\mathbb{P}\left[\sup_{0\le i\le t}(Z_i - Z_0) \ge \beta\right] &= \mathbb{P}\left[\sup_{0\le i\le t} e^{s(Z_i - Z_0)} \ge e^{s\beta}\right] \\ &\le \frac{\mathbb{E}\left[e^{s(Z_t - Z_0)}\right]}{e^{s\beta}} \\ &= \frac{\mathbb{E}\left[e^{s\sum_{r=1}^t (Z_r - Z_{r-1})}\right]}{e^{s\beta}}. \end{aligned} \tag{3.2.1}$$

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. By taking out what is known (Lemma B.6.13)

$$\mathbb{E}\left[\mathbb{E}\left[e^{s\sum_{r=1}^t (Z_r - Z_{r-1})}\,\middle|\,\mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[e^{s\sum_{r=1}^{t-1}(Z_r - Z_{r-1})}\,\mathbb{E}\left[e^{s(Z_t - Z_{t-1})}\,\middle|\,\mathcal{F}_{t-1}\right]\right].$$

The martingale property and the assumption in the statement imply that, conditioned on $\mathcal{F}_{t-1}$, 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 2.4.12), it holds almost surely that

$$\mathbb{E}\left[e^{s(Z_t - Z_{t-1})}\,\middle|\,\mathcal{F}_{t-1}\right] \le \exp\left(\frac{s^2 c_t^2/4}{2}\right) = \exp\left(\frac{c_t^2 s^2}{8}\right). \tag{3.2.2}$$

Using the tower property (Lemma B.6.16) and arguing by induction, we obtain

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

Put differently, we have proved that $Z_t - Z_0$ is sub-Gaussian with variance factor $\frac{1}{4}\sum_{r\le t} c_r^2$. By (2.4.16) (or, equivalently, by choosing $s = \beta / \frac{1}{4}\sum_{r\le t} c_r^2$ in (3.2.1)) we get the result. ∎

In Theorem 3.2.1 the *martingale difference* sequence (X_t), where $X_t := Z_t - Z_{t-1}$, is not only "pairwise uncorrelated" by Lemma 3.1.50, that is,

martingale difference

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

but it is in fact "mutually uncorrelated," that is,

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

This stronger property helps explain why $\sum_{r \leq t} X_r$ is highly concentrated. This point is the subject of Exercise 3.7, which guides the reader through a slightly different proof of the Azuma-Hoeffding inequality. Compare with Exercises 2.5 and 2.6.

3.2.2 Method of bounded differences

The power of the maximal Azuma-Hoeffding inequality (Theorem 3.2.1) 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. The trick to analyzing the concentration of $f(X)$ is to consider the Doob martingale (see Example 3.1.33)

$$Z_i = \mathbb{E}[f(X) \,|\, \mathcal{F}_i], \tag{3.2.3}$$

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

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

and

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

so that we can write

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

Intuitively, the martingale difference $Z_i - Z_{i-1}$ tracks the change in our expectation of $f(X)$ as X_i is revealed.

In fact a clever probabilistic argument relates martingale differences directly to discrete derivatives. 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

$$\begin{aligned} Z_i - Z_{i-1} &= \mathbb{E}[f(X) \,|\, \mathcal{F}_i] - \mathbb{E}[f(X) \,|\, \mathcal{F}_{i-1}] \\ &= \mathbb{E}[f(X) \,|\, \mathcal{F}_i] - \mathbb{E}[f(X^{(i)}) \,|\, \mathcal{F}_{i-1}] \\ &= \mathbb{E}[f(X) \,|\, \mathcal{F}_i] - \mathbb{E}[f(X^{(i)}) \,|\, \mathcal{F}_i] \\ &= \mathbb{E}[f(X) - f(X^{(i)}) \,|\, \mathcal{F}_i]. \end{aligned}$$

Note that we crucially used the independence of the X_ks in the second and third lines. But then, by Jensen's inequality (Lemma B.6.12),

$$|Z_i - Z_{i-1}| \leq \|D_i f\|_\infty. \tag{3.2.4}$$

Assume further that $\mathbb{E}[f(X)^2] < +\infty$. By the orthogonality of increments of martingales in L^2 (Lemma 3.1.50), we immediately obtain a bound on the variance of f

$$\mathrm{Var}[f(X)] = \mathbb{E}[(Z_n - Z_0)^2] = \sum_{i=1}^{n} \mathbb{E}\left[(Z_i - Z_{i-1})^2\right] \leq \sum_{i=1}^{n} \|D_i f\|_\infty^2. \tag{3.2.5}$$

By the maximal Azuma-Hoeffding inequality and the fact that

$$Z_i - Z_{i-1} \in [-\|D_i f\|_\infty, \|D_i f\|_\infty],$$

we also get a bound on the tail

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\sum_{i \leq n} \|D_i f\|_\infty^2}\right). \tag{3.2.6}$$

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

We emphasize that, although it may not be immediately obvious, independence plays a crucial role in the bound (3.2.4), as the next example shows.

Example 3.2.2 (A counterexample)**.** Let $f(x_1, \dots, x_n) = x_1 + \cdots + x_n$ where $x_i \in \{-1, 1\}$ for all i. Then,

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

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

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

and

$$\mathbb{E}[f(X_1, \dots, X_n) \,|\, X_1] = nX_1,$$

so that

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

In particular, the corresponding Doob martingale does not have increments bounded by $\|D_i f\|_\infty = 2$.

Fo 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 $i = 1, \dots, n-1$. Let again X_1 be a uniform random variable on $\{-1, 1\}$ and, for $i = 2, \dots, n$, 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 $1 - \varepsilon$ and otherwise is flipped. Then,

$$\begin{aligned} \mathbb{E}[f(X_1, \dots, X_n)] &= \mathbb{E}\left[X_1 + \cdots + X_n\right] \\ &= \mathbb{E}\left[X_1 \left(1 + \sum_{i=1}^{n-1} \prod_{j \le i} U_j\right)\right] \\ &= \mathbb{E}[X_1]\, \mathbb{E}\left[1 + \sum_{i=1}^{n-1} \prod_{j \le i} U_j\right] \\ &= 0, \end{aligned}$$

by the independence of X_1 and the U_is. Similarly

$$\mathbb{E}[f(X_1, \dots, X_n) \,|\, X_1] = X_1\, \mathbb{E}\left[1 + \sum_{i=1}^{n-1} \prod_{j \le i} U_j\right] = X_1 \left(\sum_{i=0}^{n-1} (1 - 2\varepsilon)^i\right),$$

so that

$$|\mathbb{E}[f(X_1,\dots,X_n)\,|\,X_1] - \mathbb{E}[f(X_1,\dots,X_n)]| = \left(\sum_{i=0}^{n-1}(1-2\varepsilon)^i\right) > 2,$$

for ε small enough and $n \geq 3$. In particular, the corresponding Doob martingale does not have increments bounded by $\|D_i f\|_\infty = 2$. ◀

Variance bounds

We give improved bounds on the variance (compared to (3.2.5)). Our first bound explicitly decomposes the variance of $f(X)$ over the contributions of its individual entries.

Theorem 3.2.3 (Tensorization of the variance)**.** *Let $X_1,\dots,X_n$ be independent random variables where X_i is $\mathcal{X}_i$-valued for all i and let $X = (X_1,\dots,X_n)$. Assume that $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ is a measurable function with $\mathbb{E}[f(X)^2] < +\infty$. Define $\mathcal{F}_i = \sigma(X_1,\dots,X_i)$, $\mathcal{G}_i = \sigma(X_1,\dots,X_{i-1},X_{i+1},\dots,X_n)$ and $Z_i = \mathbb{E}[f(X)\,|\,\mathcal{F}_i]$. Then we have*

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

Proof of Theorem 3.2.3. The key lemma is the following.

Lemma 3.2.4.

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

Proof. By the tower property (Lemma B.6.16),

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

Moreover, $\sigma(X_i)$ is independent of $\sigma(\mathcal{G}_i,\mathcal{F}_{i-1})$ so by the role of independence (Lemma B.6.14), we have

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

Combining the last two displays gives the result. ■

Again, we take advantage of the orthogonality of increments to write

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

By the lemma above,

$$\begin{aligned}(Z_i - Z_{i-1})^2 &= (\mathbb{E}[f(X) \,|\, \mathcal{F}_i] - \mathbb{E}[f(X) \,|\, \mathcal{F}_{i-1}])^2 \\ &= (\mathbb{E}[f(X) \,|\, \mathcal{F}_i] - \mathbb{E}[\mathbb{E}[f(X) \,|\, \mathcal{G}_i] \,|\, \mathcal{F}_i])^2 \\ &= (\mathbb{E}[f(X) - \mathbb{E}[f(X) \,|\, \mathcal{G}_i] \,|\, \mathcal{F}_i])^2 \\ &\leq \mathbb{E}\left[(f(X) - \mathbb{E}[f(X) \,|\, \mathcal{G}_i])^2 \,\middle|\, \mathcal{F}_i\right],\end{aligned}$$

where we used Jensen's inequality on the last line. Taking expectations and using the tower property

$$\begin{aligned}\mathrm{Var}[f(X)] &= \sum_{i=1}^{n} \mathbb{E}\left[(Z_i - Z_{i-1})^2\right] \\ &\leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[(f(X) - \mathbb{E}[f(X) \,|\, \mathcal{G}_i])^2 \,\middle|\, \mathcal{F}_i\right]\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[(f(X) - \mathbb{E}[f(X) \,|\, \mathcal{G}_i])^2\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[(f(X) - \mathbb{E}[f(X) \,|\, \mathcal{G}_i])^2 \,\middle|\, \mathcal{G}_i\right]\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}[f(X) \,|\, \mathcal{G}_i]\right].\end{aligned}$$

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 3.2.5 (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,

$$\mathrm{Var}[f(X)] \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[(f(X) - f(X^{(i)}))^2].$$

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

$$\mathrm{Var}\left[f(X) \,|\, \mathcal{G}_i\right] = \frac{1}{2}\mathbb{E}[(f(X) - f(X^{(i)}))^2 \,|\, \mathcal{G}_i],$$

where we used the independence of the X_is and X_i's. Plugging back into Theorem 3.2.3 gives the claim. ∎

Our second consequence of Theorem 3.2.3 is a Poincaré-type inequality which relates the variance of a function to its *expected* "square gradient." Compare to the much weaker (3.2.5), which involves in each term a supremum rather than an expectation.

Theorem 3.2.6 (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*

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

Proof. By Lemma 2.4.11,

$$\mathrm{Var}\left[f(X) \,|\, \mathcal{G}_i\right] \leq \frac{1}{4} D_i f(X)^2.$$

Plugging back into Theorem 3.2.3 gives the claim. ∎

Remark 3.2.7. *For comparison, a version of the* Poincaré inequality *in one dimension asserts the following: let $f : [0, T] \to \mathbb{R}$ be continuously differentiable with $f(0) = f(T) = 0$, $\int_0^T f(x)^2 + f'(x)^2 \mathrm{d}x < +\infty$ and $\int_0^T f(x)\mathrm{d}x = 0$, then*

Poincaré inequality

$$\int_0^T f(x)^2 \mathrm{d}x \leq C \int_0^T f'(x)^2 \mathrm{d}x, \tag{3.2.7}$$

where the best possible C is $T^2/4\pi^2$ (see, e.g., [SS03, Chapter 3, Exercise 11]; this case is also known as Wirtinger's inequality*). We give a quick proof for $T = 1$ with the suboptimal $C = 1$. Note that $f(x) = \int_0^x f'(y)\mathrm{d}y$ so, by Cauchy-Schwarz (Theorem B.4.8),*

$$f(x)^2 \leq x \int_0^x f'(y)^2 \mathrm{d}y \leq \int_0^1 f'(y)^2 \mathrm{d}y.$$

The result follows by integration. Intuitively, for a function with mean 0 to have a large norm, it must have a large absolute derivative somewhere.

Example 3.2.8 (Longest common subsequence)**.** Let $X_1, \dots, X_{2n}$ be independent uniform random variables in $\{-1, +1\}$. Let Z be the length of the longest common subsequence in $(X_1, \dots, X_n)$ and $(X_{n+1}, \dots, X_{2n})$, that is,

$$\begin{aligned} Z = \max \big\{ k \,:\, & \exists 1 \le i_1 < i_2 < \cdots < i_k \le n \\ & \text{and } n+1 \le j_1 < j_2 < \cdots < j_k \le 2n \\ & \text{such that } X_{i_1} = X_{j_1}, X_{i_2} = X_{i_2}, \dots, X_{i_k} = X_{j_k} \big\}. \end{aligned}$$

Then, writing $Z = f(X_1, \dots, X_{2n})$, it follows that $\|D_i f\|_\infty \le 1$. Indeed, fix $\mathbf{x} = (x_1, \dots, x_{2n})$ and let $\mathbf{x}^{i,+}$ (respectively $\mathbf{x}^{i,-}$) be $\mathbf{x}$ where the i-th component is replaced with $+1$ (respectively -1). Assume without loss of generality that $f(\mathbf{x}^{i,-}) \le f(\mathbf{x}^{i,+})$. Then $|f(\mathbf{x}^{i,+}) - f(\mathbf{x}^{i,-})| \le 1$ 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 $\mathbf{x}$, we have $\|D_i f\|_\infty \le 1$. Finally, by the bounded differences inequality (Theorem 3.2.6),

$$\mathrm{Var}[Z] \le \frac{1}{4} \sum_{i=1}^{2n} \|D_i f\|_\infty^2 \le \frac{n}{2},$$

which is much better than the obvious $\mathrm{Var}[Z] \le \mathbb{E}[Z^2] \le n^2$. Note that we did not require any information about the expectation of Z. ◀

McDiarmid's inequality

The following powerful consequence of the Azuma-Hoeffding inequality is commonly referred to as the *method of bounded differences*. Compare to (3.2.6).

Theorem 3.2.9 (McDiarmid's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables where X_i is $\mathcal{X}_i$-valued for all i, and let $X = (X_1, \dots, X_n)$. 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 $\beta > 0$*

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

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

Proof of Theorem 3.2.9. As before, we let

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

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

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

and

$$A_i = \mathbb{E}\left[\inf_{y \in \mathcal{X}_i} f(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) - f(X) \,\middle|\, \mathcal{F}_{i-1}\right].$$

Indeed, since $\sigma(X_i)$ is independent of $\mathcal{F}_{i-1}$ and $\mathcal{G}_i$, by the role of independence (Lemma B.6.14)

$$\begin{aligned}
Z_i &= \mathbb{E}\left[f(X) \,|\, \mathcal{F}_i\right] \\
&\leq \mathbb{E}\left[\sup_{y \in \mathcal{X}_i} f(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) \,\middle|\, \mathcal{F}_i\right] \\
&= \mathbb{E}\left[\sup_{y \in \mathcal{X}_i} f(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) \,\middle|\, \mathcal{F}_{i-1}, X_i\right] \\
&= \mathbb{E}\left[\sup_{y \in \mathcal{X}_i} f(X_1, \ldots, X_{i-1}, y, X_{i+1}, \ldots, X_n) \,\middle|\, \mathcal{F}_{i-1}\right],
\end{aligned}$$

and similarly for the other direction. Moreover, by definition, $B_i - A_i \leq \|D_i f\|_\infty := c_i$. The Azuma-Hoeffding inequality 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 3.2.10 (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 $\sum_{i=1}^n \mathbf{1}_{B_i}$, 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_is 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_js are independent by construction and, moreover, letting $Z_{n,m} = f(X_1, \ldots, X_m)$ we have $\|D_i f\|_\infty \leq 1$. 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 3.2.11 (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 $a = (a_1, \ldots, a_k)$ be a fixed string 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

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

However the $\mathbf{1}_{E_i}$s are not independent. So we cannot use a Chernoff bound for Poisson trials (Theorem 2.4.7). Instead we use the fact that $N_n = f(X)$ where $\|D_i f\|_\infty \leq k$, as each X_i appears in at most k substrings of length k. By the method of bounded differences, for all $b > 0$,

$$\mathbb{P}\left[\left|N_n - (n-k+1)\left(\frac{1}{s}\right)^k\right| \geq bk\sqrt{n}\right] \leq 2e^{-2b^2}.$$

◀

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 next application is more striking and hints at connections to isoperimetric considerations (which we will not explore here).

Example 3.2.12 (Concentration of measure on the hypercube)**.** For $A \subseteq \{0,1\}^n$ a subset of the hypercube and $r > 0$, we let

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

be the points at ℓ^1 distance at most r from A. Fix $\varepsilon \in (0, 1/2)$ and assume that $|A| \geq \varepsilon 2^n$. Let λ_ε 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 set A. This is an example of the *concentration of measure phenomenon*.

Claim 3.2.13.

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

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

$$f(\mathbf{x}) = \inf_{\mathbf{a} \in A} \|\mathbf{x} - \mathbf{a}\|_1,$$

has $\|D_i f\|_\infty \leq 1$. Indeed changing one coordinate of $\mathbf{x}$ can increase the ℓ^1 distance to the closest point to $\mathbf{x}$ by at most 1; in the other direction, if a one-coordinate change were to decrease f by more than 1, reversing it would produce an increase of that same amount—a contradiction. Hence McDiarmid's inequality gives

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

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

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

or, rearranging and using our assumption on A,

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

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

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

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

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

■

Claim 3.2.13 is striking for two reasons: 1) the radius $2\lambda_\varepsilon\sqrt{n}$ is much smaller than n, the diameter of $\{0,1\}^n$; and 2) it applies to *any* A (such that $|A| \geq \varepsilon 2^n$). 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

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

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

$$\frac{1}{2^n}|B(\gamma)| = \sum_{\ell=0}^{\frac{n}{2}-\gamma\sqrt{\frac{n}{4}}} \binom{n}{\ell} 2^{-n} = \mathbb{P}\left[Y_n \leq \frac{n}{2} - \gamma\sqrt{\frac{n}{4}} \right]. \tag{3.2.8}$$

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 (3.2.8),

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

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 (3.2.8). 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 3.2.13 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 $r := 2$. Each point selected in A has $\binom{n}{2}$ points within ℓ^1 distance 2. By a union bound, the probability that A_r does not cover all of $\{0,1\}^n$ is at most

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

where, in the second inequality, we considered only the first N picks in the construction of A (possibly with repeats), and in the third inequality we used $1 - z \leq e^{-z}$ for all $z \in \mathbb{R}$ (see Exercise 1.16). In particular, as $n \to +\infty$,

$$\mathbb{P}[|\{0,1\}^n \backslash A_r| > 0] < 1.$$

So for n large enough there is a set A such that $A_r = \{0,1\}^n$ where $r = 2$. ◀

Remark 3.2.14. *In fact, it can be shown that sets of the form* $\{\mathbf{x} : \|\mathbf{x}\|_1 \leq s\}$ *have the smallest "expansion" among subsets of* $\{0,1\}^n$ *of the same size, a result known as Harper's vertex isoperimetric theorem. See, for example, [BLM13, Theorem 7.6 and Exercises 7.11-7.13].*

3.2.3 ▷ *Random graphs: exposure martingale and application to the chromatic number in Erdős-Rényi model*

Exposure martingales In the context of the Erdős-Rényi graph model (Definition 1.2.2), a common way to apply the Azuma-Hoeffding inequality (Theorem 3.2.1) is to introduce an "exposure martingale." Let $G \sim \mathbb{G}_{n,p}$ 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 adding the vertices of G one at a time (together with their edges to the previous vertices). The Doob martingale

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

is known as a *vertex exposure martingale*. An alternative way to define the filtration is to consider instead the random variables $X_i = (\mathbf{1}_{\{\{i,j\} \in G\}} : 1 \leq j \leq i)$ 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 with i and a vertex preceding it. Hence, $\mathcal{H}_i = \sigma(X_2, \ldots, X_i)$ for $i = 1, \ldots, n$ (and $\mathcal{H}_1$ is trivial as it corresponds to a graph with a single vertex and no edge). This representation has an important property: the X_is are *independent* as they pertain to disjoint subsets of edges. We are then in the setting of the method of bounded differences. Re-writing $F(G) = f(X_1, \ldots, X_n)$, the vertex exposure martingale coincides with the martingale (3.2.3) used in that context.

vertex exposure martingale

As an example, consider the chromatic number $\chi(G)$, that is, 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 $\|D_i f_\chi\|_\infty$.

Lemma 3.2.15. *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 with vertex i, Lemma 3.2.15 implies that $\|D_i f_\chi\|_\infty \leq 1$. Hence, for all $0 < p < 1$ and n, by an immediate application of the McDiarmid's inequality (Theorem 3.2.9):

Claim 3.2.16.

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

edge exposure martingale

Edge exposure martingales can be defined in a manner similar to the vertex case: reveal the edges one at a time in an arbitrary order. By Lemma 3.2.15, the corresponding function also satisfies the same ℓ^∞ bound. 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 3.2.17. *Note that Claim 3.2.16 tells us nothing about the expectation of $\chi(G)$. It turns out that, up to logarithmic factors, $\mathbb{E}_{n,p_n}[\chi(G)]$ 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 $\alpha > \frac{1}{2}$, the chromatic number $\chi(G)$ 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 3.2.1).

Claim 3.2.18. *Let $p_n = n^{-\alpha}$ with $\alpha > \frac{5}{6}$ and let $G_n \sim \mathbb{G}_{n,p_n}$. 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 φ_n be the smallest integer such that

$$\mathbb{P}_{n,p_n}[\chi(G_n) \leq \varphi_n] > \frac{\varepsilon}{3}. \tag{3.2.9}$$

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

We claim that (Z_i) has increments bounded by 1.

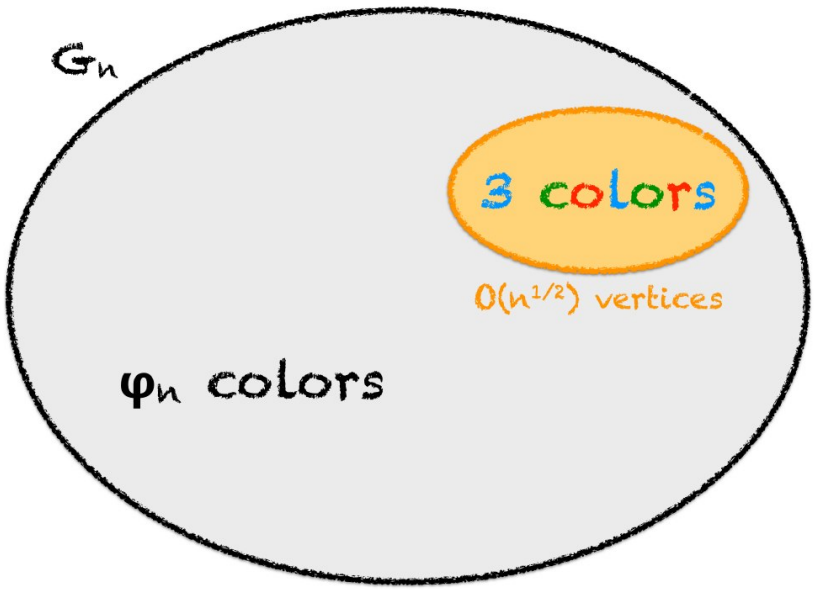

Figure 3.1: All but $O(\sqrt{n})$ vertices are colored using φ_n colors. The remaining vertices are colored using 3 additional colors.

Lemma 3.2.19. *Changing the edges incident to a single vertex can change F_n by at most* 1.

Proof. Changing the edges incident to v can increase F_n by at most 1. Indeed, if F_n increases after such a change, it must be that $v \notin U$ since in the other case the edges incident with v would not affect the colorability of $G_n \setminus U$—present or not. So we can add v to U and restore colorability. 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_ε such that $e^{-b_\varepsilon^2/2} = \frac{\varepsilon}{3}$. Then, applying the Azuma-Hoeffding inequality to $(-Z_i)$,

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

which, since $\mathbb{P}_{n,p_n}[F_n(G_n) = 0] = \mathbb{P}_{n,p_n}[\chi(G_n) \leq \varphi_n] > \frac{\varepsilon}{3}$, 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

$$\begin{aligned}
&\mathbb{P}_{n,p_n}\left[F_n(G_n) \geq 2b_\varepsilon\sqrt{n-1}\right] \\
&\quad\leq \mathbb{P}_{n,p_n}\left[F_n(G_n) - \mathbb{E}_{n,p_n}[F_n(G_n)] \geq b_\varepsilon\sqrt{n-1}\right] \\
&\quad\leq \frac{\varepsilon}{3}. \qquad (3.2.10)
\end{aligned}$$

So with probability at least $1 - \frac{\varepsilon}{3}$, we can color all vertices but $2b_\varepsilon\sqrt{n-1}$ using φ_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 3.2.20. *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,*

$$\mathbb{P}_{n,p_n}\left[\textit{every subset of } c\sqrt{n} \textit{ vertices of } G_n \textit{ can be 3-colored}\right] > 1 - \frac{\varepsilon}{3}. \quad (3.2.11)$$

Proof. We use the first moment method (Theorem 2.2.6). We refer to a subset of vertices that is not 3-colorable but such that all of its subsets are as minimal, non 3-colorable. Let Y_n be the number of such subsets of size at most $c\sqrt{n}$ in G_n.

Any minimal, non 3-colorable subset W must have degree at least 3. Indeed suppose that $w \in W$ has degree less than 3. Then $W\setminus\{w\}$ is 3-colorable by definition. 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. Hence, the probability that a subset of vertices of G_n of size ℓ is minimal, non 3-colorable is at most

$$\binom{\binom{\ell}{2}}{\frac{3\ell}{2}} p_n^{\frac{3\ell}{2}},$$

by a union bound over all subsets of edges of size $\frac{3\ell}{2}$.

By the first moment method, by the binomial bounds $\binom{n}{\ell} \leq \left(\frac{en}{\ell}\right)^\ell$ (see Appendix A) and $\binom{\ell}{2} \leq \ell^2/2$, for some $c' \in (0, +\infty)$

$$\begin{aligned}
\mathbb{P}_{n,p_n}[Y_n > 0] &\leq \mathbb{E}_{n,p_n} Y_n \\
&\leq \sum_{\ell=4}^{c\sqrt{n}} \binom{n}{\ell} \binom{\binom{\ell}{2}}{\frac{3\ell}{2}} p_n^{\frac{3\ell}{2}} \\
&\leq \sum_{\ell=4}^{c\sqrt{n}} \left(\frac{en}{\ell}\right)^\ell \left(\frac{e\ell}{3}\right)^{\frac{3\ell}{2}} n^{-\frac{3\ell\alpha}{2}} \\
&\leq \sum_{\ell=4}^{c\sqrt{n}} \left(\frac{e^{\frac{5}{2}} n^{1-\frac{3\alpha}{2}} \ell^{\frac{1}{2}}}{3^{\frac{3}{2}}}\right)^\ell \\
&\leq \sum_{\ell=4}^{c\sqrt{n}} \left(c' n^{\frac{5}{4}-\frac{3\alpha}{2}}\right)^\ell \\
&\leq O\left(n^{\frac{5}{4}-\frac{3\alpha}{2}}\right)^4 \\
&\to 0,
\end{aligned}$$

as $n \to +\infty$, where we used that $\frac{5}{4} - \frac{3\alpha}{2} < \frac{5}{4} - \frac{5}{4} = 0$ when $\alpha > \frac{5}{6}$ so that the geometric series is dominated by its first term. Therefore for n large enough $\mathbb{P}_{n,p_n}[Y_n > 0] \leq \varepsilon/3$, concluding the proof. ∎

By the choice of φ_n in (3.2.9),

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

By (3.2.10) and (3.2.11) with $c = 2b_\varepsilon$,

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

So, overall,

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

That concludes the proof. ∎

3.2.4 ▷ *Random graphs: degree sequence of preferential attachment graphs*

Let $(G_t)_{t\geq 1} \sim \mathrm{PA}_1$ be a preferential attachment graph (Definition 1.2.3). A key feature of such graphs is a power-law degree sequence: the fraction of vertices with degree d behaves like $\propto d^{-\alpha}$ for some $\alpha > 0$, that is, it has a fat tail. Recall that we restrict ourselves to the tree case. In contrast, we will show in Section 4.1.4 that a (sparse) Erdős-Rényi random graph has an asymptotically Poisson-distributed degree sequence, and therefore a much thinner tail.

Power law degree sequence Let $D_i(t)$ be the degree of the i-th vertex in G_t, denoted v_i, and let

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

be the number of vertices of degree d in G_t. By construction $N_0(t) = 0$ for all t. Define the sequence

$$f_d := \frac{4}{d(d+1)(d+2)}, \qquad d \geq 1. \tag{3.2.12}$$

Our main claim is:

Claim 3.2.21.

$$\frac{1}{t} N_d(t) \to_{\mathrm{p}} f_d, \qquad \forall d \geq 1.$$

Proof. The claim is immediately implied by the following lemmas.

Lemma 3.2.22 (Convergence of the mean)**.**

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

Lemma 3.2.23 (Concentration around the mean)**.** *For any* $\delta > 0$*,*

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

An alternative representation of the process We start with the proof of Lemma 3.2.23, which is an application of the method of bounded differences.

Proof of Lemma 3.2.23. 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 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 higher time index (but we ignore the orientations in the output). For a directed edge (i, j), we refer to i as its tail and j as its head. For all $s \geq 2$, let X_s be an independent, uniformly chosen edge extremity among the edges in G_{s-1}, that is, pick a uniform element in

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

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

We claim that $N_d(t) =: h(X_2, \ldots, X_t)$ as a function of $X_2, \ldots, X_t$ satisfies $\|D_i h\|_\infty \leq 2$. Indeed let $(x_2, \ldots, x_t)$ be a realization of $(X_2, \ldots, X_t)$ and let $y \in \mathcal{X}_s$ with $y \neq x_s$. Replacing $x_s = (i, \mathrm{end})$ with $y = (j, \mathrm{end}')$ where $i, j \in \{1, \ldots, s-1\}$ and $\mathrm{end}, \mathrm{end}' \in \{\mathrm{tail}, \mathrm{head}\}$ 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, head). But, crucially, those changes only affect the degrees of the vertices (i, end) and (j, end') in the original graph. Hence the number of vertices with degree d changes by at most 2, as claimed. For instance, returning to the example of Figure 3.2. If we replace $x_3 = (2, \mathrm{tail})$ with $y = (1, \mathrm{tail})$, 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 McDiarmid's inequality (Theorem 3.2.9), for all $\beta > 0$,

$$\mathbb{P}[|N_d(t) - \mathbb{E}N_d(t)| \geq \beta] \leq 2 \exp\left(-\frac{2\beta^2}{(2)^2(t-1)}\right),$$

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

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

That concludes the proof of the lemma. ∎

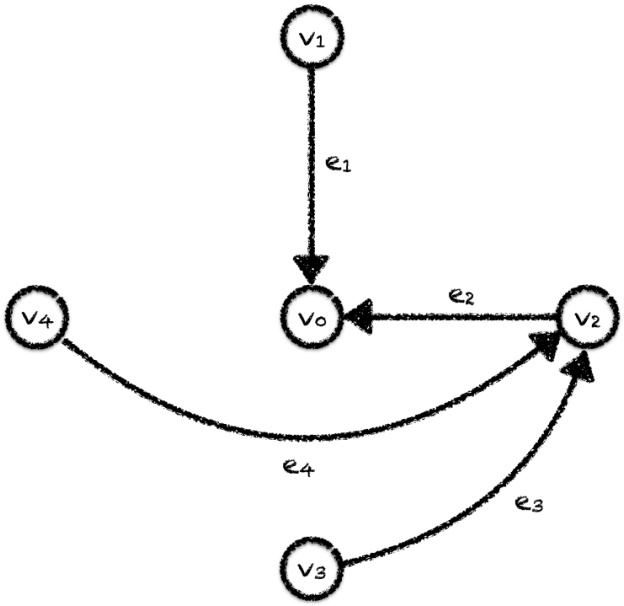

Figure 3.2: Graph obtained when $x_2 = (1, \text{head})$, $x_3 = (2, \text{tail})$ and $x_4 = (3, \text{head})$.

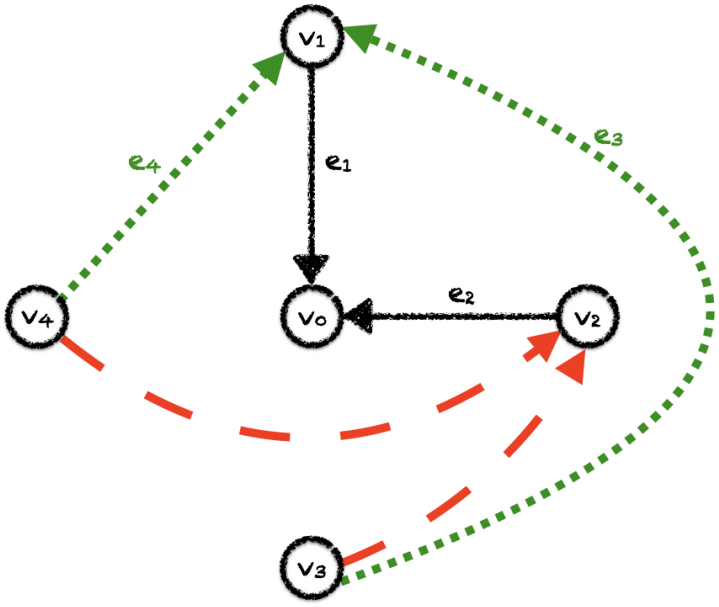

Figure 3.3: Substituting $x_3 = (2, \text{tail})$ with $y = (1, \text{tail})$ in the example of Figure 3.2 has the effect of replacing the dashed edges with the dotted edges. Note that only the degrees of vertices v_1 and v_2 are affected by this change.

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 3.2.22 does not rely on the Azuma-Hoeffding inequality but is given for completeness (and may be skipped).

Proof of Lemma 3.2.22. 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 $t \to +\infty$. Let $d \geq 1$. Observe that $\mathbb{E}N_d(t) = 0$ for $t \leq d-1$ since we need at least d edges to have a degree-d vertex. Moreover, by the description of the preferential attachment process, the following recursion holds for $t \geq d-1$

$$\mathbb{E}N_d(t+1) - \mathbb{E}N_d(t) = \underbrace{\frac{d-1}{2t}\mathbb{E}N_{d-1}(t)}_{\text{(a)}} - \underbrace{\frac{d}{2t}\mathbb{E}N_d(t)}_{\text{(b)}} + \underbrace{\mathbf{1}_{\{d=1\}}}_{\text{(c)}}. \tag{3.2.13}$$

Indeed: (a) $N_d(t)$ increases by 1 if a vertex of degree $d-1$ is picked, an event of probability $\frac{d-1}{2t}N_{d-1}(t)$ because the sum of degrees at time t is twice the number of edges (i.e., t); (b) $N_d(t)$ decreases by 1 if a vertex of degree d is picked, an event of probability $\frac{d}{2t}N_d(t)$; and (c) the last term comes from the fact that the new vertex always has degree 1. We rewrite (3.2.13) as

$$\begin{aligned}
\mathbb{E}N_d(t+1) &= \mathbb{E}N_d(t) + \frac{d-1}{2t}\mathbb{E}N_{d-1}(t) - \frac{d}{2t}\mathbb{E}N_d(t) + \mathbf{1}_{\{d=1\}} \\
&= \left(1 - \frac{d/2}{t}\right)\mathbb{E}N_d(t) + \left\{\frac{d-1}{2}\left[\frac{1}{t}\mathbb{E}N_{d-1}(t)\right] + \mathbf{1}_{\{d=1\}}\right\} \\
&=: \left(1 - \frac{d/2}{t}\right)\mathbb{E}N_d(t) + g_d(t),
\end{aligned} \tag{3.2.14}$$

where $g_d(t)$ is defined as the expression in curly brackets on the second line. We will not solve this recursion explicitly. Instead we seek to analyze its asymptotics, specifically we show that $\frac{1}{t}\mathbb{E}N_d(t) \to f_d$.

The key is to notice that the expression for $\mathbb{E}N_d(t+1)$ depends on $\frac{1}{t}\mathbb{E}N_{d-1}(t)$—so we work by induction on d. Because of the form of the recursion, the following technical lemma is what we need to proceed.

Lemma 3.2.24. *Let f, g be nonnegative functions of $t \in \mathbb{N}$ satisfying the following recursion*

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

with $g(t) \to g \in [0, +\infty)$ 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 3.2.21. We first conclude the proof of Lemma 3.2.22. First let $d = 1$. In that case, $g_1(t) = g_1 := 1$, $\alpha := 1/2$, and $t_0 := 1$. By Lemma 3.2.24,

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

Assuming by induction that $\frac{1}{t}\mathbb{E}N_{d-1}(t) \to f_{d-1}$ we get

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

as $t \to +\infty$. Using Lemma 3.2.24 with $\alpha := d/2$ and $t_0 := d-1$, we obtain

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

That concludes the proof of Lemma 3.2.22. ∎

To prove Claim 3.2.21, we combine Lemmas 3.2.22 and 3.2.23. Fix any $d, \delta, \varepsilon > 0$. Choose t' large enough that for all $t \geq t'$

$$\max\left\{\left|\frac{1}{t}\mathbb{E}N_d(t) - f_d\right|, \sqrt{\frac{2\log\delta^{-1}}{t}}\right\} \leq \varepsilon.$$

Then

$$\mathbb{P}\left[\left|\frac{1}{t}N_d(t) - f_d\right| \geq 2\varepsilon\right] \leq 2\delta,$$

for all $t \geq t'$. That proves convergence in probability. ∎

Proof of the technical lemma It remains to prove Lemma 3.2.24.

Proof of Lemma 3.2.24. By induction on t, we have

$$\begin{aligned}
f(t+1) &= \left(1 - \frac{\alpha}{t}\right) f(t) + g(t) \\
&= \left(1 - \frac{\alpha}{t}\right)\left[\left(1 - \frac{\alpha}{t-1}\right) f(t-1) + g(t-1)\right] + g(t) \\
&= \left(1 - \frac{\alpha}{t}\right) g(t-1) + g(t) + \left(1 - \frac{\alpha}{t}\right)\left(1 - \frac{\alpha}{t-1}\right) f(t-1) \\
&= \cdots \\
&= \sum_{i=0}^{t-t_0} g(t-i) \prod_{j=0}^{i-1}\left(1 - \frac{\alpha}{t-j}\right) + f(t_0) \prod_{j=0}^{t-t_0}\left(1 - \frac{\alpha}{t-j}\right),
\end{aligned}$$

or

$$f(t+1) = \sum_{s=t_0}^{t} g(s) \prod_{r=s+1}^{t} \left(1 - \frac{\alpha}{r}\right) + f(t_0) \prod_{r=t_0}^{t} \left(1 - \frac{\alpha}{r}\right), \tag{3.2.15}$$

where empty products are equal to 1. To guess the limit note that, for large s, $g(s)$ is roughly constant and that the product in the first term behaves like

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

So approximating the sum by an integral we get that $f(t+1) \approx \frac{gt}{\alpha+1}$, which is indeed consistent with the claim.

Formally, we use that there is a constant $\gamma = 0.577\ldots$ such that (see e.g. [LL10, Lemma 12.1.3])

$$\sum_{\ell=1}^{m} \frac{1}{\ell} = \log m + \gamma + \Theta(m^{-1}),$$

and that by a Taylor expansion, for $|z| \le 1/2$,

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

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

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

so, taking exponentials,

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

Hence

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

as $t \to +\infty$. Moreover

$$\begin{aligned}
\frac{1}{t}\sum_{s=\eta t}^{t} g(s) \prod_{r=s+1}^{t}\left(1-\frac{\alpha}{r}\right) &\leq \frac{1}{t}\sum_{s=\eta t}^{t}(g+\eta)\frac{s^\alpha}{t^\alpha}(1+\Theta(s^{-1})) \\
&\leq O(\eta) + (1+\Theta(t^{-1}))\frac{g}{t^{\alpha+1}}\sum_{s=\eta t}^{t} s^\alpha \\
&\leq O(\eta) + (1+\Theta(t^{-1}))\frac{g}{t^{\alpha+1}}\frac{(t+1)^{\alpha+1}}{\alpha+1} \\
&\to O(\eta) + \frac{g}{\alpha+1},
\end{aligned}$$

where we bounded the sum on the second line by an integral. Similarly,

$$\begin{aligned}
\frac{1}{t}\sum_{s=t_0}^{\eta t-1} g(s) \prod_{r=s+1}^{t}\left(1-\frac{\alpha}{r}\right) &\leq \frac{1}{t}\sum_{s=t_0}^{\eta t-1}(g+\eta)\frac{s^\alpha}{t^\alpha}(1+\Theta(s^{-1})) \\
&\leq \frac{\eta t}{t}(g+\eta)\frac{(\eta t)^\alpha}{t^\alpha}(1+\Theta(t_0^{-1})) \\
&\to O(\eta^{\alpha+1}).
\end{aligned}$$

Plugging these inequalities back into (3.2.15), 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. Letting $\eta \to 0$ concludes the proof. ∎

Remark 3.2.25. *A more quantitative result (uniform in t and d) can be derived. See, for example, [vdH17, Sections 8.5, 8.6]. See the same reference for a generalization beyond trees.*

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 3.2.1) to (multi-armed) bandit problems. These are meant as a simple model of sequential decision making with limited information where a fundamental issue is trading off between exploitation of actions that have done well in the past and exploration of actions that might perform better in the future. A typical application is online advertising, where one must decide which advertisement to display to the next visitor to a website.

In the simplest version of the (two-arm) *stochastic bandit* problem, there are two *unknown* reward distributions ν_1, ν_2 over $[0,1]$ with respective means $\mu_1 \neq \mu_2$. At each time $t = 1, \dots, n$, we request an independent sample from ν_{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 $\mu^* := \mu_1 \vee \mu_2$, our goal is to minimize

stochastic bandit

arm

$$\overline{R}_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} \mu_{I_t}\right], \tag{3.2.16}$$

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 (3.2.16) is taken over the choices $(I_t)_{t=1}^n$, which themselves depend on the random rewards $(Z_s)_{t=1}^n$. As indicated above, because ν_1 and ν_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.

pseudo-regret

One general approach that has proved effective in this type of problem is known as *optimism in the face of uncertainty*. Roughly speaking, we construct a set of plausible environments (in our case, the means of the reward distributions) that are consistent with observed data; then we make an optimal decision assuming that the true environment is the most favorable among them. A concrete implementation of this principle is the *Upper Confidence Bound (UCB)* algorithm, which we now describe. In words, we use a concentration inequality to build a confidence interval for each reward mean, and then we pick the arm with highest upper bound.

UCB

UCB algorithm

To state the algorithm formally, we will need some notation. For $i = 1, 2$, let $T_i(t)$ be the number of times arm i is pulled up to time t

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

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

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

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

$$\hat{\mu}_{i,s} = \frac{1}{s} \sum_{r \le s} X_{i,r}.$$

Since the $X_{i,s}$s are independent and $[0,1]$-valued by assumption, by Hoeffding's inequality (Theorem 2.4.10), for any $\beta > 0$

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

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

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

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

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

The argument above implies that the true mean μ_i has probability less than $1/t^{\alpha^2}$ of being higher than $\hat{\mu}_{i,T_i(t-1)} + \alpha \mathrm{H}(T_i(t-1), 1/t)$. The algorithm makes an "optimistic" decision: it chooses the higher of the two values.

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

Theorem 3.2.26 (Pseudo-regret of UCB)**.** *In the two-arm stochastic bandit problem where the rewards are in* $[0,1]$ *with distinct means,* α*-UCB with* $\alpha > 1$ *achieves*

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

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

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 Δ_*, then Hoeffding's inequality (Theorem 2.4.10) 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 $\Delta^* \frac{1}{\Delta_*^2} \log n + \frac{1}{n} \Delta_* n \approx \frac{1}{\Delta_*} \log n$. The UCB strategy, on the other hand, elegantly adapts to the gap Δ_* and the horizon n.

Analysis of the UCB algorithm

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

$$\begin{aligned}\overline{R}_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} \mu_{I_t}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{n} (\mu^* - \mu_{I_t})\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{n} \sum_{i=1,2} \mathbf{1}\{I_t = i\}\Delta_i\right] \\ &= \sum_{i=1,2} \Delta_i \mathbb{E}[T_i(n)]. \end{aligned} \tag{3.2.17}$$

Hence the problem boils down to bounding $\mathbb{E}[T_i(n)]$, the expected number of times that arm i is pulled. Note that $T_i(n)$ is a complicated function of the observations. To analyze it, we will use the following sufficient condition. Let i^* be the optimal arm, that is, the one that achieves μ^*. Intuitively, if arm $i \neq i^*$ is pulled, it is because: either our upper estimate of μ_{i^*} happens to be low or our lower estimate of μ_i happens to be high (i.e., our concentration inequality failed); or there is too much uncertainty in our estimate of μ_i (i.e., we haven't pulled arm i enough).

Lemma 3.2.27. *Under the α-UCB strategy, if arm $i \neq i^*$ is pulled at time t then at least one of the following events hold:*

$$\mathcal{E}_{t,1} = \{\hat{\mu}_{i^*,T_{i^*}(t-1)} + \alpha\,\mathrm{H}(T_{i^*}(t-1), 1/t) \leq \mu^*\}, \tag{3.2.18}$$

$$\mathcal{E}_{t,2} = \{\hat{\mu}_{i,T_i(t-1)} - \alpha\,\mathrm{H}(T_i(t-1), 1/t) > \mu_i\}, \tag{3.2.19}$$

$$\mathcal{E}_{t,3} = \left\{\alpha\,\mathrm{H}(T_i(t-1), 1/t) > \frac{\Delta_i}{2}\right\}. \tag{3.2.20}$$

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

$$\begin{aligned}\hat{\mu}_{i^*,T_{i^*}(t-1)} + \alpha\,\mathrm{H}(T_{i^*}(t-1), 1/t) &> \mu^* \\ &= \mu_i + \Delta_i \\ &\geq \hat{\mu}_{i,T_i(t-1)} + \alpha\,\mathrm{H}(T_i(t-1), 1/t).\end{aligned}$$

That implies that arm i would not be chosen. ∎

We first deal with $\mathcal{E}_{t,3}$. Let

$$u_n = \frac{2\alpha^2 \log n}{\Delta_*^2}.$$

Using the condition in Lemma 3.2.27, we get the following bound on $\mathbb{E}[T_i(n)]$.

Lemma 3.2.28. *Under the α-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 $i \neq i^*$, by definition of $T_i(n)$,

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

where we used that by Lemma 3.2.27

$$\{I_t = i\} \subseteq \mathcal{E}_{t,1} \cup \mathcal{E}_{t,2} \cup \mathcal{E}_{t,3}.$$

The condition in $\mathcal{E}_{t,3}$ can be written equivalently as

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

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 $T_i(t) = T_i(t-1) + 1$ whenever $I_t = i$, the event $\{I_t = i\} \cap \mathcal{E}_{t,3}$ can occur at most u_n times and

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

which proves the claim. ∎

It remains to bound $\mathbb{P}[\mathcal{E}_{t,1}]$ and $\mathbb{P}[\mathcal{E}_{t,2}]$ from above. This is not entirely straightforward because, while $\hat{\mu}_{i,T_i(t-1)}$ involves a sum of independent random variables, the number of terms $T_i(t-1)$ is itself a *random variable*. Moreover $T_i(t-1)$ depends on the past rewards Z_s, $s \leq t-1$, in a complex way. 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 (deterministic) values in the support of $T_i(t-1)$. That is,

$$\begin{aligned}\mathbb{P}[\mathcal{E}_{t,2}] &= \mathbb{P}[\hat{\mu}_{i,T_i(t-1)} - \alpha\,\mathrm{H}(T_i(t-1), 1/t) > \mu_i] \\ &\leq \mathbb{P}\left[\bigcup_{s\leq t-1} \{\hat{\mu}_{i,s} - \alpha\,\mathrm{H}(s, 1/t) > \mu_i\}\right]. \end{aligned} \tag{3.2.21}$$

We reformulate the previous bound as

$$\begin{aligned}&\mathbb{P}\left[\bigcup_{s\leq t-1} \{\hat{\mu}_{i,s} - \alpha\,\mathrm{H}(s, 1/t) > \mu_i\}\right] \\ &\quad= \mathbb{P}\left[\sup_{s\leq t-1} (\hat{\mu}_{i,s} - \mu_i - \alpha\,\mathrm{H}(s, 1/t)) > 0\right] \\ &\quad= \mathbb{P}\left[\sup_{s\leq t-1} \left(\frac{1}{s}\sum_{r\leq s} X_{i,r} - \mu_i - \alpha\sqrt{\frac{\log t}{2s}}\right) > 0\right] \\ &\quad= \mathbb{P}\left[\sup_{s\leq t-1} \frac{1}{\sqrt{s}}\left(\frac{1}{\sqrt{s}}\sum_{r\leq s}(X_{i,r} - \mu_i) - \alpha\sqrt{\frac{\log t}{2}}\right) > 0\right] \\ &\quad= \mathbb{P}\left[\sup_{s\leq t-1} \frac{\sum_{r=1}^{s}(X_{i,r} - \mu_i)}{\sqrt{s}} > \alpha\sqrt{\frac{\log t}{2}}\right]. \end{aligned} \tag{3.2.22}$$

Observe that the numerator on the left-hand side of the inequality on the last line is a martingale (see Example 3.1.29) with increments in $[-\mu_i, 1-\mu_i]$. But the denominator depends on s.

We try two approaches:

- We could simply use that $\sqrt{s} \geq 1$ on the denominator and apply the maximal

Azuma-Hoeffding inequality (Theorem 3.2.1) to get

$$\begin{aligned}\sum_{t=1}^{n}\mathbb{P}[\mathcal{E}_{t,2}] &\leq \sum_{t=1}^{n}\mathbb{P}\left[\sup_{s\leq t-1}\sum_{r=1}^{s}(X_{i,r}-\mu_i) > \alpha\sqrt{\frac{\log t}{2}}\right] \\ &\leq \sum_{t=1}^{n}\exp\left(-\frac{2(\alpha\sqrt{(\log t)/2})^2}{t-1}\right) \\ &\leq \sum_{t=1}^{n}\exp\left(-\alpha^2\frac{\log t}{t-1}\right). \end{aligned} \tag{3.2.23}$$

That is of order $\Theta(n)$ for any α.

- On the other hand, we could use a union bound over s and apply the maximal Azuma-Hoeffding inequality to each term to get

$$\begin{aligned}\sum_{t=1}^{n}\mathbb{P}[\mathcal{E}_{t,2}] &\leq \sum_{t=1}^{n}\sum_{s\leq t-1}\mathbb{P}\left[\sum_{r=1}^{s}(X_{i,r}-\mu_i) > \alpha\sqrt{\frac{s\log t}{2}}\right] \\ &\leq \sum_{t=1}^{n}\sum_{s\leq t-1}\exp\left(-\frac{2(\alpha\sqrt{(s\log t)/2})^2}{s}\right) \\ &= \sum_{t=1}^{n}(t-1)\exp\left(-\alpha^2\log t\right) \\ &\leq \sum_{t=1}^{n}\frac{1}{t^{\alpha^2-1}}. \end{aligned} \tag{3.2.24}$$

The series converges for $\alpha > \sqrt{2}$. Therefore, in that case, this bound is $\Theta(1)$, which is much better than our previous attempt. For $1 < \alpha \leq \sqrt{2}$ however, we get a bound of order $\Theta(n^{\alpha^2})$, which is worse than before.

It turns out that doing something "in between" the two approaches above gives a bound that significantly improves over both of them in the $1 < \alpha \leq \sqrt{2}$ regime. This is known as the slicing (or peeling) method.

Slicing method

slicing method

The *slicing method* is useful when bounding a weighted supremum. Its application is somewhat problem-specific so we will content ourselves with illustrating it in our case. Specifically, our goal is to control probabilities of the form

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

where $M_s := \sum_{r=1}^{s}(X_{i,r} - \mu_i)$, $w(s) := \sqrt{s}$, and $\beta := \alpha\sqrt{\frac{\log t}{2}}$. The idea is to divide up the supremum into *slices* $\gamma^{k-1} \le s < \gamma^k$, $k \ge 1$, where the constant $\gamma > 1$ will be optimized below. That is, fixing $K_t = \lceil \frac{\log t}{\log \gamma} \rceil$ (which roughly solves $\gamma^{K_t} = t$), by a union bound over the slices

$$\mathbb{P}\left[\sup_{1\le s<t} \frac{M_s}{w(s)} \ge \beta\right] \le \sum_{k=1}^{K_t} \mathbb{P}\left[\sup_{\gamma^{k-1}\le s<\gamma^k} \frac{M_s}{w(s)} \ge \beta\right].$$

Because $w(s)$ is increasing, on each slice separately we can bound

$$\begin{aligned}
\mathbb{P}\left[\sup_{\gamma^{k-1}\le s<\gamma^k} \frac{M_s}{w(s)} \ge \beta\right] &\le \mathbb{P}\left[\sup_{\gamma^{k-1}\le s<\gamma^k} \frac{M_s}{w(\gamma^{k-1})} \ge \beta\right] \\
&= \mathbb{P}\left[\sup_{\gamma^{k-1}\le s<\gamma^k} M_s \ge \beta w(\gamma^{k-1})\right] \\
&\le \mathbb{P}\left[\sup_{s\le\gamma^k} M_s \ge \beta w(\gamma^{k-1})\right].
\end{aligned}$$

Now we apply the maximal Azuma-Hoeffding inequality (Theorem 3.2.1) to obtain

$$\begin{aligned}
\mathbb{P}\left[\sup_{s\le\gamma^k} M_s \ge \beta w(\gamma^{k-1})\right] &\le \exp\left(-\frac{2(\beta w(\gamma^{k-1}))^2}{\gamma^k}\right) \\
&\le \exp\left(-\frac{2\beta^2}{\gamma}\right) \\
&= t^{-\alpha^2/\gamma},
\end{aligned}$$

where we used that $M_s - M_{s-1} = X_{i,s} - \mu_i \in [-\mu_i, 1-\mu_i]$, an interval of length 1. Plugging this back above we get

$$\mathbb{P}\left[\sup_{1\le s<t} \frac{M_s}{w(s)} \ge \beta\right] \le \left\lceil \frac{\log t}{\log \gamma} \right\rceil t^{-\alpha^2/\gamma}. \tag{3.2.25}$$

Now we see the tradeoff: increasing γ makes the slices larger and hence the tail inequality weaker, but it also makes the number of slices smaller which helps with the union bound.

Combining (3.2.21), (3.2.22), and (3.2.25), we have proved:

Lemma 3.2.29. *For any $\gamma > 1$, it holds that*

$$\sum_{t=1}^{n} \mathbb{P}[\mathcal{E}_{t,2}] \le \sum_{t=1}^{n} \left\lceil \frac{\log t}{\log \gamma} \right\rceil t^{-\alpha^2/\gamma},$$

and similarly for $\mathbb{P}[\mathcal{E}_{t,1}]$.

For $\alpha > 1$, we can choose $\gamma > 1$ such that $\alpha^2/\gamma > 1$. In that case, the series on the right-hand side is summable. This improves over both (3.2.23) and (3.2.24).

We are ready to prove the main result.

Proof of Theorem 3.2.26. By (3.2.17) and Lemmas 3.2.27, 3.2.28 and 3.2.29, we have

$$\overline{R}_n = \sum_{i=1,2} \Delta_i \mathbb{E}[T_i(n)] \leq \Delta_* \left(u_n + 2 \sum_{t=1}^{n} \left\lceil \frac{\log t}{\log \gamma} \right\rceil t^{-\alpha^2/\gamma} \right).$$

Recalling that $\alpha > 1$, choose $\gamma > 1$ such that $\alpha^2/\gamma > 1$. In that case, as noted above, the series on the right hand side is summable and there is $C_\alpha \in (0, +\infty)$ such that

$$\overline{R}_n \leq \Delta_*(u_n + C_\alpha).$$

That proves the claim. ∎

Remark 3.2.30. *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].*

3.2.6 Coda: Talagrand's inequality

We end this section with a celebrated concentration inequality that applies under weaker conditions than McDiarmid's inequality (Theorem 3.2.9)—but is *not* proved using the martingale method. It is known as *Talagrand's inequality*.

Bounds on $\|D_i f\|_\infty$ are often expressed in terms of a Lipschitz condition under an appropriate metric. Let $0 < c_i < +\infty$, $i = 1, \ldots, n$ and $\mathbf{c} = (c_1, \ldots, c_n)$. The $\mathbf{c}$-*weighted Hamming distance* is defined as

weighted Hamming distance

$$\rho_{\mathbf{c}}(\mathbf{x}, \mathbf{y}) := \sum_{i=1}^{n} c_i \mathbf{1}_{\{x_i \neq y_i\}},$$

for $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. The proof of the following equivalence is left as an exercise (see Exercise 3.8).

Lemma 3.2.31 (Lipschitz condition)**.** *A function $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ satisfies the Lipschitz condition*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq \rho_{\mathbf{c}}(\mathbf{x}, \mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n, \tag{3.2.26}$$

if and only if

$$\|D_i f\|_\infty \leq c_i, \quad \forall i.$$

Consider the following relaxed version of (3.2.26):

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \sum_{i=1}^{n} c_i(\mathbf{x}) \mathbf{1}_{\{x_i \neq y_i\}}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n, \tag{3.2.27}$$

where now $c_i(\mathbf{x})$ is a finite, positive function over $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. Notice the "one-sided" nature of this condition, in the sense that c_i depends on $\mathbf{x}$ but not on $\mathbf{y}$. A typical example where (3.2.27) is satisfied, but (3.2.26) is not, is given below.

We state Talagrand's inequality without proof.

Talagrand's inequality

Theorem 3.2.32 (Talagrand'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 (3.2.27) holds. Then $f(X)$ is sub-Gaussian with variance factor $\| \sum_{i \leq n} c_i^2 \|_\infty$. In fact, for all $\beta > 0$ the following upper and lower tail bounds hold*

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

and

$$\mathbb{P}[f(X) - \mathbb{E}f(X) \leq -\beta] \leq \exp\left(- \frac{\beta^2}{2\, \mathbb{E}\left[\sum_{i \leq n} c_i(X)^2 \right]} \right).$$

Compared to McDiarmid's inequality (Theorem 3.2.9), the upper tail in Theorem 3.2.32 has the sum over the coordinates *inside the supremum*, potentially a major improvement; the lower tail is even better, replacing the supremum with an *expectation*.

Example 3.2.33 (Spectral norm of a random matrix with bounded entries)**.** Let A be an $n \times n$ random matrix. We assume that the entries $A_{i,j}$, $i, j = 1, \ldots, n$, are independent, centered random variables in $[-1, 1]$. In Theorem 2.4.28, we proved an upper tail bound on the spectral norm

$$\|A\|_2 = \sup_{\mathbf{x} \in \mathbb{R}^n \setminus \{0\}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\substack{\mathbf{x} \in \mathbb{S}^{n-1} \\ \mathbf{y} \in \mathbb{S}^{n-1}}} \langle A\mathbf{x}, \mathbf{y} \rangle,$$

of such a matrix (in the more general sub-Gaussian case) using an ε-net argument. Theorem 2.4.28 also implies that $\mathbb{E}\|A\|_2 = O(\sqrt{n})$ by (B.5.1). (See Exercise 3.9 for a lower bound on the expectation.)

Here we use Talagrand's inequality (Theorem 3.2.32) directly to show concentration around the mean. For this, we need to check (3.2.27) where we think of the spectral norm as a function of n^2 independent random variables

$$\|A\|_2 = f(\{A_{i,j}\}_{i,j}).$$

Let $\mathbf{x}^*(A)$ and $\mathbf{y}^*(A)$ be unit vectors in $\mathbb{R}^n$ such that

$$\|A\|_2 = \langle A\mathbf{x}^*(A), \mathbf{y}^*(A)\rangle,$$

which exist by compactness.

Given two $n \times n$ matrices $A, \widetilde{A}$ with entries in $[-1, 1]$, we have

$$\begin{aligned}
\|A\|_2 - \|\widetilde{A}\|_2 &= \langle A\mathbf{x}^*(A), \mathbf{y}^*(A)\rangle - \sup_{\substack{\mathbf{x}\in\mathbb{S}^{n-1}\\ \mathbf{y}\in\mathbb{S}^{n-1}}} \langle \widetilde{A}\mathbf{x}, \mathbf{y}\rangle \\
&\le \langle A\mathbf{x}^*(A), \mathbf{y}^*(A)\rangle - \langle \widetilde{A}\mathbf{x}^*(A), \mathbf{y}^*(A)\rangle \\
&= \langle (A - \widetilde{A})\mathbf{x}^*(A), \mathbf{y}^*(A)\rangle \\
&\le \sum_{i,j} |A_{ij} - \widetilde{A}_{ij}||\mathbf{x}^*(A)_i||\mathbf{y}^*(A)_j| \\
&\le \sum_{i,j} \mathbf{1}_{A_{ij}\neq\widetilde{A}_{ij}} c_{ij}(A),
\end{aligned}$$

where on the last line we set

$$c_{ij}(A) := 2|\mathbf{x}^*(A)_i||\mathbf{y}^*(A)_j|,$$

and used the fact that $|A_{ij} - \widetilde{A}_{ij}| \le 2$. Note that

$$\sum_{i,j} c_{ij}(A)^2 = 4\sum_i \mathbf{x}^*(A)_i^2 \sum_j \mathbf{y}^*(A)_j^2 = 4.$$

Hence Talagrand's inequality implies that $\|A\|_2$ is sub-Gaussian with variance factor 4. ◀

3.3 Potential theory and electrical networks

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

Let (X_t) be a Markov chain with transition matrix P on a finite or countable state space V. Recall from Definition 3.1.6 that τ_B is the first visit time to $B \subseteq V$. For two disjoint subsets A, Z of V, the probability of hitting A before Z

$$h(x) = \mathbb{P}_x[\tau_A < \tau_Z], \tag{3.3.1}$$

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

harmonic function

$$h(x) = \sum_y P(x,y)h(y), \qquad \forall x \in W. \tag{3.3.2}$$

Indeed note that $h = 1$ (respectively $= 0$) on A (respectively Z) and by the Markov property (Theorem 1.1.18), after the first step of the chain, for $x \in W$

$$\begin{aligned} \mathbb{P}_x[\tau_A < \tau_Z] &= \sum_{y \notin A \cup Z} P(x,y)\, \mathbb{P}_y[\tau_A < \tau_Z] \\ &\qquad + \sum_{y \in A} P(x,y) \cdot 1 + \sum_{y \in Z} P(x,y) \cdot 0 \\ &= \sum_y P(x,y)\, \mathbb{P}_y[\tau_A < \tau_Z]. \end{aligned} \tag{3.3.3}$$

Quantities such as (3.3.1) 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 as we outline in this section. It turns out that harmonic functions and martingales are closely related. In Section 3.3.1 we elaborate on that connection.

But first we rewrite (3.3.2) 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 π, then the edge weights are $c(x,y) = \pi(x)P(x,y)$. In this notation (3.3.2) becomes

$$h(x) = \frac{1}{c(x)} \sum_{y \sim x} c(x,y)h(y), \qquad \forall x \in (A \cup Z)^c, \tag{3.3.4}$$

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 (3.3.4) is known as a voltage. 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. We develop the electrical analogy and many of its applications in Section 3.3.2.

3.3.1 Martingales, the Dirichlet problem and Lyapounov functions

To see why martingales come in, let $\mathcal{F}_t = \sigma(X_0, \ldots, X_t)$ and let $\tau^* := \tau_{W^c}$. By a first-step calculation again, for a function h satisfying (3.3.2),

$$h(X_{t\wedge\tau^*}) = \mathbb{E}\left[h(X_{(t+1)\wedge\tau^*}) \,|\, \mathcal{F}_t\right], \qquad \forall t \geq 0, \tag{3.3.5}$$

that is, $(h(X_{t\wedge\tau^*}))_t$ is a martingale with respect to $(\mathcal{F}_t)$. Indeed, on $\{\tau^* \leq t\}$,

$$\mathbb{E}[h(X_{(t+1)\wedge\tau^*}) \,|\, \mathcal{F}_t] = h(X_{\tau^*}) = h(X_{t\wedge\tau^*}),$$

and on $\{\tau^* > t\}$

$$\mathbb{E}[h(X_{(t+1)\wedge\tau^*}) \,|\, \mathcal{F}_t] = \sum_y P(X_t, y)h(y) = h(X_t) = h(X_{t\wedge\tau^*}).$$

Although the rest of Section 3.3 is concerned with reversible Markov chains, the current subsection applies to the non-reversible case as well. We give an overview of potential theory for general, countable-space, discrete-time Markov chains and its connections to martingales. As a major application, we introduce the concept of a Lyapounov function which is useful in bounding certain hitting times.

Existence and uniqueness of a harmonic extension

We begin with a special case, which will be generalized below.

Theorem 3.3.1 (Harmonic extension: existence and uniqueness)**.** *Let P be an irreducible transition matrix on a finite or countably infinite 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. Then there exists a unique extension of h to W that is harmonic on W, that is, which satisfies (3.3.2). The solution is given by*

$$h(x) = \mathbb{E}_x\left[h\left(X_{\tau_{W^c}}\right)\right].$$

Proof. We first argue about uniqueness. Suppose h is defined over all of V and satisfies (3.3.2). Let $\tau^* := \tau_{W^c}$. Then the process $\left(h\left(X_{t\wedge\tau^*}\right)\right)_t$ is a martingale by (3.3.5). Because W is finite and the chain is irreducible, we have $\tau^* < +\infty$ almost surely, as implied by Lemma 3.1.25. Moreover the process is bounded because h is bounded on W^c and W is finite. Hence by Doob's optional stopping theorem (Theorem 3.1.38 (ii))

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

which implies that h is unique, since the right-hand side depends only on the chain and the fixed values of h on W^c.

For the existence, simply define $h(x) := \mathbb{E}_x[h(X_{\tau^*})], \forall x \in W$, and use a first-step argument similarly to (3.3.3). ■

For some insights on what happens when the assumptions of Theorem 3.3.1 are not satisfied, see Exercise 3.11. For an alternative (arguably more intuitive) proof of uniqueness based on the maximum principle, see Exercise 3.12.

In the proof above it suffices to specify h on the outer boundary of W

$$\partial_{\mathrm{V}} W = \{z \in V \backslash W \, : \, \exists y \in W, P(y,z) > 0\}.$$

Introduce the *Laplacian* associated to P

Laplacian

$$\begin{aligned} \Delta f(x) &= \left[\sum_y P(x,y) f(y)\right] - f(x) \\ &= \sum_y P(x,y)[f(y) - f(x)] \\ &= \mathbb{E}_x[f(X_1) - f(X_0)], \end{aligned} \tag{3.3.6}$$

provided the expectation exists. We have proved that, under the assumptions of Theorem 3.3.1, there exists a unique solution to

$$\begin{cases} \Delta f(x) = 0 & \forall x \in W, \\ f(x) = h(x) & \forall x \in \partial_{\mathrm{V}} W, \end{cases} \tag{3.3.7}$$

and that solution is given by $f(x) = \mathbb{E}_x[h(X_{\tau_{W^c}})]$, for $x \in W \cup \partial_{\mathrm{V}} W$. The system (3.3.7), in reference to its counterpart in the theory of partial differential equations, is referred to as a *Dirichlet problem*.

Dirichlet problem

Example 3.3.2 (Simple random walk on $\mathbb{Z}^d$)**.** The Laplacian above can be interpreted as a discretized version of the standard Laplacian. For instance, for simple random walk on $\mathbb{Z}$,

$$\begin{aligned} \Delta f(x) &= \left[\sum_y P(x,y) f(y)\right] - f(x) \\ &= \sum_y P(x,y)[f(y) - f(x)] \\ &= \frac{1}{2}\{[f(x+1) - f(x)] - [f(x) - f(x-1)]\}, \end{aligned}$$

which is a discretized second derivative. More generally, for simple random walk on $\mathbb{Z}^d$, we get

$$\begin{aligned}\Delta f(x) &= \left[\sum_y P(x,y)f(y)\right] - f(x) \\ &= \sum_y P(x,y)[f(y) - f(x)] \\ &= \frac{1}{2d}\sum_{i=1}^{d}\{[f(x+\mathbf{e}_i) - f(x)] - [f(x) - f(x-\mathbf{e}_i)]\},\end{aligned}$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_d$ is the standard basis in $\mathbb{R}^d$. ◀

Theorem 3.3.1 has many applications. One of its consequences is that harmonic functions on a finite state space are constant.

Corollary 3.3.3. *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 $W = V\backslash\{z\}$. Applying Theorem 3.3.1, for all $x \in W$, $h(x) = \mathbb{E}_x[h\left(X_{\tau_{W^c}}\right)] = h(z)$. ■

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

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

does not in fact depend on x.

Proof. Because the chain is irreducible and has a finite state space, $\mathbb{E}_x[\tau_y] < +\infty$ for all x, y. By Corollary 3.3.3, 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,

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

Substituting into the definition of $h(x)$ gives

$$\begin{aligned} h(x) &= (1-\pi(x)) + \sum_z \sum_{y \neq x} \pi(y) P(x,z)\, \mathbb{E}_z[\tau_y] \\ &= (1-\pi(x)) + \sum_z P(x,z)\left(h(z) - \pi(x)\, \mathbb{E}_z[\tau_x]\right). \end{aligned}$$

Rearranging, we get

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

where we used $1/\pi(x) = \mathbb{E}_x[\tau_x^+] = 1 + \sum_z P(x,z)\, \mathbb{E}_z[\tau_x]$ by Theorem 3.1.19 and a first-step argument (recall that the first return time τ_x^+ was defined in Definition 3.1.6). ■

Potential theory for Markov chains

More generally, many quantities of interest can be expressed in the following form. Consider again a subset $W \subset V$ and the stopping time

$$\tau_{W^c} = \inf\{t \geq 0 \,:\, X_t \in W^c\}.$$

Let also $h \,:\, W^c \to \mathbb{R}_+$ and $k \,:\, W \to \mathbb{R}_+$. Define the quantity

$$u(x) := \mathbb{E}_x\left[h(X_{\tau_{W^c}}) \mathbf{1}\{\tau_{W^c} < +\infty\} + \sum_{0 \leq t < \tau_{W^c}} k(X_t)\right]. \tag{3.3.8}$$

The first term on the right-hand side is a final cost incurred when we exit W (and depends on where we do), while the second term is a unit time cost incurred along the sample path. Note that, in fact, it suffices to define h on $\partial_{\mathrm{V}} W$, the outer boundary of W if we restrict ourselves to $x \in W$. Observe also that the function $u(x)$ may take the value $+\infty$; the expectation is well-defined (in $\mathbb{R}_+ \cup \{+\infty\}$) by the nonnegativity of the terms (see Appendix B).

Example 3.3.5 (Some special cases)**.** Here are some important special cases:

- Revisiting (3.3.1), for two disjoint subsets A, Z of V, the probability
$$u(x) := \mathbb{P}_x[\tau_A < \tau_Z],$$
of hitting A before Z as a function of the starting point $x \in V$ is obtained by taking $W := (A \cup Z)^c$, $h = 1$ (respectively $= 0$) on A (respectively Z), and $k = 0$ on V. The further special case $Z = \emptyset$ leads to the *exit probability* from A *exit probability*
$$u(x) := \mathbb{P}_x[\tau_A < +\infty].$$
On the other hand, if A and Z form a disjoint partition of W^c (or $\partial_V W$ will suffice if $x \in W$), we get the *exit law* from W *exit law*
$$u(x) := \mathbb{P}_x[X_{\tau_{W^c}} \in A; \tau_{W^c} < +\infty].$$

- The *average occupation time* of $A \subseteq W$ before exiting W *average occupation time*
$$u(x) := \mathbb{E}_x \left[\sum_{0 \le t < \tau_{W^c}} \mathbf{1}_{\{X_t \in A\}} \right],$$
is obtained by taking $h = 0$ on V, and $k = 1$ (respectively $= 0$) on A (respectively on A^c). Revisiting (3.1.3), the Green function of the chain stopped at τ_{W^c}, that is,
$$u(x) := \mathscr{G}_{\tau_{W^c}}(x, y) = \mathbb{E}_x \left[\sum_{0 \le t < \tau_{W^c}} \mathbf{1}_{\{X_t = y\}} \right],$$
is obtained by taking $A = \{y\}$. Another special case is $A = W$ where we get the *mean exit time* from A *mean exit time*
$$u(x) := \mathbb{E}_x \left[\tau_{A^c} \right].$$

◀

The function u in (3.3.8) turns out to satisfy a generalized version of (3.3.7). The proof is usually called *first-step analysis* (of which we have already seen many instances). *first-step analysis*

Theorem 3.3.6 (First-step analysis)**.** *Let P be a transition matrix on a finite or countable state space V. Let W be a proper subset of V, and let $h : W^c \to \mathbb{R}_+$ and $k : W \to \mathbb{R}_+$ be bounded functions. Then the function $u \ge 0$, as defined in (3.3.8), satisfies the system of equations*
$$\begin{cases} u(x) = k(x) + \sum_y P(x, y) u(y) & \text{for } x \in W, \\ u(x) = h(x) & \text{for } x \in W^c. \end{cases} \tag{3.3.9}$$

Proof. For $x \in W^c$, by definition $u(x) = h(x)$ since $\tau_{W^c} = 0$. Fix $x \in W$. By taking out what is known (Lemma B.6.13), the tower property (Lemma B.6.16) and the Markov property (Theorem 1.1.18),

$$\begin{aligned} u(x) &= k(x) + \mathbb{E}_x \left[h(X_{\tau_{W^c}}) \mathbf{1}\{\tau_{W^c} < +\infty\} + \sum_{1 \le t < \tau_{W^c}} k(X_t) \right] \\ &= k(x) + \mathbb{E}_x \left[\mathbb{E} \left[h(X_{\tau_{W^c}}) \mathbf{1}\{\tau_{W^c} < +\infty\} + \sum_{1 \le t < \tau_{W^c}} k(X_t) \,\middle|\, \mathcal{F}_1 \right] \right] \\ &= k(x) + \mathbb{E}_x \left[u(X_1) \right], \end{aligned}$$

which gives the claim. ∎

If u is finite, the system of equations (3.3.9) can be rewritten as the *Poisson equation* (once again as an analogue of its counterpart in the theory of partial differential equations)

Poisson equation

$$\begin{cases} \Delta u = -k & \text{on } W, \\ u = h & \text{on } W^c. \end{cases} \tag{3.3.10}$$

This is well-defined for instance if W is a finite subset and P is irreducible. Indeed, as we argued in the proof of Theorem 3.3.1, the stopping time τ_{W^c} then has a finite expectation. Because h is bounded, it follows that

$$\begin{aligned} u(x) &:= \mathbb{E}_x \left[h(X_{\tau_{W^c}}) \mathbf{1}\{\tau_{W^c} < +\infty\} + \sum_{0 \le t < \tau_{W^c}} k(X_t) \right] \\ &\le \sup_{x \in W^c} h(x) + \sup_{x \in W} k(x) \sup_{x \in W} \mathbb{E}_x \left[\tau_{W^c} \right] \\ &< +\infty, \end{aligned}$$

uniformly in x. Using (3.3.6) and rearranging (3.3.9) gives (3.3.10).

Remark 3.3.7. *A more general form of the statement which can be used to study certain moment-generating functions can be found, for example, in [Ebe, Theorem 1.3].*

In a generalization of Theorem 3.3.1, our next theorem allows one to establish uniqueness of the solution of the system (3.3.10) under some conditions (which we will not detail here, but see Exercise 3.13). Perhaps even more useful, it also gives an effective approach to bound the function u from above. This is based on the following supermartingale.

Lemma 3.3.8 (Locally superharmonic functions)**.** *Let P be a transition matrix on a finite or countable state space V. Let W be a proper subset of V, and let $h : W^c \to \mathbb{R}_+$ and $k : W \to \mathbb{R}_+$ be bounded functions. Suppose the nonnegative function $\psi : V \to \mathbb{R}_+$ satisfies*

$$\Delta\psi \leq -k \qquad \text{on } W.$$

Then the process

$$N_t := \psi(X_{t\wedge\tau_{W^c}}) + \sum_{0\leq s<t\wedge\tau_{W^c}} k(X_s),$$

is a nonnegative supermartingale for any initial point $x \in V$.

Proof. Observe that: on $\{\tau_{W^c} \leq t\}$, we have $N_{t+1} = N_t$; while on $\{\tau_{W^c} > t\}$, we have $N_{t+1} - N_t = \psi(X_{t+1}) - \psi(X_t) + k(X_t)$ by cancellations in the sum. So, since $\{\tau_{W^c} > t\} \in \mathcal{F}_t$ by definition of a stopping time, it holds by taking out what is known that

$$\begin{aligned}
\mathbb{E}[N_{t+1} - N_t \,|\, \mathcal{F}_t] &= \mathbb{E}[\mathbf{1}\{\tau_{W^c} > t\}(\psi(X_{t+1}) - \psi(X_t) + k(X_t)) \,|\, \mathcal{F}_t] \\
&= \mathbf{1}\{\tau_{W^c} > t\}(\mathbb{E}[\psi(X_{t+1}) - \psi(X_t) \,|\, \mathcal{F}_t] + k(X_t)) \\
&= \mathbf{1}\{\tau_{W^c} > t\}(\Delta\psi(X_t) + k(X_t)) \\
&\leq \mathbf{1}\{\tau_{W^c} > t\}(-k(X_t) + k(X_t)) \\
&= 0,
\end{aligned}$$

where we used that, by (3.3.6) and the Markov property,

$$\mathbb{E}[\psi(X_{t+1}) - \psi(X_t) \,|\, \mathcal{F}_t] = \Delta\psi(X_t), \tag{3.3.11}$$

and that $X_t \in W$ on $\{\tau_{W^c} > t\}$. ∎

Theorem 3.3.9 (Poisson equation: bounding the solution)**.** *Let P be a transition matrix on a finite or countable state space V. Let W be a proper subset of V, and let $h : W^c \to \mathbb{R}_+$ and $k : W \to \mathbb{R}_+$ be bounded functions. Suppose the nonnegative function $\psi : V \to \mathbb{R}_+$ satisfies the system of inequalities*

$$\begin{cases} \Delta\psi \leq -k & \text{on } W, \\ \psi \geq h & \text{on } W^c. \end{cases} \tag{3.3.12}$$

Then

$$\psi \geq u, \quad \text{on } V, \tag{3.3.13}$$

where u is the function defined in (3.3.8)*.*

Proof. The system (3.3.13) holds on W^c by Theorem 3.3.6 and (3.3.12) since in that case $u(x) = h(x) \le \psi(x)$.

Fix $x \in W$. Consider the nonnegative supermartingale (N_t) in Lemma 3.3.8. By the convergence of nonnegative supermartingales (Corollary 3.1.48), (N_t) converges almost surely to a finite limit with expectation $\le \mathbb{E}_x[N_0]$. In particular, the limit $N_{\tau_{W^c}}$ is well-defined, nonnegative and finite, including on the event that $\{\tau_{W^c} = +\infty\}$. As a result,

$$\begin{aligned} N_{\tau_{W^c}} &= \lim_t \left[\psi(X_{t\wedge\tau_{W^c}}) + \sum_{0\le s<t\wedge\tau_{W^c}} k(X_s) \right] \\ &\ge h(X_{\tau_{W^c}})\mathbf{1}\{\tau_{W^c} < +\infty\} + \sum_{0\le s<\tau_{W^c}} k(X_s), \end{aligned}$$

where we used (3.3.12). Moreover, by Lemma 3.1.37, $\mathbb{E}_x\left[N_{t\wedge\tau_{W^c}}\right] \le \mathbb{E}_x[N_0]$ for all t and Fatou's lemma (see Proposition B.4.14) gives $\mathbb{E}_x\left[N_{\tau_{W^c}}\right] \le \mathbb{E}_x[N_0]$.

Hence, by definition of u,

$$\begin{aligned} u(x) &= \mathbb{E}_x \left[h(X_{\tau_{W^c}})\mathbf{1}\{\tau_{W^c} < +\infty\} + \sum_{0\le t<\tau_{W^c}} k(X_t) \right] \\ &\le \mathbb{E}_x\left[N_{\tau_{W^c}}\right] \\ &\le \mathbb{E}_x[N_0] \\ &= \psi(x), \end{aligned}$$

where, on the last line, we used that the initial state is $x \in W$. That proves the claim. ∎

Lyapounov functions

Here is an important application, bounding from above the hitting time τ_A to a set A in expectation.

Theorem 3.3.10 (Controlling hitting times via Lyapounov functions)**.** *Let P be a transition matrix on a finite or countably infinite state space V. Let A be a proper subset of V. Suppose the nonnegative function $\psi : V \to \mathbb{R}_+$ satisfies the system of inequalities*

$$\Delta\psi \le -1, \quad \text{on } A^c. \tag{3.3.14}$$

Then

$$\mathbb{E}_x\left[\tau_A\right] \le \psi(x),$$

for all $x \in V$.

Proof. Indeed, by (3.3.14) and nonnegativity (in particular on A), the function ψ satisfies the assumptions of Theorem 3.3.9 with $W = A^c$, $h = 0$ on A, and $k = 1$ on A^c. Hence, by definition of u and the claim in Theorem 3.3.9,

$$\begin{aligned}\mathbb{E}_x[\tau_A] &= \mathbb{E}_x\left[h(X_{\tau_A})\mathbf{1}\{\tau_A < +\infty\} + \sum_{0\le t<\tau_A} k(X_t)\right]\\ &= u(x)\\ &\le \psi(x).\end{aligned}$$

That establishes the claim. ∎

Recalling (3.3.11), condition (3.3.14) is equivalent to the following conditional expected decrease in ψ outside A:

$$\mathbb{E}[\psi(X_{t+1}) - \psi(X_t)\,|\,\mathcal{F}_t] \le -1, \quad \text{on } \{X_t \in A^c\}. \tag{3.3.15}$$

A nonnegative function satisfying an inequality of this type, also known as drift condition, is often referred to as a *Lyapounov function*. Intuitively, it tends to decrease along the sample path outside of A. Because it is non-negative, it cannot decrease forever and therefore the chain eventually enters A. We consider a simple example next.

Lyapounov function

Example 3.3.11 (A Markov chain on the nonnegative integers)**.** Let $(Z_t)_{t\ge1}$ be i.i.d. integrable random variables taking values in $\mathbb{Z}$ such that $\mathbb{E}[Z_1] < 0$. Let $(X_t)_{t\ge0}$ be the chain defined by $X_0 = x$ for some $x \in \mathbb{Z}_+$ and

$$X_{t+1} = (X_t + Z_{t+1})^+,$$

where recall that $z^+ = \max\{0, z\}$. In particular $X_t \in \mathbb{Z}_+$ for all t. Let $(\mathcal{F}_t)$ be the corresponding filtration. When X_t is large, the "local drift" is close to $\mathbb{E}[Z_1] < 0$. By analogy to the biased case of the gambler's ruin (Example 3.1.43), we might expect that, from a large starting point x, it will take time roughly $x/|\mathbb{E}[Z_1]|$ in expectation to "return to a neighborhood of 0." We prove something along those lines here using a Lyapounov function.

Observe that, for any $y \in \mathbb{Z}_+$, we have on the event $\{X_t = y\}$ by the Markov property

$$\begin{aligned}\mathbb{E}_x[X_{t+1} - X_t\,|\,\mathcal{F}_t] &= \mathbb{E}[(y + Z_{t+1})^+ - y]\\ &= \mathbb{E}[-y\mathbf{1}\{Z_{t+1} \le -y\} + Z_{t+1}\mathbf{1}\{Z_{t+1} > -y\}]\\ &\le \mathbb{E}[Z_{t+1}\mathbf{1}\{Z_{t+1} > -y\}]\\ &= \mathbb{E}[Z_1\mathbf{1}\{Z_1 > -y\}].\end{aligned} \tag{3.3.16}$$

For all y, the random variable $|Z_1\mathbf{1}\{Z_1 > -y\}|$ is bounded by $|Z_1|$, itself an integrable random variable. Moreover, $Z_1\mathbf{1}\{Z_1 > -y\} \to Z_1$ as $y \to +\infty$ almost surely. Hence, the dominated convergence theorem (Proposition B.4.14) implies that

$$\lim_{y\to+\infty} \mathbb{E}[Z_1\mathbf{1}\{Z_1 > -y\}] = \mathbb{E}[Z_1] < 0.$$

So for any $0 < \varepsilon < -\mathbb{E}[Z_1]$, there is $y_\varepsilon \in \mathbb{Z}_+$ large enough that $\mathbb{E}[Z_1\mathbf{1}\{Z_1 > -y\}] < -\varepsilon$ for all $y > y_\varepsilon$. Fix ε as above and define

$$A := \{0, 1, \ldots, y_\varepsilon\}.$$

We use Theorem 3.3.10 to bound τ_A in expectation. Define the Lyapounov function

$$\psi(x) = \frac{x}{\varepsilon}, \quad \forall x \in \mathbb{Z}_+.$$

On the event $\{X_t = y\}$, we rewrite (3.3.16) as

$$\begin{aligned}\mathbb{E}[\psi(X_{t+1}) - \psi(X_t) \,|\, \mathcal{F}_t] &\le \frac{\mathbb{E}[Z_1\mathbf{1}\{Z_1 > -y\}]}{\varepsilon} \\ &\le -1,\end{aligned}$$

for $y \in A^c$. This is the same as (3.3.15). Hence, we can apply Theorem 3.3.10 to get

$$\mathbb{E}_x[\tau_A] \le \psi(x) = \frac{x}{\varepsilon},$$

for all $x \ge y_\varepsilon$. ◀

A well-known, closely related result gives a criterion for positive recurrence. We state it without proof.

Theorem 3.3.12 (Foster's theorem)**.** *Let P be an irreducible transition matrix on a countable state space V. Let A be a finite, proper subset of V. Suppose the nonnegative function $\psi : V \to \mathbb{R}_+$ satisfies the system of inequalities*

$$\Delta\psi \le -1, \quad \text{on } A^c,$$

as well as the condition

$$\sum_{y\in V} P(x, y)\psi(y) < +\infty, \quad \forall x \in A.$$

Then P is positive recurrent.

3.3.2 Basic electrical network theory

We now develop the basic theory of electrical networks for the analysis of random walks. All results in this subsection (and the next one) concern *reversible* Markov chains, or random walks on networks (see Definition 1.2.7). We begin with a few definitions. Throughout, we will use the notation $h|_B$ for the function h restricted to the subset B. We also write $h \equiv c$ if h is identically equal to the constant c.

Definitions

Let $\mathcal{N} = (G, c)$ be a finite or countable network with $G = (V, E)$. 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 $r(e) := 1/c(e)$, for all $e \in E$. For an edge $e = \{x, y\}$ we overload $c(x, y) := c(e)$ and $r(x, y) := r(e)$. Both c and r are symmetric as functions of x, y. Recall that the transition matrix of the random walk on $\mathcal{N}$ satisfies

conductance
resistance

$$P(x, y) = \frac{c(x, y)}{c(x)},$$

where

$$c(x) = \sum_{z : z \sim x} c(x, z).$$

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, that is, $A = \{a\}$ 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 3.3.1, 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.

source, sink

Corollary 3.3.13 (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 *defined as follows: a function v on V such that v is harmonic on W, that is,*

voltage

$$v(x) = \frac{1}{c(x)} \sum_{y : y \sim x} c(x, y) v(y), \qquad \forall x \in W, \tag{3.3.17}$$

where

$$v(a) = v_0 \qquad \textit{and} \qquad v|_Z \equiv 0. \tag{3.3.18}$$

Moreover

$$\frac{v(x)}{v_0} = \mathbb{P}_x[\tau_a < \tau_Z], \tag{3.3.19}$$

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

Proof. Set $h(x) = v(x)$ on $A \cup Z$. Theorem 3.3.1 gives the result. ∎

Note in the definition above 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.

Let v be a voltage function on $\mathcal{N}$ with source a and sink Z. The Laplacian-based formulation of harmonicity, (3.3.7), can be interpreted in terms of flows (see Definition 1.1.13). We define the *current* function

current

$$i(x, y) := c(x, y)[v(x) - v(y)], \tag{3.3.20}$$

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 is defined on ordered pairs of vertices and is anti-symmetric, that is, $i(x, y) = -i(y, x)$. In terms of the current, the harmonicity of v is then expressed as

Ohm's law

$$\sum_{y:y\sim x} i(x, y) = 0, \qquad \forall x \in W, \tag{3.3.21}$$

that is, i is a flow on W (without capacity constraints). 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 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

Kirchhoff's node law

Kirchhoff's cycle law

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

as can be seen by substituting Ohm's law.

The *strength* of the current is defined as

strength

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

Because $a \notin W$, it does not satisfy Kirchhoff's node law and the strength is not 0 in general. The definition of $i(x, y)$ ensures that the flow out of the source is nonnegative as $\mathbb{P}_y[\tau_a < \tau_Z] \leq 1 = \mathbb{P}_a[\tau_a < \tau_Z]$ for all $y \sim a$ so that

$$i(a, y) = c(a, y)[v(a) - v(y)] = c(a, y)\,[v_0\mathbb{P}_a[\tau_a < \tau_Z] - v_0\mathbb{P}_y[\tau_a < \tau_Z]] \geq 0.$$

Note that by multiplying the voltage by a constant we obtain a current which is similarly scaled. Up to that scaling, the current 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 $\|i\| = 1$.

unit current

Summing up the previous paragraphs, 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 3.3.14 (Network reduction: birth-death chain)**.** Let $\mathcal{N}$ be the line on $\{0, 1, \dots, n\}$ 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 $1 \le x \le n - 1$. Consider the voltage function v when $v(0) = 1$ and $v(n) = 0$ with current i, which exists and is unique by Corollary 3.3.13. The desired quantity is $v(x)$.

Note that because i is a flow on $\mathcal{N}$, 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 (to the right of x) 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 $\mathcal{N}$ (with the original voltage function)

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

Ohm's law is also satisfied on every other edge (to the right of x) 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 3.4.

Because the voltage at x was not changed by this transformation, 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

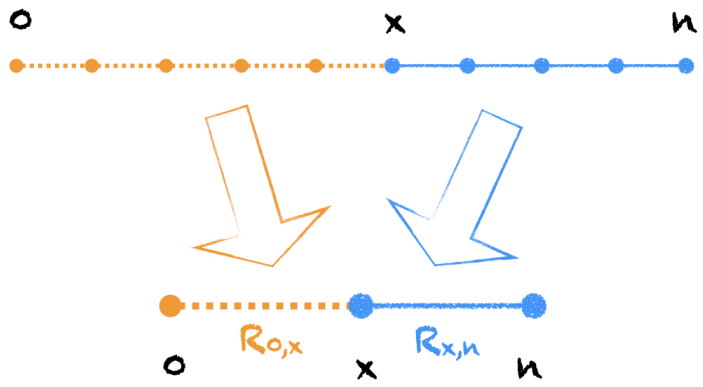

Figure 3.4: Reduced network.

the reciprocal of the resistance), that is,

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

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 (see Examples 3.1.41 and 3.1.43) corresponds to taking $c(j, j+1) = (p/q)^j$ or $r(j, j+1) = (q/p)^j$, for some $0 < p < 1$ and $q = 1 - p$. In this case we obtain
$$\mathbb{P}_x[\tau_0 < \tau_n] = \frac{\sum_{j=x}^{n-1}(q/p)^j}{\sum_{j=0}^{n-1}(q/p)^j} = \frac{(q/p)^x(1-(q/p)^{n-x})}{1-(q/p)^n} = \frac{(p/q)^{n-x}-1}{(p/q)^n - 1},$$
when $p \neq q$ (otherwise we get back the simple random walk case).z

◀

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, although to avoid complicating the notation further we will not do this here. (But see Example 3.3.22 for a simple case.)

series law, parallel law

Another useful network reduction technique is illustrated in the next example.

Example 3.3.15 (Network reduction: binary tree)**.** Let $\mathcal{N}$ be the rooted binary tree with n levels $\widehat{\mathbb{T}}_2^n$ 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

$$\mathbb{P}_x[\tau_0 < \tau_n] = (n-x)/n.$$

Indeed let v be the voltage with values 1 and 0 at $a = 0$ and $Z = \{n\}$ 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, that is, 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 3.3.14. 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 3.3.14. 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

$$\mathbb{P}[a \to Z] := \mathbb{P}_a[\tau_Z < \tau_a^+],$$

for the *escape probability*. The next lemma can be interpreted as a sort of Ohm's law between a and Z, where $c(a)\,\mathbb{P}[a \to Z]$ is the "effective conductance." (We will be more formal in Definition 3.3.19 below.)

escape probability

Lemma 3.3.16 (Effective Ohm's Law)**.** *Let v be a voltage on $\mathcal{N}$ with source a and sink Z. Let i be the associated current. Then*

$$\frac{v(a)}{\|i\|} = \frac{1}{c(a)\,\mathbb{P}[a \to Z]}. \tag{3.3.22}$$

Proof. Using the usual first-step analysis,

$$\begin{aligned}
\mathbb{P}[a \to Z] &= \sum_{x:x\sim a} P(a,x)\mathbb{P}_x[\tau_Z < \tau_a] \\
&= \sum_{x:x\sim a} \frac{c(a,x)}{c(a)}\left(1 - \frac{v(x)}{v(a)}\right) \\
&= \frac{1}{c(a)v(a)} \sum_{x:x\sim a} c(a,x)[v(a) - v(x)] \\
&= \frac{1}{c(a)v(a)} \sum_{x:x\sim a} i(a,x),
\end{aligned}$$

where we used Corollary 3.3.13 on the second line and Ohm's law on the last line. Rearranging gives the result. ∎

Recall the Green function from (3.1.3).

Theorem 3.3.17 (Probabilistic interpretation of the current)**.** *For $x \sim y$, let $N^Z_{x\to y}$ be the number of one-step transitions from x to y up to the time of the first visit to the sink Z for the random walk on $\mathcal{N}$ started at a. Let v be the voltage corresponding to the unit current i. Then the following formulas hold:*

$$v(x) = \frac{\mathscr{G}_{\tau_Z}(a,x)}{c(x)}, \qquad \forall x, \tag{3.3.23}$$

and

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

Proof. We prove the formula for the voltage by showing that $v(x)$ as defined above is harmonic on $W = V\backslash(\{a\} \cup Z)$. Note first that, for all $z \in Z$, the expected number of visits to z before reaching Z (i.e., $\mathscr{G}_{\tau_Z}(a,z)$) is 0. Or, put differently, $0 = v(z) = \frac{\mathscr{G}_{\tau_Z}(a,z)}{c(z)}$. Moreover, to compute $\mathscr{G}_{\tau_Z}(a,a)$, 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 (Theorem 3.1.8) and hence

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

and, by Lemma 3.3.16 and the fact that we are using the unit current, $v(a) = \frac{\mathscr{G}_{\tau_Z}(a,a)}{c(a)}$, as required.

To establish the formula for $x \in W$, we compute the quantity

$$\frac{1}{c(x)} \sum_{y : y \sim x} \mathbb{E}_a[N^Z_{y \to x}],$$

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

$$\frac{1}{c(x)} \sum_{y : y \sim x} \mathbb{E}_a[N^Z_{y \to x}] = \frac{\mathscr{G}_{\tau_Z}(a,x)}{c(x)}. \tag{3.3.24}$$

On the other hand, by the Markov property (Theorem 1.1.18)

$$\begin{aligned}
&\mathbb{E}_a[N^Z_{y \to x}] \\
&= \mathbb{E}_a \left[\sum_{0 \le t < \tau_Z} \mathbf{1}_{\{X_t = y, X_{t+1} = x\}} \right] \\
&= \sum_{t \ge 0} \mathbb{P}_a \left[X_t = y, X_{t+1} = x, \tau_Z > t \right] \\
&= \sum_{t \ge 0} \mathbb{P}_a[\tau_Z > t] \mathbb{P}_a[X_t = y \,|\, \tau_Z > t] \mathbb{P}_a[X_{t+1} = x \,|\, X_t = y, \tau_Z > t] \\
&= \sum_{t \ge 0} \mathbb{P}_a[\tau_Z > t] \mathbb{P}_a[X_t = y \,|\, \tau_Z > t] P(y,x) \\
&= \sum_{t \ge 0} \mathbb{P}_a[X_t = y, \tau_Z > t] P(y,x) \\
&= P(y,x)\, \mathbb{E}_a \left[\sum_{0 \le t < \tau_Z} \mathbf{1}_{\{X_t = y\}} \right] \\
&= P(y,x)\, \mathscr{G}_{\tau_Z}(a,y),
\end{aligned} \tag{3.3.25}$$

so that, summing over y, we obtain this time

$$\begin{aligned}\frac{1}{c(x)}\sum_{y:y\sim x}\mathbb{E}_a[N^Z_{y\to x}] &= \frac{1}{c(x)}\sum_{y:y\sim x}P(y,x)\,\mathscr{G}_{\tau_Z}(a,y)\\ &= \sum_{y:y\sim x}P(x,y)\frac{\mathscr{G}_{\tau_Z}(a,y)}{c(y)},\end{aligned} \tag{3.3.26}$$

where we used that $c(x,y) = c(x)P(x,y) = c(y)P(y,x)$ (see Definition 1.2.7). Equating (3.3.24) and (3.3.26) shows that $\frac{\mathscr{G}_{\tau_Z}(a,x)}{c(x)}$ is harmonic on W and hence must be equal to the voltage function by Corollary 3.3.13.

Finally, by (3.3.25),

$$\begin{aligned}\mathbb{E}_a[N^Z_{x\to y} - N^Z_{y\to x}] &= P(x,y)\,\mathscr{G}_{\tau_Z}(a,x) - P(y,x)\,\mathscr{G}_{\tau_Z}(a,y)\\ &= P(x,y)v(x)c(x) - P(y,x)v(y)c(y)\\ &= c(x,y)[v(x)-v(y)]\\ &= i(x,y).\end{aligned}$$

That concludes the proof. ■

Example 3.3.18 (Network reduction: binary tree (continued))**.** Recall the setting of Example 3.3.15. We argued that the current on side edges, that is, 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

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 $\mathbb{P}_a[\tau_a^+ < +\infty] < 1$.

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,

exhaustive sequence

by Lemma 3.1.25, $\mathbb{P}_a[\tau_{Z_n} \wedge \tau_a^+ = +\infty] = 0$ for all n by our assumptions on (G_n). Hence, the remaining possibilities are

$$\begin{aligned} 1 &= \mathbb{P}_a[\exists n,\ \tau_a^+ < \tau_{Z_n}] + \mathbb{P}_a[\forall n,\ \tau_{Z_n} < \tau_a^+] \\ &= \mathbb{P}_a[\tau_a^+ < +\infty] + \lim_n \mathbb{P}[a \to Z_n]. \end{aligned}$$

Therefore a is transient if and only if $\lim_n \mathbb{P}[a \to Z_n] > 0$. 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. Hence we define

$$\mathbb{P}[a \to \infty] := \lim_n \mathbb{P}[a \to Z_n].$$

We use Lemma 3.3.16 to characterize this limit using electrical network concepts.

But, first, here comes the key definition. In Lemma 3.3.16, $v(a)$ 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 "super-edge," Equation (3.3.22) is the analogue of Ohm's law if we interpret $c(a)\,\mathbb{P}[a \to Z]$ as an "effective conductance."

Definition 3.3.19 (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*

effective resistance

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

where the rightmost equality holds by Lemma 3.3.16. The reciprocal is called the effective conductance *and is denoted by $\mathscr{C}(a \leftrightarrow Z) := 1/\mathscr{R}(a \leftrightarrow Z)$.*

effective conductance

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

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

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

Theorem 3.3.20 (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 $\mathscr{R}(a \leftrightarrow \infty) < +\infty$.*

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 3.3.21 (Gambler's ruin chain revisited)**.** Extend the gambler's ruin chain of Example 3.3.14 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 $\{0, 1, \ldots, n-1\}$ and letting $Z_n = \{n, n+1 \ldots\}$. To compute the effective resistance $\mathscr{R}(0 \leftrightarrow Z_n)$, we use the same reduction as in Example 3.3.14. The "super-edge" between 0 and n has resistance

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

when $p \neq q$, and similarly it has resistance n in the $p = q$ case. Hence, taking a limit as $n \to +\infty$,

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

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

Example 3.3.22 (Biased walk on the b-ary tree)**.** Fix $\lambda \in (0, +\infty)$. Consider the rooted, infinite b-ary tree with conductance λ^j on all edges between level $j-1$ and j, for $j \geq 1$. 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 $\mathscr{R}(0 \leftrightarrow Z_n)$: (i) glue together all vertices of Z_n; (ii) glue together all vertices on the same level of G_n; (iii) replace parallel edges with a single edge whose conductance is the sum of the conductances; (iv) let the current on this edge be the sum of the currents; and (v) leave the voltages unchanged. It can be checked that Ohm's law and Kirchhoff's node law are still satisfied, and that 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

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

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

A generalization is provided in Example 3.3.27. ◀

3.3.3 Bounding the effective resistance via variational principles

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. We derive them for finite networks, but will later on apply them to exhaustive sequences.

Variational principles

Recall from Definition 1.1.13 that a flow θ 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: θ is anti-symmetric, that is, $\theta(x, y) = -\theta(y, x)$ for all $x \sim y$; and it satisfies the flow-conservation constraint $\sum_{y:y\sim x} \theta(x, y) = 0$ on all vertices x except those in $\{a\} \cup Z$. The strength of the flow is $\|\theta\| = \sum_{y:y\sim a} \theta(a, y)$. 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 θ is defined as

energy

$$\mathscr{E}(\theta) = \frac{1}{2} \sum_{x,y} r(x, y)\theta(x, y)^2.$$

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 3.3.23 (Thomson's principle)**.** *Let $\mathcal{N} = (G, c)$ be a finite, connected network. The effective resistance between source a and sink Z is characterized by*

$$\mathscr{R}(a \leftrightarrow Z) = \inf \left\{ \mathscr{E}(\theta) \,:\, \theta \text{ is a unit flow from } a \text{ to } Z \right\}. \tag{3.3.27}$$

The unique minimizer is the unit current.

Proof. It will be convenient to work in vector form. Let $1, \dots, n$ be the vertices of G and order the edges arbitrarily as $e_1, \dots, e_m$. (We ignore any self-loops, which have no flow.) Choose an arbitrary orientation of $\mathcal{N}$, that is, replace each edge $e_i = \{x, y\}$ with either $\vec{e}_i = (x, y)$ or (y, x). Let $\overrightarrow{G}$ be the corresponding directed graph. Think of the flow $\boldsymbol{\theta}$ as a vector with one coordinate for each oriented edge. Then the flow constraint can be written as a linear system $B\boldsymbol{\theta} = \boldsymbol{b}$. Here the matrix B has a column for each directed edge and a row for each vertex *except those in* Z. The entries of B are $B_{x,(x,y)} = 1$, $B_{y,(x,y)} = -1$, and 0 otherwise. We have already encountered this matrix: it is an oriented incidence matrix of G (see Definition 1.1.16) restricted to the rows in $V \setminus Z$. The vector $\boldsymbol{b}$ has 0s everywhere except for $b_a = 1$. Let $\boldsymbol{r}$ be the vector of resistances and let R be the diagonal matrix with diagonal $\boldsymbol{r}$. In vector form, $\mathscr{E}(\boldsymbol{\theta}) = \boldsymbol{\theta}^T R\boldsymbol{\theta}$ and the optimization problem (3.3.27) reads

$$\mathscr{E}^* = \inf\{\boldsymbol{\theta}^T R\boldsymbol{\theta} \,:\, B\boldsymbol{\theta} = \boldsymbol{b}\}.$$

We first characterize the optimal flow. We introduce the *Lagrangian*

Lagrangian

$$\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{h}) := \boldsymbol{\theta}^T R\boldsymbol{\theta} - 2\boldsymbol{h}^T(B\boldsymbol{\theta} - \boldsymbol{b}),$$

where $\boldsymbol{h}$ has an entry for all vertices *except those in* Z. For all $\boldsymbol{h}$,

$$\mathscr{E}^* \geq \inf_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{h}),$$

because those $\boldsymbol{\theta}$s with $B\boldsymbol{\theta} = \boldsymbol{b}$ make the second term vanish in $\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{h})$. Since $\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{h})$ is strictly convex as a function of $\boldsymbol{\theta}$, the solution to its minimization is characterized by the usual optimality conditions which in this case read $2R\boldsymbol{\theta} - 2B^T\boldsymbol{h} = 0$, or

$$\boldsymbol{\theta} = R^{-1}B^T\boldsymbol{h}. \tag{3.3.28}$$

Substituting into the Lagrangian and simplifying, we have proved that

$$\mathscr{E}(\boldsymbol{\theta}) \geq \mathscr{E}^* \geq -\boldsymbol{h}^T BR^{-1}B^T\boldsymbol{h} + 2\boldsymbol{h}^T\boldsymbol{b} =: \mathscr{L}^*(\boldsymbol{h}), \tag{3.3.29}$$

for all $\boldsymbol{h}$ and flow $\boldsymbol{\theta}$. This inequality is a statement of weak duality. To show that a flow $\boldsymbol{\theta}$ is optimal it suffices to find $\boldsymbol{h}$ such that $\mathscr{E}(\boldsymbol{\theta}) = \mathscr{L}^*(\boldsymbol{h})$.

Let $\boldsymbol{\theta} = \boldsymbol{i}$ be the unit current in vector form, which satisfies $B\boldsymbol{\theta} = \boldsymbol{b}$ by our choice of $\boldsymbol{b}$ and Kirchhoff's node law (i.e., (3.3.21)). The suitable dual turns out to be the corresponding voltage $\boldsymbol{h} = \boldsymbol{v}$ in vector form restricted to $V \setminus Z$. To see this, observe that $B^T\boldsymbol{h}$ is the vector of neighboring node differences

$$B^T\boldsymbol{h} = (h(x) - h(y))_{(x,y)\in\overrightarrow{G}}, \tag{3.3.30}$$

where implicitly $h|_Z \equiv 0$. Hence the optimality condition (3.3.28) is nothing but Ohm's law (i.e., (3.3.20)) in vector form. Therefore, if $\boldsymbol{i}$ is the unit current and $\boldsymbol{v}$ is the associated voltage in vector form, it holds that

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

where the first equality follows from the fact that $\boldsymbol{i}$ minimizes $\mathscr{L}(\boldsymbol{i}; \boldsymbol{v})$ by (3.3.28) and the second equality follows from the fact that $B\boldsymbol{i} = \boldsymbol{b}$. So we must have $\mathscr{E}(\boldsymbol{i}) = \mathscr{E}^*$ by weak duality (i.e., (3.3.29)).

As for uniqueness, it can be checked that two minimizers $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ satisfy

$$\mathscr{E}^* = \frac{\mathscr{E}(\boldsymbol{\theta}) + \mathscr{E}(\boldsymbol{\theta}')}{2} = \mathscr{E}\left(\frac{\boldsymbol{\theta} + \boldsymbol{\theta}'}{2}\right) + \mathscr{E}\left(\frac{\boldsymbol{\theta} - \boldsymbol{\theta}'}{2}\right),$$

by definition of the energy. The first term in the rightmost expression is greater or equal to $\mathscr{E}^*$ since 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 the only way for this to happen is if $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

To conclude the proof, it remains to compute the optimal value. The matrix $BR^{-1}B^T$ is related to the Laplacian associated to random walk on $\mathcal{N}$ (see Section 3.3.1) up to a row scaling. Multiplying by row $x \in V \setminus Z$ involves taking a conductance-weighted average of the neighboring values and subtracting the value at x, that is,

$$\begin{aligned}\left(BR^{-1}B^T\boldsymbol{v}\right)_x &= \sum_{y:(x,y)\in\overrightarrow{G}} \Big[c(x,y)(v(x) - v(y))\Big] \\ &\qquad - \sum_{y:(y,x)\in\overrightarrow{G}} \Big[c(y,x)(v(y) - v(x))\Big] \\ &= \sum_{y:y\sim x} \Big[c(x,y)(v(x) - v(y))\Big],\end{aligned}$$

where we used (3.3.30) and the facts that $r(x,y)^{-1} = c(x,y)$ and $c(x,y) = c(y,x)$, and it is assumed implicitly that $v|_Z \equiv 0$. By Corollary 3.3.13, this is zero except for the row $x = a$ where it is

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

where we used Ohm's law and the fact that the current has unit strength. We have

finally

$$\begin{aligned}\mathscr{E}^* &= \mathscr{L}^*(\boldsymbol{v}) \\ &= -\boldsymbol{v}^T B R^{-1} B^T \boldsymbol{v} + 2\boldsymbol{v}^T \boldsymbol{b} \\ &= -v(a) + 2v(a) \\ &= v(a) \\ &= \mathscr{R}(a \leftrightarrow Z),\end{aligned}$$

by (3.3.16). That concludes the proof. ∎

Observe that the convex combination $\boldsymbol{\alpha}$ minimizing the sum of squares $\sum_j \alpha_j^2$ is constant. In a similar manner, Thomson's principle (Theorem 3.3.23) 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 3.3.24 (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

$$\mathscr{R}(1 \leftrightarrow n) = \frac{2}{n},$$

from the definition of the effective resistance (Definition 3.3.19).

We now look for the optimal flow in Thomson's principle. Pushing a flow of 1 through 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 $3 \cdot (1/2)^2 = 3/4$. Taking this further, pushing a flow of $\frac{1}{n-1}$ through $\{1, n\}$ as well as through each two-edge path to n via the remaining neighbors of 1 gives the yet improved bound

$$\left(\frac{1}{n-1}\right)^2 + 2(n-2)\left(\frac{1}{n-1}\right)^2 = \frac{2n-3}{(n-1)^2} = \frac{2}{n} \cdot \frac{2n^2-3n}{2n^2-4n+2} > \frac{2}{n},$$

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 α on $\{1, n\}$ (and the rest divided up evenly among the two-edge paths), we get an energy of $\alpha^2 + 2(n-2)[\frac{1-\alpha}{n-2}]^2$ which is minimized at $\alpha = \frac{2}{n}$ where it is indeed

$$\left(\frac{2}{n}\right)^2 + \frac{2}{n-2}\left(\frac{n-2}{n}\right)^2 = \frac{2}{n}\left(\frac{2}{n} + \frac{n-2}{n}\right) = \frac{2}{n}.$$

◀

As we noted above, the matrix $BR^{-1}B^T$ in the proof of Thomson's principle is related to the Laplacian. Because $B^T\boldsymbol{h}$ is the vector of neighboring node differences, we have

$$\boldsymbol{h}^T BR^{-1}B^T\boldsymbol{h} = \frac{1}{2}\sum_{x,y} c(x,y)[h(y) - h(x)]^2,$$

where we implicitly fix $h|_Z \equiv 0$, which is called the *Dirichlet energy*. Thinking of B^T as a "discrete gradient," the Dirichlet energy can be interpreted as the weighted norm of the gradient of $\boldsymbol{h}$. The following is a "dual" to Thomson's principle. Exercise 3.15 asks for a proof.

Dirichlet energy

Theorem 3.3.25 (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*

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

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. Recall from Section 1.1.1 that, on a finite graph, a cutset separating a from Z is a set of edges Π such that any path between a and Z must include at least one edge in Π. Similarly, as defined in Section 2.3.3, on a countable, locally finite network, a cutset separating a from ∞ is a finite set of edges that must be crossed by any infinite (self-avoiding) path from a.

Corollary 3.3.26 (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*

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

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 ∞,

$$\mathscr{R}(a \leftrightarrow \infty) \geq \sum_j \left(\sum_{e \in \Pi_j} c(e) \right)^{-1}.$$

Proof. Consider the case where $\mathcal{N}$ is finite first. We will need the following claim, which follows immediately from Lemma 1.1.14: for any unit flow θ between a and Z and any cutset Π_j separating a from Z, it holds that

$$\sum_{e \in \Pi_j} |\theta(e)| \geq \|\theta\| = 1.$$

By Cauchy-Schwarz (Theorem B.4.8),

$$\begin{aligned} \sum_{e \in \Pi_j} c(e) \sum_{e' \in \Pi_j} r(e')\theta(e')^2 &\geq \left(\sum_{e \in \Pi_j} \sqrt{c(e)r(e)}\, |\theta(e)| \right)^2 \\ &= \left(\sum_{e \in \Pi_j} |\theta(e)| \right)^2 \\ &\geq 1. \end{aligned}$$

Rearranging, summing over j and using the disjointness of the cutsets,

$$\mathscr{E}(\theta) = \frac{1}{2} \sum_{x,y} r(x,y)\theta(x,y)^2 \geq \sum_{j=1}^{n} \sum_{e' \in \Pi_j} r(e')\theta(e')^2 \geq \sum_{j=1}^{n} \left(\sum_{e \in \Pi_j} c(e) \right)^{-1}.$$

Thomson's principle gives the result.

The infinite case follows from a similar argument using an exhaustive sequence. ∎

The following example is an application of Nash-Williams (Corollary 3.3.26) and Thomson's principle to recurrence.

Example 3.3.27 (Biased walk on general trees)**.** Let $\mathscr{T}$ be a locally finite tree with root 0. Consider again the biased walk from Example 3.3.22, that is, the conductance is λ^j on all edges between level $j-1$ and j. Recall the branching number $\mathrm{br}(\mathscr{T})$ from Definition 2.3.10.

Assume $\lambda > \mathrm{br}(\mathscr{T})$. For any $\varepsilon > 0$, there is a cutset Π such that $\sum_{e\in\Pi} \lambda^{-|e|} \leq \varepsilon$. By Nash-Williams,

$$\mathscr{R}(0 \leftrightarrow \infty) \geq \left(\sum_{e\in\Pi} c(e)\right)^{-1} \geq \varepsilon^{-1}.$$

Since ε is arbitrary, the walk is recurrent by Theorem 3.3.20.

Suppose instead that $\lambda < \mathrm{br}(\mathscr{T})$ and let $\lambda < \lambda_* < \mathrm{br}(\mathscr{T})$. By the proof of Claim 2.3.11, for all $n \geq 1$, there exist $\varepsilon > 0$ and a unit flow ϕ_n from 0 to the n-level vertices ∂_n with capacity constraints $|\phi_n(x,y)| \leq \varepsilon^{-1}\lambda_*^{-|e|}$ for all edges $e = \{x,y\}$, where $|e|$ is the graph distance from the root to the endvertex of e furthest from it. Then, letting $F_m = \{e : |e| = m\}$, the energy of the flow is

$$\begin{aligned}
\mathscr{E}(\phi_n) &= \frac{1}{2}\sum_{x,y} r(x,y)\phi_n(x,y)^2 \\
&\leq \sum_{m=1}^{n} \lambda^m \sum_{e=\{x,y\}\in F_m} |\phi_n(x,y)|\varepsilon^{-1}\lambda_*^{-|e|} \\
&= \varepsilon^{-1}\sum_{m=1}^{n}\left(\frac{\lambda}{\lambda_*}\right)^m \sum_{e=\{x,y\}\in F_m} |\phi_n(x,y)| \\
&\leq \varepsilon^{-1}\sum_{m=1}^{+\infty}\left(\frac{\lambda}{\lambda_*}\right)^m \\
&< +\infty,
\end{aligned}$$

where, on the fourth line, we used Lemma 1.1.14 together with the fact that ϕ_n is a unit flow and F_m is a cutset separating 0 and ∂_n. Thomson's principle implies that $\mathscr{R}(0 \leftrightarrow \partial_n)$ is uniformly bounded in n. The walk is transient by Theorem 3.3.20. ◀

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

Corollary 3.3.28. *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 with a, then $\mathbb{P}[a \to Z]$ 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 3.3.29 (Rayleigh's principle)**.** *Let $\mathcal{N}$ and $\mathcal{N}'$ be two networks on the same finite, connected graph G such that, for each edge in G, the resistance in $\mathcal{N}'$ is greater than it is in $\mathcal{N}$. Then, for any source a and sink Z,*

$$\mathscr{R}_{\mathcal{N}}(a \leftrightarrow Z) \leq \mathscr{R}_{\mathcal{N}'}(a \leftrightarrow Z).$$

Proof. Compare the energies of an arbitrary flow on $\mathcal{N}$ and $\mathcal{N}'$, and apply Thomson's principle. ■

Note that this corollary implies the previous one by thinking of an absent edge as one with infinite resistance.

Flows to infinity

Combining Theorem 3.3.20 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 θ from source a to ∞* on a countable, locally finite network: θ 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 $\mathscr{E}(\theta)$ of such a flow is well defined in $[0, +\infty]$.

flow to ∞

Theorem 3.3.30 (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 ∞ of finite energy.*

Proof. 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 ∞ on $\mathcal{N}$ 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 $\mathscr{R}(a \leftrightarrow Z_n) \leq B$ for all n and transience follows from Theorem 3.3.20.

Proving the other direction involves producing a flow to ∞. Suppose a is transient and let (G_n) be an exhaustive sequence as above. Then Theorem 3.3.20 implies that $\mathscr{R}(a \leftrightarrow Z_n) \leq \mathscr{R}(a \leftrightarrow \infty) < B$ for some $B < +\infty$ and Thomson's principle guarantees in turn the existence of a flow θ_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 ∞ 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 (Proposition B.4.14),

$\mathbb{E}_a Y_n(x) \to \mathbb{E}_a Y_\infty(x)$ where $Y_\infty(x)$ is the total number of visits to x. Moreover $\mathbb{E}_a Y_\infty(x) < +\infty$ by transience and (3.1.2). By (3.3.23), $\mathbb{E}_a Y_n(x) = c(x)v_n(x)$. So we can now define

$$v_\infty(x) := \lim_n v_n(x) < +\infty,$$

and then

$$\begin{aligned} i_\infty(x,y) &:= c(x,y)[v_\infty(x) - v_\infty(y)] \\ &= \lim_n c(x,y)[v_n(x) - v_n(y)] \\ &= \lim_n i_n(x,y), \end{aligned}$$

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_∞. Note that by construction of i_ℓ

$$\begin{aligned} \frac{1}{2}\sum_{x,y\in G_n} c(x,y)i_\infty(x,y)^2 &= \lim_{\ell\geq n}\frac{1}{2}\sum_{x,y\in G_n} c(x,y)i_\ell(x,y)^2 \\ &\leq \limsup_{\ell\geq n} \mathscr{E}(i_\ell) \\ &< B, \end{aligned}$$

uniformly in n. Because the left-hand side converges to the energy of i_∞ as $n \to +\infty$, we are done. ∎

We give an application to Pólya's theorem in Section 3.3.4.

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 3.3.31 (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 ϕ from the vertices of G to the vertices of G' is a* rough embedding *if there are constants $\alpha, \beta < +\infty$ and a map Φ defined on the edges of G such that:*

rough embedding

1. *for every edge $e = \{x, y\}$ in G, $\Phi(e)$ is a non-empty 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 β edges in G whose image under Φ contains e'.*

The map ϕ need not in general be a bijection.

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

roughly equivalent

Example 3.3.32 (Independent-coordinate random walk)**.** Let $\mathcal{N} = \mathbb{L}^d$ with unit resistances and let $\mathcal{N}'$ be the network corresponding to the *independent-coordinate random walk*

$$(Y_t^{(1)}, \ldots, Y_t^{(d)}),$$

where each coordinate $(Y_t^{(i)})$ is an independent simple random walk on $\mathbb{Z}$ started at 0. For example the neighborhood of the origin in $\mathcal{N}'$ is $\{(x_1, \ldots, x_d) : x_i \in \{-1, 1\},\ \forall i\}$. Note that $\mathcal{N}'$ contains only those points of $\mathbb{Z}^d$ with coordinates of identical parities.

Despite encoding quite different random walks, we claim that the networks $\mathcal{N}$ and $\mathcal{N}'$ are roughly equivalent.

- *$\mathcal{N}$ to $\mathcal{N}'$:* Consider the map ϕ which associates to each $x \in \mathcal{N}$ a closest point in $\mathcal{N}'$ chosen in some arbitrary manner. For Φ, 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 ϕ which associates to each $x \in \mathcal{N}'$ the corresponding point x in $\mathcal{N}$. Construct Φ similarly to the previous case.

Exercise 3.19 asks for a rigorous proof of rough equivalence. See also Exercise 3.20 for an important generalization of this example. ◀

Our main result about roughly equivalent networks is that they have the same type.

Theorem 3.3.33 (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 θ be a unit flow from some a to ∞ of finite energy. The existence of this flow is guaranteed by Theorem 3.3.30. Let ϕ, Φ be a rough embedding from $\mathcal{N}$ to $\mathcal{N}'$ with parameters α and β.

The basic idea of the proof is to map the flow θ onto $\mathcal{N}'$ using Φ. Because flows are directional, it will be convenient to think of edges as being directed. For

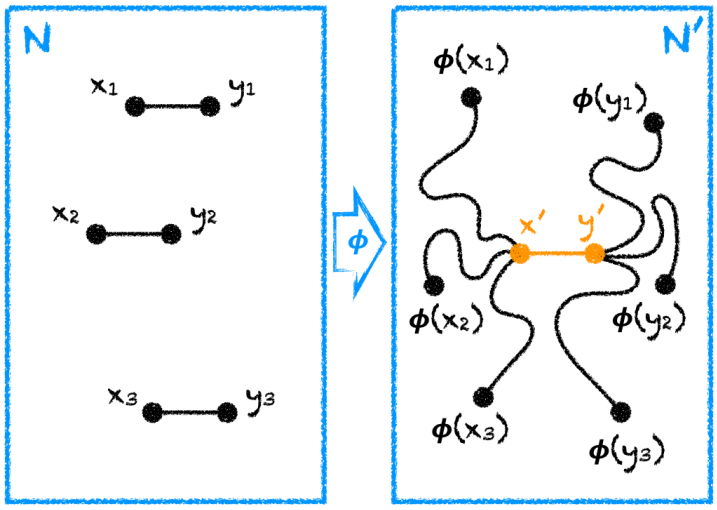

Figure 3.5: The flow on (x', y') is the sum of the flows on (x_1, y_1), (x_2, y_2), and (x_3, y_3).

$e = \{x, y\}$ in $\mathcal{N}$, let $\overrightarrow{\Phi}(x, y)$ be the path $\Phi(e)$ oriented from $\phi(x)$ to $\phi(y)$. So $(x', y') \in \overrightarrow{\Phi}(x, y)$ means 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 $\overrightarrow{\Phi}(y, x)$.) Then define, for x', y' with $\{x', y'\}$ in $\mathcal{N}'$,

$$\theta'(x', y') := \sum_{(x,y):(x',y') \in \overrightarrow{\Phi}(x,y)} \theta(x, y). \tag{3.3.31}$$

See Figure 3.5.

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

1. *(Anti-symmetry)* By construction, $\theta'(y', x') = -\theta'(x', y')$, that is, θ' is anti-symmetric, because θ itself is anti-symmetric. We used the fact that $\overrightarrow{\Phi}(y, x)$ is $\overrightarrow{\Phi}(x, y)$ oriented in the opposite direction.

2. *(Flow conservation)* Next we check the flow-conservation constraints. Fix z' in $\mathcal{N}'$. By Condition 2 in Definition 3.3.31, there are finitely many edges e in $\mathcal{N}$ such that $\Phi(e)$ 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 (u', z'), (z', w') be the directed edges incident with z' on $\overrightarrow{\Phi}(x, y)$. Observe that, in the definition of θ', (y, x) contributes $\theta(y, x) = -\theta(x, y)$ to $\theta'(z', u')$ and (x, y) contributes $\theta(x, y)$ to $\theta'(z', w')$. So these contributions cancel out in the flow-conservation constraint for z', that is, in the sum $\sum_{v':v'\sim z'} \theta'(z', v')$.
- If instead $e = \{x, y\}$ is such that $\phi(x) = z'$, let (z', w') be the first edge on the path $\overrightarrow{\Phi}(x, y)$. Edge (x, y) contributes $\theta(x, y)$ to $\theta'(z', w')$. A similar statement applies to $\phi(y) = z'$ by changing the role of x and y. This case also applies to $\phi(x) = \phi(y) = z'$.

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

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

Because θ is a flow, the sum in parentheses is 0 if $z \neq a$ and 1 otherwise. So the right-hand side is 0 unless $a \in \phi^{-1}(\{z'\})$ in which case it is 1.

We have shown that θ' is a unit flow from $\phi(a)$ to ∞. It remains to bound the energy of θ'. By (3.3.31), Cauchy-Schwarz, and Condition 2 in Definition 3.3.31,

$$\begin{aligned}
\theta'(x', y')^2 &= \left[\sum_{(x,y):(x',y')\in\overrightarrow{\Phi}(x,y)} \theta(x, y) \right]^2 \\
&\leq \left[\sum_{(x,y):(x',y')\in\overrightarrow{\Phi}(x,y)} 1 \right] \left[\sum_{(x,y):(x',y')\in\overrightarrow{\Phi}(x,y)} \theta(x, y)^2 \right] \\
&\leq \beta \sum_{(x,y):(x',y')\in\overrightarrow{\Phi}(x,y)} \theta(x, y)^2.
\end{aligned}$$

Summing over all pairs and using Condition 1 in Definition 3.3.31 gives

$$\begin{aligned}
\frac{1}{2} \sum_{x',y'} r'(x', y')\theta'(x', y')^2 &\leq \beta \frac{1}{2} \sum_{x',y'} r'(x', y') \sum_{(x,y):(x',y')\in\overrightarrow{\Phi}(x,y)} \theta(x, y)^2 \\
&= \beta \frac{1}{2} \sum_{x,y} \theta(x, y)^2 \sum_{(x',y')\in\overrightarrow{\Phi}(x,y)} r'(x', y') \\
&\leq \alpha\beta \frac{1}{2} \sum_{x,y} r(x, y)\theta(x, y)^2,
\end{aligned}$$

which is finite by assumption. That concludes the proof. ∎

As an application, we give a second proof of Pólya's theorem in Section 3.3.4.

Other applications

So far we have emphasized applications to recurrence. Here we show that electrical network theory can also be used to bound commute times. In Section 3.3.5, we give further applications beyond random walks on graphs.

An application of Corollary 3.1.24 gives another probabilistic interpretation of the effective resistance—and a useful formula.

Theorem 3.3.34 (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*

commute time

$$\mathbb{E}_x[\tau_{x,y}] = \mathbb{E}_x[\tau_y] + \mathbb{E}_y[\tau_x] = c_{\mathcal{N}}\,\mathscr{R}(x \leftrightarrow y),$$

where $c_{\mathcal{N}} = 2\sum_{e=\{x,y\}\in\mathcal{N}} c(e)$.

Proof. This follows immediately from Corollary 3.1.24 and the definition of the effective resistance (Definition 3.3.19). Specifically,

$$\begin{aligned}\mathbb{E}_x[\tau_y] + \mathbb{E}_y[\tau_x] &= \frac{1}{\pi_x\,\mathbb{P}_x[\tau_y < \tau_x^+]}\\ &= \frac{1}{(2\sum_{e=\{x,y\}\in\mathcal{N}} c(e))^{-1}c(x)\,\mathbb{P}_x[\tau_y < \tau_x^+]}\\ &= c_{\mathcal{N}}\,\mathscr{R}(x \leftrightarrow y).\end{aligned}$$

∎

Example 3.3.35 (Random walk on the torus)**.** Consider random walk on the d-dimensional torus $\mathbb{L}_n^d$ 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 $\mathbb{L}_n^d$. To use the commute time identity (Theorem 3.3.34), note that by symmetry $\mathbb{E}_x[\tau_y] = \mathbb{E}_y[\tau_x]$ so that

$$\mathbb{E}_x[\tau_y] = \frac{1}{2}\,c_{\mathcal{N}}\,\mathscr{R}(x \leftrightarrow y) = dn^d\,\mathscr{R}(x \leftrightarrow y). \tag{3.3.32}$$

where we used that the number of vertices is n^d and the graph is $2d$-regular.

To simplify, assume n is odd and identify the vertices of $\mathbb{L}_n^d$ with the box

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

in $\mathbb{L}^d$ centered at $x = 0$. Let $\partial B_j^\infty = \{z \in \mathbb{L}^d : \|z\|_\infty = j\}$ and let Π_j be the set of edges between ∂B_j^∞ and ∂B_{j+1}^∞. Note that on B the ℓ^1 norm of y is at most k (the graph distance between $x = 0$ and y). Since the ℓ^∞ norm is at least $1/d$ times the ℓ^1 norm on $\mathbb{L}^d$, there exists $J = O(k)$ such that all Π_js, $j \le J$, are cutsets separating x from y. By the Nash-Williams inequality

$$\mathscr{R}(x \leftrightarrow y) \ge \sum_{0 \le j \le J} |\Pi_j|^{-1} = \sum_{0 \le j \le J} \Omega\left(j^{-(d-1)}\right) = \begin{cases} \Omega(\log k), & d = 2 \\ \Omega(1), & d \ge 3. \end{cases}$$

From (3.3.32), we get:

Claim 3.3.36.

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

◀

Remark 3.3.37. *The bounds in the previous example are tight up to constants. See [LPW06, Proposition 10.13]. Note that the case $d \ge 3$ does not in fact depend on the distance k.*

See Exercise 3.22 for an application of the commute time identity to cover times.

3.3.4 ▷ *Random walks: Pólya's theorem, two ways*

The following is a classical result.

Theorem 3.3.38 (Pólya's theorem)**.** *Random walk on $\mathbb{L}^d$ is recurrent for $d \le 2$ and transient for $d \ge 3$.*

We prove the theorem for $d = 2, 3$ using the tools developed in the previous subsection. The other cases follow by Rayleigh's principle (Corollary 3.3.29). There are elementary proofs of this result. But we showed above that the electrical network approach has the advantage of being robust to the details of the lattice. For a different argument, see Exercise 2.10.

The case $d = 2$ follows from the Nash-Williams inequality (Corollary 3.3.26) by letting Π_j be the set of edges connecting vertices of ℓ^∞ norm j and $j+1$. Using the fact that all conductances are 1, that $|\Pi_j| = O(j)$, and that $\sum_j j^{-1}$ diverges, recurrence is established by Theorem 3.3.20.

First proof

Now consider the case $d = 3$ and let $a = 0$ be the origin.

method of random paths

We construct a finite-energy flow to ∞ using the *method of random paths*. Note that a simple way to produce a unit flow to ∞ is to push a flow of 1 through an infinite path (which, recall, are self-avoiding by definition). Taking this a step further, let μ be a probability measure on infinite paths and define the anti-symmetric function

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

where Γ is a random path distributed according to μ, oriented away from 0. (We will give an explicit construction below where the appropriate formal probability space will be clear.) Observe that $\sum_{y\sim x}[\mathbf{1}_{(x,y)\in\Gamma} - \mathbf{1}_{(y,x)\in\Gamma}] = 0$ for any $x \neq 0$ because vertices visited by Γ are entered and exited exactly once. That same sum is 1 at $x = 0$. Hence θ is a unit flow to ∞. For edge $e = \{x, y\}$, consider the following "edge marginal" of μ:

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

where we used that a path Γ cannot visit both (x, y) and (y, x) by definition. Then we get the following bound.

Claim 3.3.39 (Method of random paths)**.**

$$\mathscr{E}(\theta) \leq \sum_e \mu(e)^2. \tag{3.3.33}$$

For a measure μ 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 μ as follows. Let $\boldsymbol{U}$ be a uniformly random point on the unit sphere in $\mathbb{R}^3$ and let γ be the ray from 0 to ∞ going through $\boldsymbol{U}$. Imagine centering a unit cube around each point in $\mathbb{Z}^3$ whose edges are aligned with the axes. Then γ traverses an infinite number of such cubes. Let Γ be the corresponding path in the lattice $\mathbb{L}^3$. To see that this procedure indeed produces a path observe that γ, 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 μ distributes its mass among sufficiently spread out paths, we bound the probability that a vertex is visited by Γ. 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

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

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

Lemma 3.3.40. *There are $O(j^2)$ edges with an endpoint at ℓ^2 distance within $[j, j+1]$ from the origin.*

Proof. Consider an open ball of ℓ^2 radius $1/2$ centered around each vertex of ℓ^2 norm within $[j, j+1]$. Those balls are non-intersecting and have total volume $\Theta(N_j)$, where N_j is the number of such vertices. On the other hand, the volume of the shell of ℓ^2 inner and outer radii $j - 1/2$ and $j + 3/2$ centered around the origin (where all those balls lie) 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 those bounds into (3.3.33), we get

$$\mathscr{E}(\theta) \leq \sum_j O(j^2) \cdot \left[O(1/j^2)\right]^2 = O\left(\textstyle\sum_j j^{-2}\right) < +\infty.$$

Transience follows from Theorem 3.3.30. (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 3.17.)

Second proof

We briefly describe a second proof based on the independent-coordinate random walk. Consider the networks $\mathcal{N}$ and $\mathcal{N}'$ in Example 3.3.32. Because they are roughly equivalent (Definition 3.3.31), they have the same type by Theorem 3.3.33. 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 transient if and only if the expected number of visits to 0 is finite (see (3.1.2)). By independence of the coordinates, this expectation can be written as

$$\sum_{t\geq 0}\left(\mathbb{P}\left[Y_{2t}^{(1)} = 0\right]\right)^d = \sum_{t\geq 0}\left(\binom{2t}{t}2^{-2t}\right)^d = \sum_{t\geq 0}\Theta(t^{-d/2}),$$

where we used Stirling's formula (see Appendix A). The rightmost sum is finite if and only if $d \geq 3$. That implies random walk on $\mathcal{N}'$ is transient under that condition. By rough equivalence, the same is true of $\mathcal{N}$.

3.3.5 ▷ *Randomized algorithms: Wilson's method for generating uniform spanning trees*

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

With a slight abuse of notation, we use $e \in G$ to indicate that e is an edge of G.

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. Such a tree has $|V| - 1$ edges. A *uniform spanning tree* is a spanning tree T chosen uniformly at random among all spanning trees of G.

uniform spanning tree

We make some simple observations first. Because G is connected, it has at least one spanning tree by Corollary 1.1.6. Moreover, for any edge $e \in G$, there always exists at least one spanning tree including it. To see this, let T' be any spanning tree of G, which exists by the previous observation. If $e \notin T'$, then we obtain a new spanning tree by adding e to T' and removing one edge $\neq e$ in the cycle created. As a consequence, the probability of inclusion $\mathbb{P}[e \in T]$ in a uniform spanning tree T cannot be 0. It is however possible for $\mathbb{P}[e \in T]$ to equal to 1 if removing e disconnects the graph. Such an edge is called a *bridge*.

bridge

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

Claim 3.3.41. *For a uniform spanning tree T of a connected graph G,*

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

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 3.3.41 is not trivial. The proof relies on the electrical network perspective. The key is a remarkable formula for the inclusion of an edge in a uniform spanning tree.

Theorem 3.3.42 (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 $e = \{x, y\}$*

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

Before explaining how this formula arises, we show that it implies Claim 3.3.41.

Proof of Claim 3.3.41. Recall that $\mathbb{P}[e' \in T] \neq 0$. By the law of total probability,

$$\mathbb{P}[e \in T] = \mathbb{P}[e \in T \,|\, e' \in T]\,\mathbb{P}[e' \in T] + \mathbb{P}[e \in T \,|\, e' \notin T]\,\mathbb{P}[e' \notin T],$$

so, since $\mathbb{P}[e' \in T] + \mathbb{P}[e' \notin T] = 1$, we can instead prove

$$\mathbb{P}[e \in T \,|\, e' \notin T] \geq \mathbb{P}[e \in T]. \tag{3.3.34}$$

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 (in the form of Corollary 3.3.28),

$$\mathscr{R}_{\mathcal{N}'}(x \leftrightarrow y) \geq \mathscr{R}_{\mathcal{N}}(x \leftrightarrow y),$$

and Kirchhoff's resistance formula (Theorem 3.3.42) gives (3.3.34). ∎

Remark 3.3.43. *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 $\mathcal{A} \subseteq 2^E$ is said to be* increasing *if $\omega \cup \{e\} \in \mathcal{A}$ whenever $\omega \in \mathcal{A}$. The event $\mathcal{A}$ 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 $\mathcal{A}$ or neither is. The law $\mathbb{P}_T$ of T has* negative associations *in the sense that for any two increasing events $\mathcal{A}$ and $\mathcal{B}$ that depend only on disjoint sets of edges, we have $\mathbb{P}_T[\mathcal{A} \cap \mathcal{B}] \leq \mathbb{P}_T[\mathcal{A}]\mathbb{P}_T[\mathcal{B}]$. See [LP16, Exercise 4.6].*

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

$$\mathscr{R}(x \leftrightarrow y) = \frac{v(x)}{\|i\|} = c(e)(v(x) - v(y)) = i(x, y), \tag{3.3.35}$$

where we used Ohm's law (i.e., (3.3.20)) as well as the fact that $c(e) = 1$, $v(y) = 0$, and $\|i\| = 1$. Note that $\|i\|$ and $i(x, y)$ are *not* the same quantity: although $\|i\| = 1$, $i(x, y)$ is only the current along the edge to y. Furthermore by the probabilistic interpretation of the current (Theorem 3.3.17), with $Z = \{y\}$,

$$i(x, y) = \mathbb{E}_x[N^Z_{x\to y} - N^Z_{y\to x}] = \mathbb{P}_x\left[(x, y) \text{ is traversed before } \tau_y\right]. \tag{3.3.36}$$

Indeed, started at x, $N^Z_{y\to x} = 0$ and $N^Z_{x\to y} \in \{0, 1\}$. 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 also Exercise 3.23.)

Wilson's method It will be somewhat easier 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 $\mathcal{P} = x_0 \sim \ldots \sim x_k$ be a walk in $\mathcal{N}$. The loop erasure of $\mathcal{P}$ 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 subwalk $x_{\ell+1} \sim \cdots \sim x_j$ from $\mathcal{P}$, and repeat. The result is self-avoiding, that is, a path, and is denoted by $\mathrm{LE}(\mathcal{P})$. *loop erasure*

Let v_0 be an arbitrary vertex of G, which we refer to as the root, and let T_0 be the subtree made up of v_0 alone. Starting with the root, order arbitrarily the vertices of G as $v_0, \ldots, v_{n-1}$. 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 $\mathcal{P}$ be the resulting walk.

2. Add the loop erasure $\mathrm{LE}(\mathcal{P})$ 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 3.3.44. *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 3.3.42. From (3.3.35) and (3.3.36), it suffices to prove that, for $e = \{x, y\}$,

$$\mathbb{P}_x\left[(x, y) \text{ is traversed before } \tau_y\right] = \mathbb{P}[e \in T],$$

where the probability on the left-hand side refers to random walk on $\mathcal{N}$ with unit resistances started at x and the probability on the right-hand side refers to a uniform spanning tree T on $\mathcal{N}$. Generate T using Wilson's method started at root $v_0 = y$ with the choice $v_1 = x$. If the walk from x to y during the first iteration of Wilson's

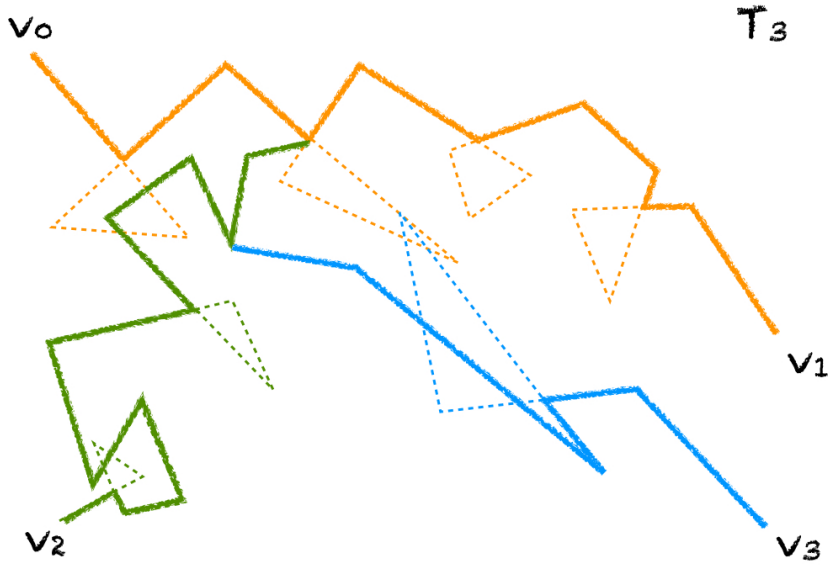

Figure 3.6: An illustration of Wilson's method. The dashed lines indicate erased loops.

method includes (x, y), then the loop erasure is simply $x \sim y$ and e is in T. On the other hand, if the walk from x to y does not include (x, y), then e cannot be used at a later stage because it would create a cycle. That immediately proves the theorem. ■

It remains to prove the claim.

Proof of Claim 3.3.44. 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 $G = (V, E)$ and root v_0. To each vertex $x \neq v_0$ in V, we assign an independent stack of "colored directed edges"

$$\mathcal{S}_0^x := (\langle x, Y_1^x \rangle_1, \langle x, Y_2^x \rangle_2, \ldots)$$

where each Y_j^x is chosen independently at random from the distribution $P(x, \cdot)$. In particular all Y_j^xs are neighbors of x in $\mathcal{N}$. The index j in $\langle x, Y_j^x \rangle_j$ is the *color* of the edge. It keeps track of the position of the edge in the original stack. (Picture $\mathcal{S}^x$ as a spring-loaded plate dispenser located on vertex x.)

color

We consider a process which involves popping edges off the stacks. We use the notation $\mathcal{S}^x$ to denote the *current* stack at x. The initial assignment of the stack is $\mathcal{S}^x := \mathcal{S}_0^x$ as above. Given the current stacks $(\mathcal{S}^x)_x$, we call *visible graph* the (colored) directed graph over V with edges $\mathrm{Top}(\mathcal{S}^x)$ for all $x \neq v_0$, where $\mathrm{Top}(\mathcal{S}^x)$ is the first edge in the current stack $\mathcal{S}^x$. The latter are referred to as *visible edges*. We denote the current visible graph by $\overrightarrow{G}_\odot$.

visible graph

visible edge

Note that $\overrightarrow{G}_\odot$ has out-degree 1 for all $x \neq v_0$ and the root has out-degree 0. In particular all undirected cycles in $\overrightarrow{G}_\odot$ are in fact directed cycles, and we refer to them simply as cycles. (Indeed a set of edges forming an undirected cycle that is not directed must have a vertex of out-degree 2.) Also recall the following characterization from Corollary 1.1.8: an acyclic, undirected subgraph with $|V|$ vertices and $|V|-1$ edges is a spanning tree of G. Hence if there is no cycle in $\overrightarrow{G}_\odot$, then it must be a spanning tree (as an undirected graph) where, furthermore, all edges point towards the root. Such a tree is also known as a *spanning arborescence*. Once that happens, we are done.

spanning arborescence

As the name suggests, a cycle popping algorithm proceeds by popping cycles in $\overrightarrow{G}_\odot$ off the tops of the stacks until a spanning arborescence is produced. That is, at every iteration, if $\overrightarrow{G}_\odot$ 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 $\overrightarrow{G}_\odot$ is revealed. See Figure 3.7 for an illustration.

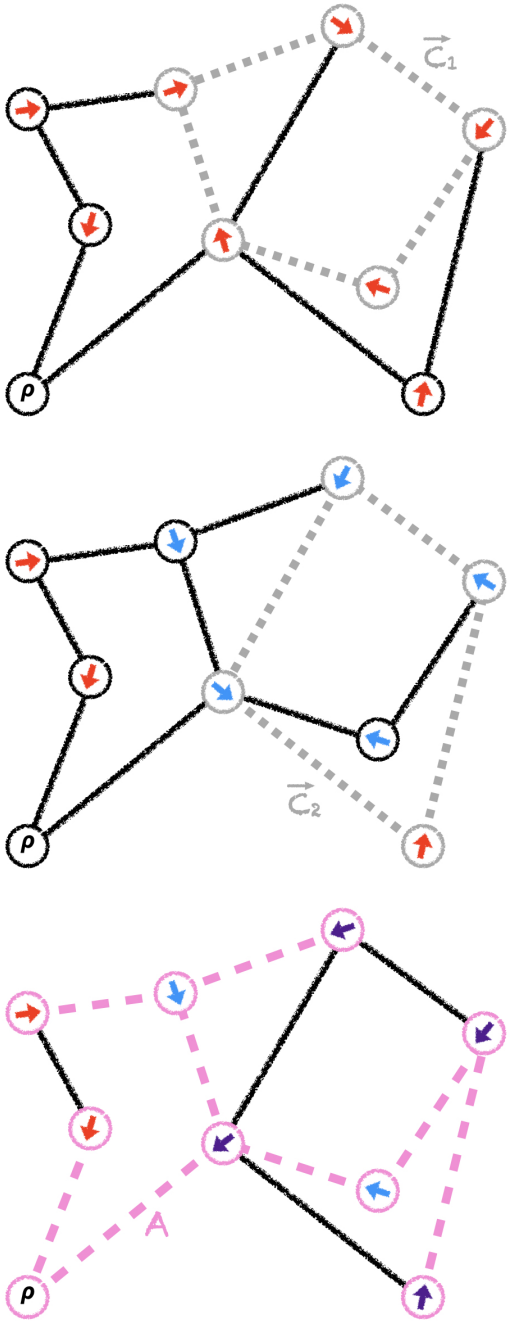

Figure 3.7: A realization of a cycle popping algorithm (from top to bottom). In all three figures, the underlying graph is G while the arrows depict the visible edges.

With these definitions in place, the proof of the claim involves the following steps.

(i) *Wilson's method is a cycle popping algorithm.* Recasting Wilson's method, we can think of the initial stacks $(\mathcal{S}_0^x)$ as corresponding to picking—ahead of time—all potential transitions in the random walks. With this representation, the algorithm 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. A key observation: 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 as they are closed. 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 resulting loop-erased path are not popped—note that they are part of the final arborescence.

(ii) *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, that is, 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 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 $\overrightarrow{G}_\odot$ if there exists a sequence of colored cycles $\overrightarrow{C}_1, \dots, \overrightarrow{C}_r = \overrightarrow{C}$ that can be popped in that order starting from $\overrightarrow{G}_\odot$. Note that, by this definition, $\overrightarrow{C}_1$ is a cycle in $\overrightarrow{G}_\odot$. Now we claim that if $\overrightarrow{C}'_1$ were popped first instead of $\overrightarrow{C}_1$, producing the new visible graph $\overrightarrow{G}'_\odot$, then $\overrightarrow{C}$ would still be poppable for $\overrightarrow{G}'_\odot$. This claim implies that, in any cycle popping algorithm, either an infinite number of cycles are popped or eventually all poppable cycles are popped—independently of the order—producing the same outcome. (Note that, while the same cycle may be popped more than once, the same *colored* cycle cannot.)

colored cycle

poppable cycle

To prove the claim, note first that if $\overrightarrow{C}'_1 = \overrightarrow{C}$ or if $\overrightarrow{C}'_1$ does not share a vertex with any of $\overrightarrow{C}_1, \dots, \overrightarrow{C}_r$ 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 $\overrightarrow{C}_1, \dots, \overrightarrow{C}_{j-1}$ 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 $\overrightarrow{C}_1, \ldots, \overrightarrow{C}_{j-1}$.

(iii) *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 3.1.25). To sum up, all cycle popping algorithms terminate and produce the same spanning arborescence. It remains to compute the distribution of the outcome.

(iv) *The arborescence has the desired distribution.* Let $\mathcal{A}$ be the spanning arborescence produced by any cycle popping algorithm on the stacks $(\mathcal{S}_0^x)$. 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, that is, 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 "popped edges" in $\mathcal{C}$ and the "final edges" in $\mathcal{A}$, that is,

$$\prod_{\vec{e} \in \mathcal{C} \cup \mathcal{A}} P(\vec{e}) = \Psi(\mathcal{A}) \prod_{\overrightarrow{C} \in \mathcal{C}} \Psi\left(\overrightarrow{C}\right),$$

where the function Ψ returns the product of the transition probabilities of a set of directed edges. Thanks to the product form on the right-hand side, summing over all possible $\mathcal{C}$s gives that the probability of producing $\mathcal{A}$ is proportional to $\Psi(\mathcal{A})$.

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 $\mathcal{A}$ because these are determined by $\mathcal{C}$. Secondly, we need for the collection

of possible $\mathcal{C}$s *not to vary with* $\mathcal{A}$. But it is clear that any arborescence could lie under any $\mathcal{C}$.

To see that we are done, let T be the undirected spanning tree corresponding to the outcome, $\mathcal{A}$, of Wilson's method. Then, because $P(x,y) = \frac{c(x,y)}{c(x)}$, we get

$$\Psi(\mathcal{A}) = \frac{W(T)}{\prod_{x \neq v_0} c(x)},$$

where note that the denominator does not depend on T. So if we forget the orientation of $\mathcal{A}$ which is determined by the root (i.e., sum over all choices of root), we get a spanning tree whose distribution is proportional to $W(T)$, as required. ■

Exercises

Exercise 3.1 (Reflection)**.** Give a rigorous proof of Theorem 3.1.9 through a formal application of the strong Markov property (i.e., specify f_t and F_t in Theorem 3.1.8).

Exercise 3.2 (Time of k-th return)**.** Give a rigorous proof of (3.1.1) through a formal application of the strong Markov property (i.e., specify f_t and F_t in Theorem 3.1.8).

Exercise 3.3 (Tightness of Matthews' bounds)**.** Show that the bounds (3.1.6) and (3.1.7) are tight up to smaller order terms for the coupon collector problem (Example 2.1.4). [Hint: State the problem in terms of the cover time of a random walk on the complete graph with self-loops.]

Exercise 3.4 (Pólya's urn: a suprisingly simple formula)**.** Consider the setting of Example 3.1.49. Prove that

$$\mathbb{P}[G_t = m+1] = \binom{t}{m} \frac{m!(t-m)!}{(t+1)!}.$$

[Hint: Consider the probability of one particular sequence of outcomes producing the desired event.]

Exercise 3.5 (Optional stopping theorem)**.** Give a rigorous proof of the remaining cases of the optional stopping theorem (Theorem 3.1.38).

Exercise 3.6 (Supermartingale inequality)**.** Let (M_t) be a nonnegative, supermartingale. Show that, for any $b > 0$,

$$\mathbb{P}\left[\sup_{s \geq 0} M_s \geq b\right] \leq \frac{\mathbb{E}[M_0]}{b}.$$

[Hint: Mimic the proof of the submartingale case.]

Exercise 3.7 (Azuma-Hoeffding: a second proof)**.** This exercise leads the reader through an alternative proof of the Azuma-Hoeffding inequality.

(i) Show that for all $x \in [-1, 1]$ and $a > 0$

$$e^{ax} \leq \cosh a + x \sinh a.$$

(ii) Use a Taylor expansion to show that for all x

$$\cosh x \leq e^{x^2/2}.$$

(iii) 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

$$\mathbb{E}\left[X_{i_1} \cdots X_{i_k}\right] = 0, \qquad \forall\, 1 \leq k \leq n,\ \forall\, 1 \leq i_1 < \cdots < i_k \leq n. \quad (3.3.37)$$

Show, using (i) and (ii), that for all $b > 0$

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq b\right] \leq \exp\left(-\frac{b^2}{2\sum_{i=1}^n c_i^2}\right).$$

(iv) Prove that (iii) implies a variant of the Azuma-Hoeffding inequality (Theorem 3.2.1) for bounded increments.

(v) Show that the random variables in Exercise 2.6 (after centering) do not satisfy (3.3.37) (without using the claim in part (ii) of that exercise).

Exercise 3.8 (Lipschitz condition)**.** Give a rigorous proof of Lemma 3.2.31.

Exercise 3.9 (Lower bound on expected spectral norm)**.** Let A be an $n \times n$ random matrix. Assume that the entries $A_{i,j}$, $i, j = 1, \ldots, n$, are independent, centered random variables in $[-1, 1]$. Suppose further that there is $0 < \sigma^2 < +\infty$ such that $\mathrm{Var}[A_{ij}] \geq \sigma^2$ for all i, j. Show that there is $0 < c < +\infty$ such that

$$\mathbb{E}\|A\| \geq c\sqrt{n},$$

for n large enough. [Hint: Use the fact that $\|A\|^2 \geq \|A\mathbf{e}_1\|^2$ together with Chebyshev's inequality.]

Exercise 3.10 (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 3.11 (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 $h(0) = 1$.

(i) When $p > 1/2$, show that there is more than one bounded extension of h to $\mathbb{Z}_+ \backslash \{0\}$ that is harmonic on $\mathbb{Z}_+ \backslash \{0\}$. [Hint: Consider $\mathbb{P}_x[\tau_0 = +\infty]$.]

(ii) When $p \leq 1/2$, show that there exists a unique bounded extension of h to $\mathbb{Z}_+ \backslash \{0\}$ that is harmonic on $\mathbb{Z}_+ \backslash \{0\}$.

Exercise 3.12 (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.

(i) Let $h : V \to \mathbb{R}$ be a function on V. Prove the maximum principle: if h is harmonic on W, that is, it satisfies

$$h(x) = \frac{1}{c(x)} \sum_{y \sim x} c(x, y)h(y), \qquad \forall x \in W,$$

and if h achieves its supremum on W, then h is constant on $W \cup \partial_{\mathrm{V}} W$, where

$$\partial_{\mathrm{V}} W = \{z \in V \setminus W \, : \, \exists y \in W, y \sim z\}.$$

(ii) 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 (i) to prove that $h_1 \equiv h_2$.

Exercise 3.13 (Poisson equation: uniqueness)**.** Show that u is the unique solution of the system in Theorem 3.3.6 under the conditions of Theorem 3.3.1. [Hint: Use Theorem 3.3.9 and mimic the proof of Theorem 3.3.1.]

Exercise 3.14 (Effective resistance: metric)**.** Show that effective resistances between pairs of vertices form a metric.

Exercise 3.15 (Dirichlet principle: proof)**.** Prove Theorem 3.3.25.

Exercise 3.16 (Martingale problem)**.** Let V be countable, let (X_t) be a stochastic process adapted to $(\mathcal{F}_t)$ and taking values in V, and let P be a transition probability on V with associated Laplacian operator Δ. Show that the following are equivalent:

(i) The process (X_t) is a Markov chain with transition probability P.

(ii) For any bounded measurable function $f : V \to \mathbb{R}$, the process

$$M_t^f = f(X_t) - \sum_{s=0}^{t-1} \Delta f(X_s),$$

is a martingale with respect to $(\mathcal{F}_t)$.

Exercise 3.17 (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 $\mathscr{R}(0 \leftrightarrow Z_n) = \Theta(\log n)$. [Hint: Use the Nash-Williams inequality and the method of random paths.]

Exercise 3.18 (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.

(i) Show that there are at least $2d - n$ vertices $x \neq a, z$ such that $a \sim x \sim z$ is a path.

(ii) Prove that
$$\mathbb{E}_a(\tau_{a,z}) \leq \frac{2dn}{2d-n}.$$

Exercise 3.19 (Independent-coordinate random walk)**.** Give a rigorous proof that the two networks in Example 3.3.32 are roughly equivalent.

Exercise 3.20 (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 $x, y \in V$ *rough isometry*

$$\alpha^{-1} d(x, y) - \beta \leq d'(\phi(x), \phi(y)) \leq \alpha d(x, y) + \beta,$$

where d and d' are the graph distances on G and G' respectively, and furthermore all vertices in G' are within distance β 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 3.21 (Random walk on the cycle: hitting time)**.** Use the commute time identity (Theorem 3.3.34) to compute $\mathbb{E}_x[\tau_y]$ in Example 3.3.35 in the case $d = 1$. Give a second proof using a direct martingale argument.

Exercise 3.22 (Random walk on the binary tree: cover time)**.** As in Example 3.3.15, let $\mathcal{N}$ be the rooted binary tree with n levels $\widehat{\mathbb{T}}_2^n$ and equal conductances on all edges.

(i) Show that the maximal hitting time $\mathbb{E}_a\tau_b$ is achieved for a and b such that their most recent common ancestor is the root 0. Furthermore argue that in that case $\mathbb{E}_a[\tau_b] = \mathbb{E}_a[\tau_{a,0}]$, where recall that $\tau_{a,0}$ is the commute time between a and 0.

(ii) Use the commute time identity (Theorem 3.3.34) and Matthews' cover time bounds (Theorem 3.1.27) to give an upper bound on the mean cover time of the order of $O(n^2 2^n)$.

Exercise 3.23 (Markov chain tree theorem)**.** Let P be the transition matrix of a finite, irreducible Markov chain with stationary distribution π. Let G be the directed graph corresponding to the positive transitions of P. For an arborescence $\mathcal{A}$ of G, define its weight as

$$\Psi(\mathcal{A}) = \prod_{\vec{e} \in \mathcal{A}} P(\vec{e}).$$

Consider the following process on spanning arborescences over G. Let ρ be the root of the current spanning arborescence $\mathcal{A}$. Pick an outgoing edge $\vec{e} = (\rho, x)$ of ρ according to $P(\rho, \cdot)$. Edge $\vec{e}$ is not in $\mathcal{A}$ by definition of an arborescence. 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.

(i) Show that this chain is irreducible.

(ii) Show that Ψ is a stationary measure for this chain.

(iii) Prove the *Markov chain tree theorem*: The stationary distribution π of P is proportional to

$$\pi_x = \sum_{\mathcal{A} \,:\, \mathrm{root}(\mathcal{A}) = x} \Psi(\mathcal{A}).$$

Bibliographic Remarks

Section 3.1 Picking up where Appendix B leaves off, Sections 3.1.1 and 3.1.3 largely follow the textbooks [Wil91] and [Dur10], which contain excellent introductions to martingales. The latter also covers Markov chains, and includes the proofs we skipped here. Theorem 3.1.11 is proved in [Dur10, Theorem 4.3.2]. Many more results like Corollary 3.1.24 can be derived from the occupation measure identity; see for example [AF, Chapter 2]. The upper bound in Theorem 3.1.27 was first proved by Matthews [Mat88].

Section 3.2 The Azuma-Hoeffding inequality is due to Hoeffding [Hoe63] and Azuma [Azu67]. The version of the inequality in Exercise 3.7 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]. In its current form, it appears in [McD89]. Example 3.2.11 is taken from [MU05, Section 12.5]. The presentation in Section 3.2.3 follows [AS11, Section 7.3]. Claim 3.2.16 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 3.2.21 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], [CL06], or [vdH17]. Example 3.2.12 borrows from [BLM13, Section 7.1] and [Pet, Section 6.3]. General references on the concentration of measure phenomenon and concentration inequalities are [Led01] and [BLM13]. See [BCB12] or [LS20] for an introduction to bandit problems; or [AJKS22] for an introduction to the sample complexity of the more general reinforcement learning problem. 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 [vH16]. Section 3.2.6 is based on [vH16, Section 4.3]. In particular, a proof of Talagrand's inequality (Theorem 3.2.32) can be found there. See also [AS11, Chapter 7] or [BLM13, Chapter 7].

Section 3.3 Section 3.3.1 is based partly on [Nor98, Sections 4.1-2], [Ebe, Sections 0.3, 1.1-2, 3.1-2], and [Bre17, Sections 7.3, 17.1]. The material in Sections 3.3.2-3.3.5 borrows from [LPW06, Chapters 9, 10], [AF, Chapters 2, 3] and, especially, [LP16, Sections 2.1-2.6, 4.1-4.2, 5.5]. Foster's theorem (Theorem 3.3.12)

is from [Fos53]. 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 for example [BV04]. The Nash-Williams inequality is due to Nash-Williams [NW59]. The result in Example 3.3.27 is due to R. Lyons [Lyo90]. Theorem 3.3.33 is due to Kanai [Kan86]. The commute time identity was proved by Chandra, Raghavan, Ruzzo, Smolensky and Tiwari [CRR$^+$89]. 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]. 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 had previously been established by Pemantle [Pem91] using the Aldous-Broder method. For more on negative correlation in uniform spanning trees, see for example [LP16, 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].